ON GENERALIZATION ERROR BOUNDS OF NOISY GRADIENT METHODS FOR NON-CONVEX LEARNING

Jian Li * Tsinghua University

Xuanyuan Luo † Tsinghua University

Mingda Qiao ‡ Stanford University

ABSTRACT

Generalization error (also known as the out-of-sample error) measures how well the hypothesis learned from training data generalizes to previously unseen data. Proving tight generalization error bounds is a central question in statistical learning theory. In this paper, we obtain generalization error bounds for learning general non-convex objectives, which has attracted significant attention in recent years. We develop a new framework, termed *Bayes-Stability*, for proving *algorithm-dependent* generalization error bounds. The new framework combines ideas from both the PAC-Bayesian theory and the notion of algorithmic stability. Applying the Bayes-Stability method, we obtain new data-dependent generalization bounds for stochastic gradient Langevin dynamics (SGLD) and several other noisy gradient methods (e.g., with momentum, mini-batch and acceleration, Entropy-SGD). Our result recovers (and is typically tighter than) a recent result in Mou et al. (2018) and improves upon the results in Pensia et al. (2018). Our experiments demonstrate that our data-dependent bounds can distinguish randomly labelled data from normal data, which provides an explanation to the intriguing phenomena observed in Zhang et al. (2017a). We also study the setting where the total loss is the sum of a bounded loss and an additional ℓ_2 regularization term. We obtain new generalization bounds for the continuous Langevin dynamic in this setting by developing a new Log-Sobolev inequality for the parameter distribution at any time. Our new bounds are more desirable when the noise level of the process is not very small, and do not become vacuous even when T tends to infinity.

1 INTRODUCTION

Non-convex stochastic optimization is the major workhorse of modern machine learning. For instance, the standard supervised learning on a model class parametrized by $\mathbb{R}^d$ can be formulated as the following optimization problem:

$$\min_{w\in\mathbb{R}^d} \mathop{\mathbb{E}}_{z\sim\mathcal{D}} [F(w,z)],$$

where w denotes the model parameter, $\mathcal{D}$ is an unknown data distribution over the instance space $\mathcal{Z}$, and $F : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ is a given objective function which may be non-convex. A learning algorithm takes as input a sequence $S = (z_1, z_2, \ldots, z_n)$ of n data points sampled i.i.d. from $\mathcal{D}$, and outputs a (possibly randomized) parameter configuration $\hat{w} \in \mathbb{R}^d$.

A fundamental problem in learning theory is to understand the *generalization performance* of learning algorithms—is the algorithm guaranteed to output a model that generalizes well to the data distribution $\mathcal{D}$? Specifically, we aim to prove upper bounds on the *generalization error* $\mathrm{err}_{\mathrm{gen}}(S) = \mathcal{L}(\hat{w},\mathcal{D}) - \mathcal{L}(\hat{w},S)$, where $\mathcal{L}(\hat{w},\mathcal{D}) = \mathbb{E}_{z\sim\mathcal{D}}[\mathcal{L}(\hat{w},z)]$ and $\mathcal{L}(\hat{w},S) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\hat{w},z_i)$ are the population and empirical losses, respectively. We note that the loss function $\mathcal{L}$ (e.g., the 0/1 loss) could be different from the objective function F (e.g., the cross-entropy loss) used in the training process (which serves as a surrogate for the loss $\mathcal{L}$).

*lijian83@mail.tsinghua.edu.cn

†luo-xy19@mails.tsinghua.edu.cn

‡mqiao@stanford.edu

Classical learning theory relates the generalization error to various complexity measures (e.g., the VC-dimension and Rademacher complexity) of the model class. Directly applying these classical complexity measures, however, often fails to explain the recent success of over-parametrized neural networks, where the model complexity significantly exceeds the amount of available training data (see e.g., Zhang et al. (2017a)). By incorporating certain data-dependent quantities such as margin and compressibility into the classical framework, some recent work (e.g., Bartlett et al. (2017); Arora et al. (2018); Wei & Ma (2019)) obtains more meaningful generalization bounds in the deep learning context.

An alternative approach to generalization is to provide algorithm-dependent bounds. One celebrated example along this line is the algorithmic stability framework initiated by Bousquet & Elisseeff (2002). Roughly speaking, the generalization error can be bounded by the stability of the algorithm (see Section 2 for the details). Using this framework, Hardt et al. (2016) study the stability (hence the generalization) of stochastic gradient descent (SGD) for both convex and non-convex functions. Their work motivates recent study of the generalization performance of several other gradient-based optimization methods: Kuzborskij & Lampert (2018); London (2016); Chaudhari et al. (2017); Raginsky et al. (2017); Mou et al. (2018); Pensia et al. (2018); Chen et al. (2018).

In this paper, we study the algorithmic stability and generalization performance of various iterative gradient-based method, with certain continuous noise injected in each iteration, in a non-convex setting. As a concrete example, we consider the stochastic gradient Langevin dynamics (SGLD) (see Raginsky et al. (2017); Mou et al. (2018); Pensia et al. (2018)). Viewed as a variant of SGD, SGLD adds an isotropic Gaussian noise at every update step:

$$W_t = W_{t-1} - \eta_t g_t(W_{t-1}) + \sqrt{\frac{\eta_t}{\beta}} N(0; I_d), \tag{1}$$

where $g_t(W_{t-1})$ denotes either the full gradient or the gradient over a mini-batch sampled from training dataset. We also study a continuous version of (1), which is the dynamic defined by the following stochastic differential equation (SDE):

$$dW_t = -\nabla F(W_t)\, dt + \sqrt{\frac{2}{\beta}}\, dB_t, \tag{2}$$

where B_t is the standard Brownian motion.

1.1 Related Work

Most related to our work is the study of algorithm-dependent generalization bounds of stochastic gradient methods. Hardt et al. (2016) first study the generalization performance of SGD via algorithmic stability. They prove a generalization bound that scales linearly with the number of iterations, when the loss function is convex, but their results for general non-convex optimization are more restricted. London (2017) and Rivasplata et al. (2018) also combine ideas from both PAC-Bayesian and algorithm stability. However, these works are essentially different from ours. In London (2017), the prior and posterior are distributions on the hyperparameter space instead of distributions on the hypothesis space. Rivasplata et al. (2018) study the hypothesis stability measured by the distance on the hypothesis space in a setting where the returned hypothesis (model parameter) is perturbed by a Gaussian noise. Our work is a follow-up of the recent work by Mou et al. (2018), in which they provide generalization bounds for SGLD from both stability and PAC-Bayesian perspectives. Another closely related work by Pensia et al. (2018) derives similar bounds for noisy stochastic gradient methods, based on the information theoretic framework of Xu & Raginsky (2017). However, their bounds scale as $O(\sqrt{T/n})$ (n is the size of the training dataset) and are sub-optimal even for SGLD.

We acknowledge that besides the algorithm-dependent approach that we follow, recent advances in learning theory aim to explain the generalization performance of neural networks from many other perspectives. Some of the most prominent ideas include bounding the network capacity by the norms of weight matrices Neyshabur et al. (2015); Liang et al. (2019), margin theory Bartlett et al. (2017); Wei et al. (2019), PAC-Bayesian theory Dziugaite & Roy (2017); Neyshabur et al. (2018); Dziugaite & Roy (2018), network compressibility Arora et al. (2018), and over-parametrization Du et al. (2019); Allen-Zhu et al. (2019); Zou et al. (2018); Chizat et al. (2019). Most of these results are stated in the context of neural networks (some are tailored to networks with specific architecture), whereas our work addresses generalization in non-convex stochastic optimization in general. We

also note that some recent work provides explanations for the phenomenon reported in Zhang et al. (2017a) from a variety of different perspectives (e.g., Bartlett et al. (2017); Arora et al. (2018; 2019)).

Welling & Teh (2011) first consider stochastic gradient Langevin dynamics (SGLD) as a sampling algorithm in the Bayesian inference context. Raginsky et al. (2017) give a non-asymptotic analysis and establish the finite-time convergence guarantee of SGLD to an approximate global minimum. Zhang et al. (2017b) analyze the hitting time of SGLD and prove that SGLD converges to an approximate local minimum. These results are further improved and generalized to a family of Langevin dynamics based algorithms by the subsequent work of Xu et al. (2018).

1.2 Overview of Our Results

In this paper, we provide generalization guarantees for the noisy variants of several popular stochastic gradient methods.

The Bayes-Stability method and data-dependent generalization bounds. We develop a new method for proving generalization bounds, termed as Bayes-Stability, by incorporating ideas from the PAC-Bayesian theory into the stability framework. In particular, assuming the loss takes value in $[0, C]$, our method shows that the generalization error is bounded by $2C\mathbb{E}_z[\sqrt{2\mathrm{KL}(P; Q_z)}]$ and $2C\mathbb{E}_z[\sqrt{2\mathrm{KL}(Q_z; P)}]$, where P is a prior distribution independent of the training set S, and Q_z is the expected posterior distribution conditioned on $z_n = z$ (i.e., the last training data is z). The formal definition and the results can be found in Definition 5 and Theorem 7.

Inspired by Lever et al. (2013), instead of using a fixed prior distribution, we bound the KL-divergence from the posterior to a distribution-dependent prior. This enables us to derive the following generalization error bound that depends on the expected norm of the gradient along the optimization path:

$$\mathrm{err}_{\mathrm{gen}} = O\left(\frac{C}{n}\sqrt{\mathbb{E}_S\left[\sum_{t=1}^{T}\frac{\epsilon_t^2}{\sigma_t^2}g_e(t)\right]}\right). \tag{3}$$

Here S is the dataset and $g_e(t) = \mathbb{E}_{W_{t-1}}[\frac{1}{n}\sum_{i=1}^{n}\|\nabla F(W_{t-1}; z_i)\|^2]$ is the expected empirical squared gradient norm at step t; see Theorem 11 for the details.

Compared with the previous $O\left(\frac{LC}{n}\sqrt{\sum_t \frac{\epsilon_t^2}{\sigma_t^2}}\right)$ bound in (Mou et al., 2018, Theorem 1), where L is the global Lipschitz constant of the loss, our new bound (3) depends on the data distribution and is typically tighter (as the gradient norm is at most L). In modern deep neural networks, the worst-case Lipschitz constant L can be quite large, and typically much larger than the expected empirical gradient norm along the optimization trajectory. Specifically, in the later stage of the training, the expected empirical gradient is small (see Figure 1(d) for the details). Hence, our generalization bound does not grow much even if we train longer at this stage.

Our new bound also offers an explanation to the difference between training on correct and random labels observed by Zhang et al. (2017a). In particular, we show empirically that the sum of expected squared gradient norm (along the optimization path) is significantly higher when the training labels are replaced with random labels (Section 3.1, Figure 1, Appendix C.2).

We would also like to mention the PAC-Bayesian bound (for SGLD with ℓ_2-regularization) proposed by Mou et al. (2018). (This bound is different from what we mentioned before; see Theorem 2 in their paper.) Their bound scales as $O(1/\sqrt{n})$ and the numerator of their bound has a similar sum of gradient norms (with a decaying weight if the regularization coefficient > 0). Their bound is based on the PAC-Bayesian approach and holds with high probability, while our bound only holds in expectation.

Extensions. We remark that our technique allows for an arguably simpler proof of (Mou et al., 2018, Theorem 1); the original proof is based on SDE and Fokker-Planck equation. More importantly, our technique can be easily extended to handle mini-batches and a variety of general settings as follows.

1. Extension to other gradient-based methods. Our results naturally extends to other noisy stochastic gradient methods including momentum due to Polyak (1964) (Theorem 26), Nes-

terov's accelerated gradient method in Nesterov (1983) (Theorem 26), and Entropy-SGD proposed by Chaudhari et al. (2017) (Theorem 27).

2. Extension to general noises. The proof of the generalization bound in Mou et al. (2018) relies heavily on that the noise is Gaussian[1], which makes it difficult to generalize to other noise distributions such as the Laplace distribution. In contrast, our analysis easily carries over to the class of log-Lipschitz noises (i.e., noises drawn from distributions with Lipschitz log densities).
3. Pathwise stability. In practice, it is also natural to output a certain function of the entire optimization path, e.g., the one with the smallest empirical risk or a weighted average. We show that the same generalization bound holds for all such variants (Remark 12). We note that the analysis in an independent work of Pensia et al. (2018) also satisfies this property, yet their bound is $O\left(\sqrt{C^2L^2n^{-1}\sum_{t=1}^{T}\eta_t^2/\sigma_t^2}\right)$ (see Corollary 1 in their work), which scales at a slower $O(1/\sqrt{n})$ rate (instead of $O(1/n)$) when dealing with C-bounded loss[2].

Generalization bounds with ℓ_2 regularization via Log-Sobolev inequalities. We also study the setting where the total objective function F is the sum of a C-bounded differentiable objective F_0 and an additional l_2 regularization term $\frac{\lambda}{2}\|w\|_2^2$. In this case, F can be treated as a perturbation of a quadratic function, and the continuous Langevin dynamics (CLD) is well understood for quadratic functions. We obtain two generalization bounds for CLD, both via the technique of Log-Sobolev inequalities, a powerful tool for proving the convergence rate of CLD. One of our bounds is as follows (Theorem 15):

$$\mathrm{err}_{\mathrm{gen}} \quad \frac{2e^{4C}CL}{n}\sqrt{1-\exp\left(-\frac{T}{e^{8C}}\right)}. \tag{4}$$

The above bound has the following advantages:

1. Applying $e^{-x} \geq 1-x$, one can see that our bound is at most $O(\sqrt{T}/n)$, which matches the previous bound in (Mou et al., 2018, Proposition 3[3]).
2. As time T grows, the bound is upper bounded by and approaches to $2e^{4C}CLn^{-1}\sqrt{\pi}$ (unlike the previous $O(\sqrt{T}/n)$ bound that goes to infinity as $T \to +\infty$).
3. If the noise level is not so small (i.e., σ is not very large), the generalization bound is quite desirable.

Our analysis is based on a Log-Sobolev inequality (LSI) for the parameter distribution at time t, whereas most known LSIs only hold for the stationary distribution of the Markov process. We prove the new LSI by exploiting the variational formulation of the entropy formula.

2 PRELIMINARIES

Notations. We use D to denote the data distribution. The training dataset $S = (z_1, \dots, z_n)$ is a sequence of n independent samples drawn from D. $S, S' \in Z^n$ are called neighboring datasets if and only if they differ at exactly one data point (we could assume without loss of generality that $z_n \neq z_n'$). Let $F(w,z)$ and $L(w,z)$ be the objective and the loss functions, respectively, where $w \in \mathbb{R}^d$ denotes a model parameter and $z \in Z$ is a data point. Define $F(w,S) = \frac{1}{n}\sum_{i=1}^{n} F(w,z_i)$ and $F(w,D) = \mathbb{E}_{z\sim D}[F(w,z)]$; $L(w,S)$ and $L(w,D)$ are defined similarly. A learning algorithm A takes as input a dataset S and outputs a parameter $w \in \mathbb{R}^d$ randomly. Let G be the set of all possible mini-batches. $G_n = \{B \in G : n \in B\}$ denotes the collection of mini-batches that contain the n-th data point, while $\overline{G_n} = G \setminus G_n$. Let $\mathrm{diam}(A) = \sup_{x,y\in A}\|x-y\|_2$ denote the diameter of a set A.

[1] In particular, their proof leverages the Fokker-Planck equation, which describes the time evolution of the density function associated with the Langevin dynamics and can only handle Gaussian noise.

[2] They assume the loss is sub-Gaussian. By Hoeffding's lemma, C-bounded random variables are sub-Gaussian with parameter $\frac{C}{2}$.

[3] The proof of their $O(\sqrt{T}/n)$ bound can be easily extended to our setting with ℓ_2 regularization.

De nition 1 (L-lipschitz). A function $F : R^d \ Z \ ! \ R$ is L-lipschitz if and only if $jF(w_1; z) \ F(w_2; z)j \ L \ kw_1 \ w_2k_2$ holds for any $w_1; w_2 \ 2 \ R^d$ and $z \ 2 \ Z$.

De nition 2 (Expected generalization error). The expected generalization error of a learning algorithm A is de ned as

$$err_{gen} := \underset{S \ D^n}{E}[err_{gen}(S)] = \underset{S \ D^n;A}{E}[L(A(S); D) \ L \ (A(S); S)]:$$

Algorithmic Stability. Intuitively, a learning algorithm that is stable (i.e., a small perturbation of the training data does not affect its output too much) can generalize well. In the seminal work of Bousquet & Elisseeff (2002) (see also Hardt et al. (2016)), the authors formally de ned algorithmic stability and established a close connection between the stability of a learning algorithm and its generalization performance.

De nition 3 (Uniform stability). (Bousquet & Elisseeff (2002); Elisseeff et al. (2005)) A randomized algorithm A is $_n$-uniformly stable w.r.t. loss L, if for all neighboring sets $S; S^0 \ 2 \ Z^n$, it holds that

$$\sup_{z2Z} jE_A[L(w_S; z)] \ E_A[L(w_{S^0}; z)]j \ _n;$$

where w_S and w_{S^0} denote the outputs of A on S and S^0 respectively.

Lemma 4 (Generalization in expectation). (Hardt et al. (2016)) Suppose a randomized algorithm A is $_n$-uniformly stable. Then $jerr_{gen}j \ _n$.

3 BAYES-STABILITY METHOD

In this section, we incorporate ideas from the PAC-Bayesian theory (see e.g., Lever et al. (2013)) into the algorithmic stability framework. Combined with the technical tools introduced in previous sections, the new framework enables us to prove tighter data-dependent generalization bounds.

First, we de ne the posterior of a dataset and the posterior of a single data point.

De nition 5 (Single-point posterior). Let Q_S be the posterior distribution of the parameter for a given training dataset $S = (z_1; : : : ; z_n)$. In other words, it is the probability distribution of the output of the learning algorithm on dataset S (e.g., for T iterations of SGLD in (1), Q_S is the pdf of W_T). The single-point posterior $Q_{(i;z)}$ is de ned as

$$Q_{(i;z)} = \underset{(z_1;:::;z_{i \ 1};z_{i+1};:::;z_n)}{E} Q_{(z_1;:::;z_{i \ 1};z;z_{i+1};:::;z_n)} :$$

For convenience, we make the following natural assumption on the learning algorithm:

Assumption 6 (Order-independent). For any xed dataset $S = (z_1; : : : ; z_n)$ and any permutation p, Q_S is the same as Q_{S^p}, where $S^p = (z_{p_1}; : : : ; z_{p_n})$.

Assumption 6 implies $Q_{(1;z)} = \ = Q_{(n;z)}$, so we use Q_z as a shorthand for $Q_{(i;z)}$ in the following. Note that this assumption can be easily satis ed by letting the learning algorithm randomly permute the training data at the beginning. It is also easy to verify that both SGD and SGLD satisfy the order-independent assumption.

Now, we state our new Bayes-stability framework, which holds for any prior distribution P over the parameter space that is independent of the training dataset S.

Theorem 7 (Bayes-Stability). Suppose the loss function $l(w; z)$ is C-bounded and the learning algorithm is order-independent (Assumption 6). Then for any prior distribution P not depending on S, the generalization error is bounded by both $2C \ E_z \sqrt{\frac{}{2KL(P; Q_z)}}$ and $2C \ E_z \sqrt{\frac{}{2KL(Q_z; P)}}$.

Remark 8. Our Bayes-Stability framework originates from the algorithmic stability framework, and hence is similar to the notions of uniform stability and leave-one-out error (see Elisseeff et al. (2003)). However, there are important differences. Uniform stability is a distribution-independent property, while Bayes-Stability can incorporate the information of the data distribution (through the prior P). Leave-one-out error measures the loss of a learned model on an unseen data point, yet Bayes-Stability focuses on the extent to which a single data point affects the outcome of the learning algorithm (compared to the prior).

To establish an intuition, we first apply this framework to obtain an expectation generalization bound for (full) gradient Langevin dynamics (GLD), which is a special case of SGLD in (1) (i.e., GLD uses the full gradient $r_w F(W_{t-1}; S)$ as $g_t(W_{t-1})$).

Theorem 9. Suppose that the loss function l is C-bounded. Then we have the following expected generalization bound for T iterations of GLD:

$$\text{err}_{\text{gen}} \leq \frac{\sqrt{2}C}{n} \sqrt{\mathop{\mathbb{E}}_{S \sim D^n} \left[\sum_{t=1}^{T} \frac{\sigma_t^2}{2} g_e(t) \right]};$$

where $g_e(t) = \mathbb{E}_{w \sim W_{t-1}}[\frac{1}{n} \sum_{i=1}^{n} \|r_w F(w; z_i)\|_2^2]$ is the empirical squared gradient norm, and W_t is the parameter at step t of GLD.

Proof The proof builds upon the following technical lemma, which we prove in Appendix A.2.

Lemma 10. Let $(W_0, \ldots, W_T)$ and $(W_0', \ldots, W_T')$ be two independent sequences of random variables such that for each $t \in \{0, \ldots, T\}$, W_t and W_t' have the same support. Suppose W_0 and W_0' follow the same distribution. Then,

$$\text{KL}(W_{\leq T}; W'_{\leq T}) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w_{<t} \sim W_{<t}} [\text{KL}(W_t | W_{<t} = w_{<t}; W_t' | W'_{<t} = w_{<t})];$$

where $W_{\leq t}$ denotes $(W_0, \ldots, W_t)$ and $W_{<t}$ denotes $W_{\leq t-1}$.

Define $P = \mathbb{E}_{\bar{S} \sim D^{n-1}}[Q_{(\bar{S}; 0)}]$, where 0 denotes the zero data point (i.e., $F(w; 0) = 0$ for any w). Theorem 7 shows that

$$\text{err}_{\text{gen}} \leq 2C \mathop{\mathbb{E}}_{z} \sqrt{2\text{KL}(Q_z; P)}. \tag{5}$$

By the convexity of KL-divergence, for a fixed $z \in Z$, we have

$$\text{KL}(Q_z; P) = \text{KL}\left(\mathop{\mathbb{E}}_{\bar{S}}[Q_{(\bar{S}; z)}]; \mathop{\mathbb{E}}_{\bar{S}}[Q_{(\bar{S}; 0)}] \right) \leq \mathop{\mathbb{E}}_{\bar{S}} \left[\text{KL}\left(Q_{(\bar{S}; z)}; Q_{(\bar{S}; 0)} \right) \right]. \tag{6}$$

Let $(W_t)_{t \geq 0}$ and $(W_t')_{t \geq 0}$ be the training process of GLD for $S = (\bar{S}, z)$ and $S' = (\bar{S}, 0)$, respectively. Note that for a fixed $w_{<t}$, both $W_t | W_{<t} = w_{<t}$ and $W_t' | W'_{<t} = w_{<t}$ are Gaussian distributions. Since $\text{KL}(N(\mu_1, \sigma^2 I); N(\mu_2, \sigma^2 I)) = \frac{\|\mu_1 - \mu_2\|_2^2}{2\sigma^2}$ (see Lemma 18 in Appendix A.2).

$$\text{KL}(W_t | W_{<t} = w_{<t}; W_t' | W'_{<t} = w_{<t}) = \frac{\sigma_t^2 \|r_w F(w_{t-1}; z)\|_2^2}{2n^2}.$$

Applying Lemma 10 and $\text{KL}(W_T; W_T') \leq \text{KL}(W_{\leq T}; W'_{\leq T})$ gives

$$\text{KL}(Q_S; Q_{S'}) \leq \frac{1}{2n^2} \sum_{t=1}^{T} \sigma_t^2 \mathop{\mathbb{E}}_{w \sim W_{t-1}} \|r_w F(w; z)\|_2^2.$$

Recall that W_{t-1} is the parameter at step $t-1$ using $S = (\bar{S}, z)$ as dataset. In this case, we can rewrite z as z_n since it is the n-th data point of S. Note that SGLD satisfies the order-independent assumption, we can rewrite z as z_i for all $i \in [n]$. Together with (5), (6), and using $\frac{1}{n} \sum_{i=1}^{n} \sqrt{x_i} \leq \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i}$, we can prove this theorem. ∎

More generally, we give the following bound for SGLD. The proof is similar to that of Theorem 9; the difference is that we need to bound the KL-divergence between two Gaussian mixtures instead of two Gaussians. This proof is more technical and deferred to Appendix A.3.

Theorem 11. Suppose that the loss function l is C-bounded and the objective function f is L-lipschitz. Assume that the following conditions hold:

1. Batch size b of $n/2$.

2. Learning rate $\eta_t = \eta_t = (20L)$.

Then, the following expected generalization error bound holds for T iterations of SGLD (1):

$$\text{err}_{\text{gen}} \quad \frac{8\sqrt{2}C}{n} \sqrt{\mathop{\mathbb{E}}_{S \sim D^n} \left[\sum_{t=1}^{T} \frac{\eta_t^2}{\sigma_t^2} g_e(t) \right]}; \qquad \text{(empirical norm)}$$

where $g_e(t) = \mathbb{E}_{w \sim W_{t-1}} [\frac{1}{n} \sum_{i=1}^{n} \|\nabla F(w, z_i)\|_2^2]$ is the empirical squared gradient norm, and W_t is the parameter at step t of SGLD.

Furthermore, based on essentially the same proof, we can obtain the following bound that depends on the *population* gradient norm:

$$\text{err}_{\text{gen}} \quad \frac{8\sqrt{2}C}{n} \sqrt{\mathop{\mathbb{E}}_{S' \sim D^n} \left[\sum_{t=1}^{T} \frac{\eta_t^2}{\sigma_t^2} \mathop{\mathbb{E}}_{w \sim W'_{t-1}} \mathop{\mathbb{E}}_{z \sim D} \|\nabla F(w, z)\|_2^2 \right]}.$$

The full proofs of the above results are postponed to Appendix A, and we provide some remarks about the new bounds.

Remark 12. In fact, our proof establishes that the above upper bound holds for the two sequences W_T and W'_T: $\text{KL}(W_T; W'_T) \le \frac{8\sqrt{2}}{n^2} \sum_{t=1}^{T} \frac{\eta_t^2}{\sigma_t^2} g_e(t)$. Hence, our bound holds for any sufficiently regular function over the parameter sequences: $\text{KL}(f(W_T); f(W'_T)) \le \frac{8\sqrt{2}}{n^2} \sum_{t=1}^{T} \frac{\eta_t^2}{\sigma_t^2} g_e(t)$. In particular, our generalization error bound automatically extends to several variants of SGLD, such as outputting the average of the trajectory, the average of the suffix of certain length, or the exponential moving average.

Remark 13 (High-probability bounds). By relaxing the expected squared gradient norm term to n^2 and using the uniform stability framework, our proof can be adapted to recover the $\tilde{O}(C\sqrt{T}/n)$ bound in (Mou et al., 2018, Theorem 1). Then, we can apply the recent results of Feldman & Vondrak (2019) to provide a generalization error bound of $\tilde{O}(LC\sqrt{T}/n + 1/\sqrt{n})$ that holds with high probability. (Here $\tilde{O}$ hides poly-logarithmic factors.) When T is at least linear in n, the additional $1/\sqrt{n}$ term is not dominating.

3.1 Experiment

Distinguish random from normal. Inspired by Zhang et al. (2017a), we run both GLD (Figure 1) and SGLD (Appendix C.2) to fit both normal data and randomly labeled data (see Appendix C for more experiment details). As shown in Figure 1 and Figure 3 in Appendix C.2, a larger random label portion p leads to both much higher generalization error and much larger generalization error bound. Moreover, the shapes of the curves of our bounds look quite similar to those of the generalization error curves.

Note that in (b) and (c) of Figure 1, the scales in the y-axis are different. We list some possible reasons that may explain why our bound is larger than the actual generalization error. (1) as we explained in Remark 12, our bounds (Theorem 9 and 11) hold for any trajectory-based output, and are much stronger than upper bounds for the last point on the trajectory. (2) The constant we can prove in Lemma 21 may not be very tight. (3) The variance of Gaussian noise σ_t^2 is not large enough in our experiment. However, if we choose a larger variance, fitting the random labeled training data becomes quite slow. Hence, we use a small data size $n = 1(10000)$ for the above reason. We also run an extra experiment for GLD on the full MNIST dataset ($n = 60000$) without label corruption (see Figure 2 in the Appendix C). We can see that our bound is non-vacuous (since GLD—which computes the full gradients—took a long time to converge, we stopped when we achieved 90% training accuracy).[4]

[4] We highlight another difficulty in proving non-vacuous generalization error bounds when the data are randomly labeled. Consider a 10-class classification setting where all the labels are random. For any sufficiently small data size, there is always a deep neural network that perfectly fits the dataset. Thus, the training error is zero while the population error is 90%. In this case, any valid generalization error bound should be larger than 0.9. Then, the theoretical bound would still be vacuous even if it is only loose by a factor of 2.

(a) (b) (c) (d)

Figure 1: Training MLP with GLD ($\epsilon_t = 0.2/\sqrt{t}$) on a smaller version of MNIST with different random label portion p. (a) shows the training accuracy. (b) shows the generalization error, i.e., the gap between the 0/1 loss[01] on the training data and on the test data. (c) plots our bound in Theorem 9. (d) shows that for $p = 0$, the gradient norms become much smaller at later stages of training.

Relax the step size constraint. The condition on the step size in Theorem 11 may seem restrictive in the practical use[5]. We provide several ways to relax this constraint:

1. The proof of Theorem 11 still goes through if we replace c with $\max_{i \in [n]} \|\nabla F(W_{t-1}; z_i)\|_2$ in the constraint.
2. The maximum gradient norm can be controlled by gradient clipping, i.e., multiplying $\frac{\min(c, \|\nabla F(W_{t-1}; z_i)\|_2)}{\|\nabla F(W_{t-1}; z_i)\|_2}$ to each $\nabla F(W_{t-1}; z_i)$.
3. Replacing the constant 20 with 2 in this constraint will only increase the constant of our bound from 8.12 to 84.4.

We also provide an experiment combining the above ideas to make our Theorem 11 applicable in the practical use (see Figure 4 in Appendix C).

4 Generalization of CLD and GLD with ℓ_2 regularization

In this section, we study the generalization error of Continuous Langevin Dynamics (CLD) with ℓ_2 regularization. Throughout this section, we assume that the objective function over training set S is defined as $F(w; S) = F_0(w; S) + \frac{\lambda}{2}\|w\|_2^2$, and moreover, the following assumption holds.

Assumption 14. The loss function ℓ and the original objective F_0 are C-bounded. Moreover, F_0 is differentiable and L-lipschitz.

The Continuous Langevin Dynamics is defined by the following SDE:

$$dW_t = -\nabla F(W_t; S)\, dt + \sqrt{\tfrac{2}{\beta}}\, dB_t; \quad W_0 \sim \mu_0; \tag{CLD}$$

where $(B_t)_{t \geq 0}$ is the standard Brownian motion on $\mathbb{R}^d$ and the initial distribution μ_0 is the centered Gaussian distribution in $\mathbb{R}^d$ with covariance $\frac{1}{\lambda} I_d$. We show that the generalization error of CLD is upper bounded by $O\big(e^{4C}\, n^{-1} \sqrt{\beta/T}\big)$, which is independent of the training time T (Theorem 15). Furthermore, as T goes to infinity, we have a tighter generalization error bound $O\big(C^2 n^{-1}\big)$ (Theorem 39 in Appendix B). We also study the generalization of Gradient Langevin Dynamics (GLD), which is the discretization of CLD:

$$W_{k+1} = W_k - \epsilon \nabla F(W_k; S) + \sqrt{\tfrac{2\epsilon}{\beta}}\, \xi_k; \tag{GLD}$$

where ξ_k is the standard Gaussian random vector in $\mathbb{R}^d$. By leveraging a result developed in Raginsky et al. (2017), we show that, as ϵ^2 tends to zero, GLD has the same generalization as CLD (see Theorems 15 and 39). We first formally state our first main result in this section.

[5] The condition $\epsilon_t = O(\alpha_t/L)$ is also required in (Mou et al., 2018, Theorem 1)

Theorem 15. Under Assumption 14, CLD (with initial probability measure $d\nu_0 = \frac{1}{Z} e^{-\frac{\|w\|^2}{2}} dw$) has the following expected generalization error bound:

$$\mathrm{err}_{\mathrm{gen}} \quad \frac{2e^{4\ C}\ CL}{n} \sqrt{1 \quad \exp \quad \frac{T}{e^{8\ C}}} : \tag{7}$$

In addition, if L is M-smooth and non-negative, by setting > 2, > 0 and $2\ (0; 1 \wedge \frac{}{8M^2}$, GLD (running K iterations with the same $_0$ as CLD) has the expected generalization error bound:

$$\mathrm{err}_{\mathrm{gen}} \quad 2C \sqrt{\frac{p}{2KC_1^{\ 2}} + \frac{2CLe^{4\ C}}{n} \quad 1 \quad \exp \quad \frac{K}{e^{8\ C}}} \quad ; \tag{8}$$

where C_1 is a constant that only depends on M, , , b, L and d.

The following lemma is crucial for establishing the above generalization bound for CLD. In particular, we need to establish a Log-Sobolev inequality for $_t$, the parameter distribution at time t, for every time step $t > 0$. In contrast, most known LSIs only characterize the stationary distribution of the Markov process. The proof of the lemma can be found in Appendix B.

Lemma 16. Under Assumption 14, let $_t$ be the probability measure of W_t in CLD (with $d\ _0 = \frac{1}{Z} e^{-\frac{\|w\|^2}{2}} dw$). Let be a probability measure that is absolutely continuous with respect to $_t$. Suppose $d\ _t = \ _t(w)\, dw$ and $d\ = \ (w)\, dw$. Then, it holds that

$$\mathrm{KL}(\ ;\ _t) \quad \frac{\exp(8\ C\ )}{2} \int_{R^d} r \log \frac{(w)}{_t(w)}^{\ 2}\ (w)\, dw:$$

We sketch the proof of Theorem 15, and the complete proof is relegated to Appendix B.

Proof Sketch of Theorem 15. Suppose S and S^0 are two neighboring datasets. Let $(W_t)_{t\ 0}$ and $(W_t^0)_{t\ 0}$ be the process of CLD running on S and S^0, respectively. Let $_t$ and $_t$ be the pdf of W_t^0 and W_t. Let $F_S(w)$ denote $F(w; S)$. We have

$$\begin{aligned}
\frac{d}{dt}\mathrm{KL}(\ _t;\ _t) &= \frac{1}{} \int_{R^d}\ _t\ r \log \frac{_t}{_t}^{\ 2} dw + \int_{R^d}\ _t \langle r \log \frac{_t}{_t};\ r\ F_S \quad r\ F_{S^0} \rangle\, dw \\
&\quad \frac{1}{2} \int_{R^d}\ _t\ r \log \frac{_t}{_t}^{\ 2} dw + \frac{}{2} \int_{R^d}\ _t\ \| r\ F_S \quad r\ F_{S^0}\|_2^2\, dw: \\
&\quad \frac{}{e^{8\ C}}\mathrm{KL}(\ _t;\ _t) + \frac{2\ L^{\ 2}}{n^2} \qquad \text{(Lemma 16)}
\end{aligned}$$

Solving this inequality gives $\mathrm{KL}(\ _t;\ _t)$ $\sqrt{\frac{1}{n^2} 2\ L^{\ 2} e^{8\ C}\ (1 \quad e^{\ t=e^{\ 8\ C}})}$. Hence the generalization error of CLD can be bounded by $2C \sqrt{\frac{1}{2}\mathrm{KL}(\ _T;\ _T)}$, which proves the rst part. The second part of the theorem follows from Lemma 36 in Appendix B. ■

Our second generalization bound for CLD (Theorem 39 in Appendix B) is

$$\mathrm{err}_{\mathrm{gen}} \quad \frac{8\ C^{\ 2}}{n} + 4C \exp \quad \frac{T}{e^{4\ C}} \quad \sqrt{C}:$$

The high level idea to prove this bound is very similar to that in Raginsky et al. (2017). We rst observe that the (stationary) Gibbs distribution has a small generalization error. Then, we bound the distance from $_t$ to . In our setting, we can use the Holley-Stroock perturbation lemma which allows us to bound the Logarithmic Sobolev constant, and we can thus bound the above distance easily.

5 Future Directions

In this paper, we prove new generalization bounds for a variety of noisy gradient-based methods. Our current techniques can only handle continuous noises for which we can bound the KL-divergence.

One future direction is to study the discrete noise introduced in SGD (in this case the KL-divergence may not be well de ned). For either SGLD or CLD, if the noise level is small (i.e., is large), it may take a long time for the diffusion process to reach the stable distribution. Hence, another interesting future direction is to consider the local behavior and generalization of the diffusion process in nite time through the techniques developed in the studies of metastability (see e.g., Bovier et al. (2005); Bovier & den Hollander (2006); Tzen et al. (2018)). In particular, the technique may be helpful for further improving the bounds in Theorems 15 and 39 (when is not very large).

6 ACKNOWLEDGEMENT

We would like to thank Liwei Wang for several helpful discussions during various stages of the work. The research is supported in part by the National Natural Science Foundation of China Grant 61822203, 61772297, 61632016, 61761146003, and the Zhongguancun Haihua Institute for Frontier Information Technology and Turing AI Institute of Nanjing.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6155–6166, 2019.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning (ICML)*, pp. 254–263, 2018.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an in nitely wide neural net. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8139–8148, 2019.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

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

Olivier Bousquet and Andr Elisseeff. Stability and generalization. *Journal of Machine Learning Research (JMLR)*, 2:499–526, 2002.

Anton Bovier and Frank den Hollander. Metastability: a potential theoretic approach. In *International Congress of Mathematicians*, volume 3, pp. 499–518. Eur. Math. Soc. Zrich, 2006.

Anton Bovier, Vronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes ii: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7 (1):69–99, 2005.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations (ICLR)*, 2017.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2933–2943, 2019.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent nds global minima of deep neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1675–1685, 2019.

John Duchi. Derivations for linear algebra and optimization. *Berkeley, California*, 3, 2007.

Gintare Karolina Dziugaite and Daniel Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of entropy-SGD and data-dependent priors. In International Conference on Machine Learning (ICML), pp. 1377–1386, 2018.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In Uncertainty in Arti cial Intelligence (UAI), 2017.

André Elisseeff, Massimiliano Pontil, et al. Leave-one-out error and stability of learning algorithms with applications. NATO science series sub series iii computer and systems sciences, 190:111–130, 2003.

Andre Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. Journal of Machine Learning Research (JMLR), 6(Jan):55–79, 2005.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In Conference on Learning Theory (COLT), pp. 1270–1279, 2019.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: stability of stochastic gradient descent. In International Conference on Machine Learning (ICML), pp. 1225–1234, 2016.

Richard Holley and Daniel Stroock. Logarithmic sobolev inequalities and stochastic ising models. Journal of statistical physics, 46(5):1159–1194, 1987.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classi cation with deep convolutional neural networks. In Advances in Neural Information Processing Systems (NeurIPS), pp. 1097–1105, 2012.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In International Conference on Machine Learning (ICML), pp. 2815–2824, 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter pac-bayes bounds through distribution-dependent priors. Theoretical Computer Science, 473:4–28, 2013.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In International Conference on Arti cial Intelligence and Statistics (AISTATS), pp. 888–896, 2019.

Ben London. Generalization bounds for randomized learning with application to stochastic gradient descent. In NIPS Workshop on Optimizing the Optimizers, 2016.

Ben London. A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. In Advances in Neural Information Processing Systems (NeurIPS), pp. 2931–2940, 2017.

Georg Menz, André Schlichting, et al. Poincaré and logarithmic sobolev inequalities by decomposition of the energy landscape. The Annals of Probability, 42(5):1809–1884, 2014.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In Conference on Learning Theory (COLT), pp. 605–638, 2018.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $O(1=k^2)$. In Dokl. Akad. Nauk SSSR, volume 269, pp. 543–547, 1983.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Conference on Learning Theory (COLT), pp. 1376–1401, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In International Conference on Learning Representations (ICLR), 2018.

Grigorios A Pavliotis. Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations, volume 60. Springer, 2014.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In International Symposium on Information Theory (ISIT), pp. 546–550, 2018.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 4(5):1–17, 1964.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In Conference on Learning Theory (COLT), pp. 1674–1703, 2017.

Hannes Risken. Fokker-planck equation. The Fokker-Planck Equation, pp. 63–95. Springer, 1996.

Omar Rivasplata, Csaba Szepesvári, John S Shawe-Taylor, Emilio Parrado-Hernandez, and Shiliang Sun. Pac-bayes bounds for stable algorithms with instance-dependent priors. In Advances in Neural Information Processing Systems (NeurIPS), pp. 9214–9224, 2018.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In International Conference on Machine Learning (ICML), pp. 1139–1147, 2013.

Flemming Topsoe. Some inequalities for information divergence and related measures of discrimination. IEEE Transactions on Information Theory, 46(4):1602–1609, 2000.

Belinda Tzen, Tengyuan Liang, and Maxim Raginsky. Local optimality and generalization guarantees for the langevin algorithm via empirical metastability. In Conference On Learning Theory (COLT), pp. 857–875, 2018.

Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. In Advances in Neural Information Processing Systems (NeurIPS), pp. 9722–9733, 2019.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In Advances in Neural Information Processing Systems (NeurIPS), pp. 9709–9721, 2019.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In International Conference on Machine Learning (ICML), pp. 681–688, 2011.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In Advances in Neural Information Processing Systems (NeurIPS), pp. 2524–2533, 2017.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In Advances in Neural Information Processing Systems (NeurIPS), pp. 3126–3137, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In International Conference on Learning Representations (ICLR), 2017a.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In Conference on Learning Theory (COLT), pp. 1980–2022, 2017b.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. arXiv preprint arXiv:1811.08888, 2018.

A PROOFS IN SECTION 3

A.1 BAYES-STABILITY FRAMEWORK

Lemma 17. Under Assumption 6, for any prior distribution P not depending on the dataset $S = (z_1, \dots, z_n)$, the generalization error is upper bounded by

$$\mathop{\mathbb{E}}_{z} \mathop{\mathbb{E}}_{w \sim P} L(w; z) - \mathop{\mathbb{E}}_{w \sim Q_z} L(w; z) + \mathop{\mathbb{E}}_{z} \mathop{\mathbb{E}}_{w \sim P} L(w) - \mathop{\mathbb{E}}_{w \sim Q_z} L(w) \Big|,$$

where $L(w)$ denotes the population loss $L(w; D)$.

Proof of Lemma 17 Let $\mathrm{err}_{\mathrm{train}} = \mathbb{E}_S \mathbb{E}_{w \sim Q_S} L(w; S)$ and $\mathrm{err}_{\mathrm{test}} = \mathbb{E}_S \mathbb{E}_{w \sim Q_S} L(w)$. We can rewrite generalization error as $\mathrm{err}_{\mathrm{gen}} = \mathrm{err}_{\mathrm{test}} - \mathrm{err}_{\mathrm{train}}$, where

$$\begin{aligned}
\mathrm{err}_{\mathrm{test}} &= \mathop{\mathbb{E}}_{z} \mathop{\mathbb{E}}_{w} \mathop{\mathbb{E}}_{Q(1;z)} L(w) = \mathop{\mathbb{E}}_{z} \mathop{\mathbb{E}}_{w} \mathop{\mathbb{E}}_{Q_z} L(w) && \text{(Assumption 6)} \\
&= \mathop{\mathbb{E}}_{z} \int_{\mathbb{R}^d} (Q_z(w) - P(w)) L(w)\, dw + \int_{\mathbb{R}^d} P(w) L(w)\, dw.
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{err}_{\mathrm{train}} &= \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{w} \mathop{\mathbb{E}}_{Q_S} L(w; z_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{z} \mathop{\mathbb{E}}_{w} \mathop{\mathbb{E}}_{Q(i;z)} L(w; z) = \mathop{\mathbb{E}}_{z} \mathop{\mathbb{E}}_{w} \mathop{\mathbb{E}}_{Q_z} L(w; z) && \text{(Assumption 6)} \\
&= \mathop{\mathbb{E}}_{z} \int_{\mathbb{R}^d} (Q_z(w) - P(w)) L(w; z)\, dw + \int_{\mathbb{R}^d} P(w) \mathop{\mathbb{E}}_{z} L(w; z)\, dw && (P \text{ is a prior}) \\
&= \mathop{\mathbb{E}}_{z} \int_{\mathbb{R}^d} (Q_z(w) - P(w)) L(w; z)\, dw + \int_{\mathbb{R}^d} P(w) L(w)\, dw. && \text{(definition of } f(w))
\end{aligned}$$

Thus, we have

$$\begin{aligned}
|\mathrm{err}_{\mathrm{gen}}| &= |\mathrm{err}_{\mathrm{test}} - \mathrm{err}_{\mathrm{train}}| \\
&= \mathop{\mathbb{E}}_{z} \int_{\mathbb{R}^d} (Q_z(w) - P(w)) L(w)\, dw - \mathop{\mathbb{E}}_{z} \int_{\mathbb{R}^d} (Q_z(w) - P(w)) L(w; z)\, dw \\
&\le \Big| \mathop{\mathbb{E}}_{z} \mathop{\mathbb{E}}_{w \sim Q_z} L(w; z) - \mathop{\mathbb{E}}_{w \sim P} L(w; z) \Big| + \Big| \mathop{\mathbb{E}}_{z} \mathop{\mathbb{E}}_{w \sim Q_z} L(w) - \mathop{\mathbb{E}}_{w \sim P} L(w) \Big|.
\end{aligned}$$

■

Now we are ready to prove Theorem 7, which we restate in the following.

Theorem 7 (Bayes-Stability). Suppose the loss function $L(w; z)$ is C-bounded and the learning algorithm is order-independent (Assumption 6), then for any prior distribution P not depending on S, the generalization error is bounded by both $2C \mathop{\mathbb{E}}_{z} \sqrt{\frac{1}{2n} \mathrm{KL}(P; Q_z)}$ and $2C \mathop{\mathbb{E}}_{z} \sqrt{\frac{1}{2n} \mathrm{KL}(Q_z; P)}$.

Proof By Lemma 17,

$$\begin{aligned}
|\mathrm{err}_{\mathrm{gen}}| &\le \Big| \mathop{\mathbb{E}}_{z} \mathop{\mathbb{E}}_{w \sim P} L(w; z) - \mathop{\mathbb{E}}_{w \sim Q_z} L(w; z) \Big| + \Big| \mathop{\mathbb{E}}_{z} \mathop{\mathbb{E}}_{w \sim P} L(w) - \mathop{\mathbb{E}}_{w \sim Q_z} L(w) \Big| \\
&\le \mathop{\mathbb{E}}_{z} [2C \cdot \mathrm{TV}(P; Q_z) + 2C \cdot \mathrm{TV}(P; Q_z)] && (C\text{-boundedness}) \\
&\le 4C \mathop{\mathbb{E}}_{z} \left[\sqrt{\frac{1}{2} \mathrm{KL}(P; Q_z)} \right] && (\text{Pinsker's inequality})
\end{aligned}$$

The other bound follows from a similar argument. ■

A.2 TECHNICAL LEMMAS

Now we turn to the proof of Theorem 11. The following lemma allows us to reduce the proof of algorithmic stability to the analysis of a single update step.

Lemma 10. Let $(W_0, \dots, W_T)$ and $(W_0', \dots, W_T')$ be two independent sequences of random variables such that for each $t \in \{0, \dots, T\}$, W_t and W_t' have the same support. Suppose W_0 and W_0' follow the same distribution. Then,

$$KL(W_{\le T} \,;\, W'_{\le T}) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w_{<t}} \left[KL(W_t | W_{<t} = w_{<t} \,;\, W_t' | W'_{<t} = w_{<t}) \right],$$

where $W_{\le t}$ denotes $(W_0, \dots, W_t)$ and $W_{<t}$ denotes $W_{\le t-1}$.

Proof By the chain rule of the KL-divergence,

$$KL(W_{\le t} \,;\, W'_{\le t}) = KL(W_{<t} \,;\, W'_{<t}) + \mathop{\mathbb{E}}_{w_{<t}} \left[KL(W_t | W_{<t} = w_{<t} \,;\, W_t' | W'_{<t} = w_{<t}) \right].$$

The lemma follows from a summation over $t = 1, \dots, T$. ■

The following lemma (see e.g., (Duchi, 2007, Section 9)) gives a closed-form formula for the KL-divergence between two Gaussian distributions.

Lemma 18. Suppose that $P = N(\mu_1, \Sigma_1)$ and $Q = N(\mu_2, \Sigma_2)$ are two Gaussian distributions on $\mathbb{R}^d$. Then,

$$KL(P;Q) = \frac{1}{2} \left[\operatorname{tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1} (\mu_2 - \mu_1) - d + \ln \frac{\det(\Sigma_2)}{\det(\Sigma_1)} \right].$$

The following lemma (Topsoe, 2000, Theorem 3) helps us to upper bound the KL-divergence.

Definition 19. Let P and Q be two probability distributions on $\mathbb{R}^d$. The directional triangular discrimination from P to Q is defined as

$$\Delta(P;Q) = \sum_{k=0}^{\infty} 2^k \, \Delta\left(2^{-k} P + (1 - 2^{-k}) Q \,;\, Q\right),$$

where

$$\Delta(P;Q) = \int_{\mathbb{R}^d} \frac{(P(w) - Q(w))^2}{P(w) + Q(w)} \, dw.$$

Lemma 20. For any two probability distributions P and Q on $\mathbb{R}^d$,

$$KL(P;Q) \le \ln 2 \cdot \Delta(P;Q).$$

Recall that $\mathcal{G}$ is the set of all possible mini-batches, $\mathcal{G}_n = \{B \in \mathcal{G} : n \in B\}$ denotes the collection of mini-batches that contain n, while $\overline{\mathcal{G}_n} = \mathcal{G} \setminus \mathcal{G}_n$. $\operatorname{diam}(A) = \sup_{x,y \in A} \|x - y\|$ denotes the diameter of set A. The following technical lemma upper bounds the KL-divergence between two Gaussian mixtures induced by sampling a mini-batch from neighbouring datasets.

Lemma 21. Suppose that batch size $b = n/2$. $\{\mu_B : B \in \mathcal{G}\}$ and $\{\mu'_B : B \in \mathcal{G}\}$ are two collections of points in $\mathbb{R}^d$ labeled by mini-batches of size b that satisfy the following conditions for some constant $\Delta \in [0, \infty)$:

1. $\|\mu_B - \mu'_B\| \le \Delta$ for $B \in \mathcal{G}_n$ and $\mu_B = \mu'_B$ for $B \in \overline{\mathcal{G}_n}$.
2. $\operatorname{diam}(\{\mu_B : B \in \mathcal{G}\} \cup \{\mu'_B : B \in \mathcal{G}\}) \le 1$.

Let $p_{\mu,\sigma}$ denote the Gaussian distribution $N(\mu, \frac{\sigma^2}{2} I_d)$. Let $P = \frac{1}{|\mathcal{G}|} \sum_{B \in \mathcal{G}} p_{\mu_B,\sigma}$ and $P' = \frac{1}{|\mathcal{G}|} \sum_{B \in \mathcal{G}} p_{\mu'_B,\sigma}$ be two mixture distributions over all mini-batches. Then,

$$KL(P;P') \le \frac{8.23 b^2 \Delta^2}{\sigma^2 n^2}.$$

Proof of Lemma 21 By Lemma 20, $\mathrm{KL}(P; P^0)$ is bounded by

$$\ln 2 \quad (P; P^0) = \ln 2 \sum_{k=0}^{1} 2^k \quad 2^{\ k} P + (1 \quad 2^{\ k}) P^0; P^0$$
$$= \ln 2 \sum_{k=0}^{1} 2^k \int_{R^d} \frac{4^{\ k}(P(w) \quad P^0(w))^2}{2^{\ k} P(w) + (2 \quad 2^{\ k}) P^0(w)} \, dw:$$

The numerator of the above integrand is upper bounded by

$$4^{\ k}(P \quad P^0)^2 = 4^{\ k} \left(\frac{1}{jGj} \sum_{B 2 G} (p_{B;} \quad p_{\frac{0}{B};}) \right)^2$$
$$= \frac{4^{\ k} jG_n j^2}{jGj^2} \left(\frac{1}{jG_n j} \sum_{B 2 G_n} (p_{B;} \quad p_{\frac{0}{B};}) \right)^2 \qquad (9)$$
$$\frac{4^{\ k} b^2}{n^2} \left(\frac{1}{jG_n j} \sum_{B 2 G_n} (p_{B;} \quad p_{\frac{0}{B};}) \right)^2;$$

while the denominator can be lower bounded as follows:

$$2^{\ k} P + (2 \quad 2^{\ k}) P^0 \quad \frac{2^{\ k}}{jGj} \sum_{B 2 \overline{G_n}} p_{B;} + \frac{2 \quad 2^{\ k}}{jGj} \sum_{B 2 \overline{G_n}} p_{\frac{0}{B};}$$
$$= \frac{2}{jGj} \sum_{B 2 \overline{G_n}} p_{B;} \qquad (\ _B = \ _{\frac{0}{B}} \text{ for } B 2 \overline{G_n})$$
$$= \frac{1}{j\overline{G_n}j} \quad \frac{2(n \quad b)}{n} \sum_{B 2 \overline{G_n}} p_{B;}$$
$$\frac{1}{j\overline{G_n}j} \sum_{B 2 \overline{G_n}} p_{B;}; \qquad (b \quad n=2)$$

which implies, by the convexity of $1=x$, that

$$\frac{1}{2^{\ k} P + (2 \quad 2^{\ k}) P^0} \quad \frac{1}{\frac{1}{j\overline{G_n}j} \sum_{B 2 \overline{G_n}} p_{B;}} \quad \frac{1}{j\overline{G_n}j} \sum_{B 2 \overline{G_n}} \frac{1}{p_{B;}}: \qquad (10)$$

Inequalities (9) and (10) together imply

$$2^{\ k} P + (1 \quad 2^{\ k}) P^0; P^0 \quad \frac{4^{\ k} b^2}{n^2 j\overline{G_n}j j G_n j} \sum_{A 2 \overline{G_n}} \sum_{B 2 G_n} \int_{R^d} \frac{(p_{B;}(w) \quad p_{\frac{0}{B};}(w))^2}{p_{A;}(w)} \, dw: \qquad (11)$$

Now we bound the right-hand side of (11) for $x e d$ and B. By applying a translation and a rotation, we can assume without loss of generality that $_A = 0$, and the last $d \quad 2$ coordinates of $_B$ and $_{\frac{0}{B}}$ are all zero. Note that the integral is unchanged when we project the space to the two-dimensional subspace corresponding to the rst two coordinates. Thus, it suf ces to prove a bound for $d = 2$. We rewrite (11) as

$$(2^{\ k} P + (1 \quad 2^{\ k}) P^0; P^0) \quad \frac{4^{\ k} b^2}{n^2 j\overline{G_n}j j G_n j} \sum_{A 2 \overline{G_n}} \sum_{B 2 G_n} \frac{1}{2} \int_{R^2} \frac{e^{\ k \frac{w \quad B}{}k^2} \quad e^{\ \frac{w \quad \frac{0}{B}}{}^2}}{e^{\ k \frac{w}{}k^2}} \, dw: \qquad (12)$$

Let I be the integral in the right-hand side of (12). Note that $\frac{\sigma^2}{B} \le 0.1$ and $\frac{\sigma^2}{B} \le 1$. Let $\gamma = \frac{\sigma}{\sqrt{B}}$ and $r = \frac{w}{\gamma}$. Our goal is to bound $\max_{\gamma^2 \in [0,0.1]} (I/\gamma^2)$. Let $(x)^+ = \max(x, 0)$. Since

$$I \le \gamma^2 \int_0^{\infty} \frac{\max_{y \in [(r-0.1)^+, r+0.1]} (e^{-y^2/2} - e^{-(y+\gamma)^2/2})^2}{e^{-r^2/2}} 2\pi r \, dr.$$

We have

$$\max_{\gamma^2 \in [0,0.1]} \frac{I}{\gamma^2} \le \int_0^{\infty} e^{r^2/2} 2\pi r \max_{y \in [(r-0.1)^+, r+0.1]} \max_{\gamma^2 \in [0,0.1]} \frac{(e^{-y^2/2} - e^{-(y+\gamma)^2/2})^2}{\gamma^2} dr.$$

Let $\phi(y, \gamma) = (\frac{e^{-y^2/2} - e^{-(y+\gamma)^2/2}}{\gamma})^2$, we make two claims which we will prove later:

1. For all $y, \gamma \ge 0$, $\phi(y, \gamma) \le \frac{2}{e}$.
2. For all $y \ge \frac{1}{\sqrt{2}}$, $\phi(y, \gamma)$ is non-increasing in γ.

The above claims imply that:

1. For any $r \in [0, \frac{1}{\sqrt{2}} + 0.1]$, $\max_{y \in [(r-0.1)^+, r+0.1], \gamma^2 \in [0,0.1]} [\phi(y, \gamma)] \le \frac{2}{e}$.
2. For any $r \in [\frac{1}{\sqrt{2}} + 0.1, +\infty)$, we have

$$\begin{aligned} \max_{y \in [(r-0.1)^+, r+0.1], \gamma^2 \in [0,0.1]} \phi(y, \gamma) &\le \max_{y \in [(r-0.1)^+, r+0.1]} \lim_{\gamma \to 0} [\phi(y, \gamma)] \\ &= \max_{y \in [(r-0.1)^+, r+0.1]} 4y^2 e^{-2y^2} \\ &= 4(r - 0.1)^2 e^{-2(r-0.1)^2}. \end{aligned}$$

The last step holds since $\forall y^2 e^{-2y^2}$ is decreasing on $[\frac{1}{\sqrt{2}}, +\infty)$.

Thus we have

$$\max_{\gamma^2 \in [0,0.1]} \frac{I}{\gamma^2} \le \int_0^{\frac{1}{\sqrt{2}}+0.1} e^{r^2/2} 2\pi r \cdot \frac{2}{e} dr + \int_{\frac{1}{\sqrt{2}}+0.1}^{+\infty} e^{r^2/2} 2\pi r \cdot 4(r-0.1)^2 e^{-2(r-0.1)^2} dr \le 18.6487\pi^2.$$

Plugging the above into (12) gives

$$(2^k P + (1 - 2^k) P^0, P^0) \le \frac{4^k b^2}{n^2 \sigma^2} \cdot \sigma^2 \max_{\gamma^2 \in [0,0.1]} (I/\gamma^2) \le \frac{4^k b^2}{n^2} \cdot 18.6487\pi^2.$$

We conclude that

$$\begin{aligned} KL(P; P^0) &\le \ln 2 \sum_{k=0}^{\infty} 2^{-k} (2^k P + (1 - 2^k) P^0, P^0) \\ &\le \frac{37.2974 b^2 \sigma^2 \ln 2}{n^2} \le \frac{8.23 b^2 \sigma^2}{n^2}. \end{aligned}$$

Finally, we prove the two claims used above:

1. For all $y, \gamma \ge 0$, let $h(x) = e^{-x^2/2}$, we have $e^{-y^2/2} - e^{-(y+\gamma)^2/2} = \int_y^{y+\gamma} h'(t) \, dt$. Since $|h'(t)| = |e^{-t^2/2}(-2t)| = e^{-t^2/2}(2t)$. Let $\frac{\partial}{\partial t}|h'(t)| = 0$, we have $t = 1/\sqrt{2}$. Thus, $|h'(t)| \le e^{-1/2}\sqrt{2}$ and $e^{-y^2/2} - e^{-(y+\gamma)^2/2} \le \gamma\sqrt{2}/e$.
2. Suppose $y \ge 1/\sqrt{2}$, we have

$$\begin{aligned} \frac{\partial}{\partial \gamma} \left(\frac{e^{-y^2/2} - e^{-(y+\gamma)^2/2}}{\gamma} \right) &= \frac{e^{-(y+\gamma)^2/2}(2y\gamma + 2\gamma^2 + 1) - e^{-y^2/2}}{\gamma^2} \\ &= \frac{e^{-y^2/2}}{\gamma^2} [e^{-2y\gamma - \gamma^2}(2y\gamma + 2\gamma^2 + 1) - 1]. \end{aligned}$$

Let $g(y, \gamma) = e^{-2y\gamma - \gamma^2}(2y\gamma + 2\gamma^2 + 1) - 1$. Note that

(a) $\lim_{\to 0} \frac{\partial}{\partial}\left(\frac{e^{-y^2} - e^{-(y+\ )^2}}{}\right) = 0$.

(b) $\frac{\partial}{\partial} g(1/\sqrt{2};\ ) = -4e^{-(\ +\sqrt{2})^2}(\ +\sqrt{2}) < 0$ and $g(1/\sqrt{2}; 0) = 0$.

It implies that $\frac{\partial}{\partial}\left(\frac{e^{-(1/\sqrt{2})^2} - e^{-(1/\sqrt{2}+\ )^2}}{}\right)$ 0 for > 0. Since $\frac{\partial}{\partial y} g(y;\ ) = 4\ ^2 e^{-(\ +2y)}(\ + y) < 0$ for y 0, we conclude that for any y $1/\sqrt{2}$:

$$\frac{\partial}{\partial}\left(\frac{e^{-y^2} - e^{-(y+\ )^2}}{}\right)^{!} \quad \frac{\partial}{\partial}\left(\frac{e^{-(1/\sqrt{2})^2} - e^{-(1/\sqrt{2}+\ )^2}}{}\right)^{!} \quad 0.$$

■

A.3 Main Theorem

Recall that SGLD on dataset S is defined as

$$W_t \quad W_{t\ 1} \quad {}_t \nabla_w F(W_{t\ 1}; S_{B_t}) + \sqrt{\frac{t}{2}} N(0; I_d).$$

Here ${}_t$ is the step size. $B_t = \{i_1; \ldots; i_b\}$ is a subset of $\{1; \ldots; n\}$ of size b, and $S_{B_t} = (z_{i_1}; \ldots; z_{i_b})$ is the mini-batch indexed by B_t. Recall that $F(w; S)$ denotes $\frac{1}{|S|}\sum_{i=1}^{|S|} F(w; z_i)$. We restate and prove Theorem 11 in the following.

Theorem 11. Suppose that the loss function is C-bounded and the objective function F is L-lipschitz. Assume that the following conditions hold:

1. Batch size b $n/2$.
2. Learning rate ${}_t$ ${}_t \le 1/(2L)$.

Then, the following expected generalization error bound holds for T iterations of SGLD (1):

$$\text{err}_{\text{gen}} \quad \frac{8\sqrt{2}C}{n}\sqrt{\mathop{\mathbb{E}}_{S\,D^n}\left[\sum_{t=1}^{T} \frac{{}_t^2}{2}\, g_e(t)\right]}; \qquad \text{(empirical norm)}$$

where $g_e(t) = \mathbb{E}_{w\ W_{t\ 1}}[\frac{1}{n}\sum_{i=1}^{n} \|\nabla F(w; z_i)\|_2^2]$ is the empirical squared gradient norm, and W_t is the parameter at step t of SGLD.

Proof By Theorem 7, we have

$$\text{err}_{\text{gen}} \quad 2C \mathop{\mathbb{E}}_{\bar{S}} \sqrt{2\text{KL}(Q_{\bar{S}}; P)} \tag{13}$$

for any prior distribution P. In particular, we define the prior as $P(w) = \mathbb{E}_{\bar{S}\,D^{n\ 1}}[P_{\bar{S}}(w)]$, where $P_{\bar{S}}(w) = Q_{(\bar{S};0)}$. By the convexity of KL-divergence,

$$\text{KL}(Q_{\bar{S}}; P) = \text{KL}\left(\mathop{\mathbb{E}}_{\bar{S}}[Q_{(\bar{S};\bar{S})}]; \mathop{\mathbb{E}}_{\bar{S}}[Q_{(\bar{S};0)}]\right) \quad \mathop{\mathbb{E}}_{\bar{S}}\left[\text{KL}\left(Q_{(\bar{S};\bar{S})}; Q_{(\bar{S};0)}\right)\right]. \tag{14}$$

Fix a data point $z \in \mathcal{Z}$. Let $(W_t)_{t\ 0}$ and $(W_t^0)_{t\ 0}$ be the training process of SGLD for $S = (\bar{S}; z)$ and $S^0 = (\bar{S}; 0)$, respectively. Fix a time step t and $w_{<t} = (w_0; \ldots; w_{t\ 1})$. Let P_t and P_t^0 denote the distribution of W_t and W_t^0 conditioned on $W_{<t} = w_{<t}$ and $W_{<t}^0 = w_{<t}$, respectively. By the definition of SGLD, we have $P_t = \frac{1}{|G|}\sum_{B \in G} p_{\ B}$ and $P_t^0 = \frac{1}{|G|}\sum_{B \in G} p_{\ \frac{0}{B}}$, where $_B = w_{t\ 1} \quad {}_t \nabla_w F(w_{t\ 1}; S_B)$, $\frac{0}{B} = w_{t\ 1} \quad {}_t \nabla_w F(w_{t\ 1}; S_B^0)$, and p denotes the Gaussian distribution $N(\ ; \frac{{}_t^2}{2} I_d)$. We note that:

1. $\left\| \frac{0}{B} \quad {}_B \right\| \quad \frac{{}_t \|\nabla F(w_{t\ 1}; z)\|_2}{b}$ for $B \in G_n$ and $_B = \frac{0}{B}$ for $B \in \overline{G_n}$.

2. $\mathrm{diam}(\{\theta^0_B : B \in G\} \cup \{\theta_B : B \in G\}) \le 2\eta_t L \le \eta/10$.

By applying Lemma 21 with $\alpha = \frac{\eta_t \|\nabla F(w_{t-1};z)\|_2}{b}$ and $\beta = \eta_t$,

$$\mathrm{KL}(P_t; P_t^0) \le \frac{8\cdot 2^3 \eta_t^2 \|\nabla F(w_{t-1};z)\|_2^2}{\eta_t^2 n^2}.$$

By Lemma 10,

$$\mathrm{KL}(W_T; W^0{}_T) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w_{<t}} \mathop{\mathbb{E}}_{W_{<t}} \left[\mathrm{KL}(P_t; P_t^0)\right] \le \sum_{t=1}^{T} \mathop{\mathbb{E}}_{w} \mathop{\mathbb{E}}_{W_{t-1}} \left[\frac{8\cdot 2^3 \eta_t^2 \|\nabla F(w;z)\|_2^2}{\eta_t^2 n^2}\right];$$

which implies that

$$\mathrm{KL}(Q_S; Q_{S'}) = \mathrm{KL}(W_T; W_T^0) \le \mathrm{KL}(W_T; W^0{}_T) \le \frac{8\cdot 2^3}{n^2} \sum_{t=1}^{T} \frac{\eta_t^2}{\eta_t^2} \mathop{\mathbb{E}}_{w} \mathop{\mathbb{E}}_{W_{t-1}} \left[\|\nabla F(w;z)\|_2^2\right].$$

Together with (13) and (14), we have

$$\begin{aligned}
\mathrm{err}_{\mathrm{gen}} &\le 2C \mathop{\mathbb{E}}_{z} \sqrt{2 \mathop{\mathbb{E}}_{S} \left[\frac{8\cdot 2^3}{n^2} \sum_{t=1}^{T} \frac{\eta_t^2}{\eta_t^2} \mathop{\mathbb{E}}_{w} \mathop{\mathbb{E}}_{W_{t-1}} \left[\|\nabla F(w;z)\|_2^2\right]\right]} \\
&\le 2C \sqrt{2 \mathop{\mathbb{E}}_{S} \left[\frac{8\cdot 2^3}{n^2} \sum_{t=1}^{T} \frac{\eta_t^2}{\eta_t^2} \mathop{\mathbb{E}}_{w} \mathop{\mathbb{E}}_{W_{t-1}} \left[\|\nabla F(w;z_n)\|_2^2\right]\right]} : \qquad \text{(concavity of } \sqrt{\cdot}\text{)}
\end{aligned}$$

Since SGLD is order-independent, we can replace $\nabla F(w;z_n)$ with $\nabla F(w;z_i)$ for any $i \in [n]$ in the right-hand side of the above bound. Our theorem then follows from the concavity of $\sqrt{\cdot}$. Furthermore, if we bound $\mathrm{KL}(P; Q_z)$ instead of $\mathrm{KL}(Q_z; P)$ in the above proof, we obtain the following bound that depends on the population squared gradient norm

$$\mathrm{err}_{\mathrm{gen}} \le \frac{8\cdot 12C}{n} \sqrt{\mathop{\mathbb{E}}_{S} \left[\sum_{t=1}^{T} \frac{\eta_t^2}{\eta_t^2} \mathop{\mathbb{E}}_{w} \mathop{\mathbb{E}}_{W_{t-1}^0} \mathop{\mathbb{E}}_{z} \mathop{\mathbb{E}}_{D} \|\nabla F(w;z)\|_2^2\right]}.$$

■

A.4 Extension to General Noises

We can extend the generalization bounds in previous sections, which require the noise to be Gaussian, to other general noises, namely the family of log-lipschitz noises.

Definition 22 (Log-Lipschitz Noises) A probability distribution on $\mathbb{R}^d$ with density p is L-log-lipschitz if and only if $\|\nabla \ln p(w)\| \le L$ holds for any $w \in \mathbb{R}^d$. A random variable ζ is called an L-log-lipschitz noise if and only if it is drawn from an L-log-lipschitz distribution.

The analog of SGLD, noisy momentum method (Definition 24), and noisy NAG (Definition 25) can be naturally defined by replacing the Gaussian noise ζ_t at each iteration with an independent L-log-lipschitz noise in the definition.

The following lemma is an analog of Lemma 21 under log-lipschitz noises. Recall that G denotes a collection of mini-batches of size b. Lemma 23 readily implies the analogs of Theorems 11, 26 and 27 under more general noise distributions.

Lemma 23. Suppose that batch size $n=2$ and N is an L_{noise}-log-lipschitz distribution on $\mathbb{R}^d$. $\{\mathcal{B} : \mathcal{B} \in G\}$ and $\{\mathcal{B}' : \mathcal{B} \in G\}$ are two collections of points in $\mathbb{R}^d$ that satisfy the following conditions for some constant $\Delta \in (0, \frac{1}{L_{noise}})$:

1. $\|\mathcal{B} - \mathcal{B}'\| \le \Delta$ for $\mathcal{B} \in G_n$ and $\mathcal{B} = \mathcal{B}'$ for $\mathcal{B} \in \overline{G_n}$.
2. $\mathrm{diam}(\{\mathcal{B} : \mathcal{B} \in G\} \cup \{\mathcal{B}' : \mathcal{B} \in G\}) \le 1$.

For $w \in \mathbb{R}^d$, let $p(w)$ denote the distribution of $w + \epsilon$ when ϵ is drawn from N. Let $P = \frac{1}{|G|^n} \sum_{\mathcal{B} \in G^n} p_{\mathcal{B}}$ and $P' = \frac{1}{|G|^n} \sum_{\mathcal{B} \in G^n} p_{\mathcal{B}'}$ be mixture distributions over all mini-batches. Then,

$$\mathrm{KL}(P; P') \le \frac{C_0 b^2 \Delta^2}{n^2}$$

for some constant C_0 that only depends on L_{noise}.

Proof of Lemma 23. Following the same argument as in the proof of Lemma 21, we have

$$\mathrm{KL}(P; P') \le \ln 2 \sum_{k=0}^{\infty} 2^k \mathrm{KL}(2^{-k} P + (1 - 2^{-k}) P'; P') \tag{15}$$

where

$$\mathrm{KL}(2^{-k} P + (1 - 2^{-k}) P'; P') \le \frac{4^{-k} b^2}{n^2 |\overline{G_n}| |G_n|} \sum_{A \in \overline{G_n}} \sum_{\mathcal{B} \in G_n} \int_{\mathbb{R}^d} \frac{(p_{\mathcal{B}}(w) - p_{\mathcal{B}'}(w))^2}{p_{\mathcal{A}}(w)} dw. \tag{16}$$

Fixed $\mathcal{A} \in \overline{G_n}$ and $\mathcal{B} \in G_n$. Let p_{noise} denote the density of the noise distribution N. Since $\|\mathcal{A} - \mathcal{B}\| \le 1$ and p_{noise} is L_{noise}-log-lipschitz, we have

$$p_{\mathcal{B}}(w) = p_{noise}(w - \mathcal{B}) \ge p_{noise}(w - \mathcal{A}) \, e^{-L_{noise} \|\mathcal{A} - \mathcal{B}\|} \ge e^{-L_{noise}} p_{\mathcal{A}}(w).$$

Similarly, since $\|\mathcal{B} - \mathcal{B}'\| \le \Delta$, we have

$$e^{-L_{noise}} p_{\mathcal{B}}(w) \le p_{\mathcal{B}'}(w) \le e^{L_{noise}} p_{\mathcal{B}}(w).$$

Then, it follows from $L_{noise} \ge 1$ that

$$(p_{\mathcal{B}}(w) - p_{\mathcal{B}'}(w))^2 \le (e^{L_{noise}} - 1)^2 p_{\mathcal{B}}(w)^2 \le 2 L_{noise}^2 \Delta^2 p_{\mathcal{B}}(w)^2.$$

Therefore, the integral on the right-hand side of (16) can be upper bounded as follows:

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{(p_{\mathcal{B}}(w) - p_{\mathcal{B}'}(w))^2}{p_{\mathcal{A}}(w)} dw &\le 2 L_{noise}^2 \Delta^2 \int_{\mathbb{R}^d} \frac{p_{\mathcal{B}}(w)^2}{p_{\mathcal{A}}(w)} dw \\ &\le 2 L_{noise}^2 \Delta^2 \int_{\mathbb{R}^d} p_{\mathcal{B}}(w) \, e^{L_{noise}} dw \\ &= 2 L_{noise}^2 \Delta^2 e^{L_{noise}}. \end{aligned}$$

Plugging the above inequality into (15) and (16) gives

$$\mathrm{KL}(2^{-k} P + (1 - 2^{-k}) P'; P') \le \frac{4^{-k} b^2}{n^2 |\overline{G_n}| |G_n|} \sum_{\mathcal{A} \in \overline{G_n}} \sum_{\mathcal{B} \in G_n} 2 L_{noise}^2 \Delta^2 e^{L_{noise}} = L_{noise}^2 \Delta^2 e^{L_{noise}} \frac{4^{-k} b^2 \Delta^2}{n^2}.$$

and

$$\mathrm{KL}(P; P') \le \ln 2 \sum_{k=0}^{\infty} 2^k L_{noise}^2 \Delta^2 e^{L_{noise}} \frac{4^{-k} b^2 \Delta^2}{n^2} = 2 \ln 2 \, L_{noise}^2 \Delta^2 e^{L_{noise}} \frac{b^2 \Delta^2}{n^2}.$$

■

A.5 EXTENSION TO OTHER GRADIENT-BASED METHODS

A.5.1 STABILITY BOUND FOR MOMENTUM AND NESTEROVS ACCELERATED GRADIENT

We adopt the formulation of Classical Momentum and Nesterov's Accelerated Gradient (NAG) methods in Sutskever et al. (2013) and consider the noisy versions of them.

Definition 24 (Noisy Momentum Method). Noisy Momentum Method on objective function $F(w;z)$ and dataset S is defined as

$$\begin{aligned} V_t &\leftarrow \gamma V_{t-1} - \alpha_t \nabla_w F(W_{t-1}; S_{B_t}) + \xi_t \\ W_t &\leftarrow W_{t-1} + V_t \end{aligned}$$

Definition 25 (Noisy Nesterovs Accelerated Gradient). Noisy Nesterovs Accelerated Gradient (NAG) on objective function $F(w;z)$ and dataset S is defined as

$$\begin{aligned} V_t &\leftarrow \gamma V_{t-1} - \alpha_t \nabla_w F(W_{t-1} + \gamma V_{t-1}; S_{B_t}) + \xi_t; \\ W_t &\leftarrow W_{t-1} + V_t. \end{aligned}$$

In both definitions, α_t is the step size, mini-batch B_t is drawn uniformly from G, ξ_t is a Gaussian noise drawn from $N(0; \frac{\alpha_t^2}{2} I_d)$, and $\gamma \in [0,1]$ is the momentum coefficient.

Theorem 26. Under the same assumptions on the loss function, objective function, batch size and learning rate as in Theorem 11, the generalization bounds in Theorem 11 still hold for noisy momentum method and noisy NAG.

Proof of Theorem 26. For any time step t and $w_{<t} = (w_0, w_1, \ldots, w_{t-1})$, let P_t and P_t^0 denote the distribution of W_t and W_t^0 conditioned on $W_{<t} = w_{<t}$ and $W_{<t}^0 = w_{<t}$, respectively. By definition, we have $P_t = \frac{1}{|G|} \sum_{B \in G} p_B$ and $P_t^0 = \frac{1}{|G|} \sum_{B \in G} p_B^0$.

If $t = 1$, for both noisy momentum method and noisy NAG, we have

$$\begin{aligned} \mu_B &= w_{t-1} - \alpha_t \nabla_w F(w_{t-1}; S_B); \\ \mu_B^0 &= w_{t-1} - \alpha_t \nabla_w F(w_{t-1}; S_B^0). \end{aligned}$$

For $t > 1$, if noisy momentum method is used, we have

$$\begin{aligned} \mu_B &= w_{t-1} + \gamma(w_{t-1} - w_{t-2}) - \alpha_t \nabla_w F(w_{t-1}; S_B); \\ \mu_B^0 &= w_{t-1} + \gamma(w_{t-1} - w_{t-2}) - \alpha_t \nabla_w F(w_{t-1}; S_B^0). \end{aligned}$$

Similarly, the following holds under noisy NAG:

$$\begin{aligned} \mu_B &= w_{t-1} + \gamma(w_{t-1} - w_{t-2}) - \alpha_t \nabla_w F(w_{t-1} + \gamma(w_{t-1} - w_{t-2}); S_B); \\ \mu_B^0 &= w_{t-1} + \gamma(w_{t-1} - w_{t-2}) - \alpha_t \nabla_w F(w_{t-1} + \gamma(w_{t-1} - w_{t-2}); S_B^0). \end{aligned}$$

In either case, it can be verified that the conditions of Lemma 21 hold for $\sigma = \frac{2\alpha_t L}{b}$ and $\sigma = \alpha_t$. The rest of the proof is the same as the proof of Theorem 11. ■

A.5.2 STABILITY BOUND FOR ENTROPY-SGD

In the Entropy-SGD algorithm due to Chaudhari et al. (2017), instead of directly optimizing the original objective $F(w)$, we minimize the negative local entropy, defined as follows:

$$E(w; \gamma) = -\log \int_{R^d} \exp\left(-F(w') - \frac{\gamma}{2}\|w - w'\|_2^2\right) dw' \tag{17}$$

Intuitively, a wider local minimum has a lower loss (i.e., $E(w;\gamma)$) than sharper local minima. See Chaudhari et al. (2017) for more details. The Entropy-SGD algorithm invokes standard SGD to minimize the negative local entropy. However, the gradient of negative local entropy

$$\nabla_w E(w;\gamma) = w - \mathop{E}_{w' \sim P}[w'], \quad P(w') \propto \exp\left(-F(w') - \frac{\gamma}{2}\|w - w'\|_2^2\right) \tag{18}$$

is hard to compute. Thus, the algorithm uses exponential averaging to estimate the gradient in the SGLD loop; see Algorithm 1 for more details.

We have the following generalization bound for Entropy-SGD.

Algorithm 1: Entropy-SGD

Input: Training set $S = (z_1, \ldots, z_n)$ and loss function $f(w; z)$.
Hyper-parameters: Scope γ, SGD learning rate η, SGLD step size η' and batch size b.

```
for t = 1 to T do
    //SGD iteration
    W_{t;0} ← μ_{t;0} ← W_{t-1;K+1};
    for k = 0 to K − 1 do
        //SGLD iteration
        B_{t;k} ← mini-batch with size b;
        W_{t;k+1} ← W_{t;k} − η'∇_w g(W_{t;k}; S_{B_{t;k}}) + η'γ(W_{t-1;K+1} − W_{t;k}) + √η' ε, ε ~ N(0, ½ I_d);
        μ_{t;k+1} ← (1 − α) μ_{t;k} + α W_{t;k};
    end
    W_{t;K+1} ← W_{t;K} − η(W_{t;K} − μ_{t;K});
end
return W_{T;K+1};
```

Theorem 27. Suppose that the loss function f is C-bounded and the objective function F is L-lipschitz. If batch size $b = n/2$ and $\sqrt{\eta'}\varepsilon = \varepsilon/(2L)$, the following expected generalization error bound holds for Entropy-SGD:

$$\text{err}_{\text{gen}} \le \frac{8\sqrt{12}C\sqrt{\eta'}}{\varepsilon n}\sqrt{\mathbb{E}_S\left[\sum_{t=1}^{T}\sum_{k=0}^{K-1} g_e(t;k)\right]}; \qquad \text{(empirical norm)}$$

where $g_e(t;k) = \mathbb{E}_{w\ W_{t;k}}[\frac{1}{n}\sum_{i=1}^{n}\|\nabla F(w; z_i)\|_2^2]$ is the empirical squared gradient norm, and $W_{t;k}$ denotes the training process with respect to S.

Since $g_e(t;k)$ is at most L^2, it further implies the generalization error of Entropy-SGD is bounded by $O\left(\frac{C\sqrt{\eta'}L}{\varepsilon n}\sqrt{TK}\right)$.

Proof of Theorem 27 Define the history before time step $(t;k)$ as follows:

$$W_{(t;k)} = (W_{0;0}; \ldots; W_{0;K+1}; \ldots; W_{t-1;0}; \ldots; W_{t-1;K+1}; W_{t;0}; \ldots; W_{t;k}). \tag{19}$$

Since μ is only determined by W, we only need to focus on W. This proof is similar to the proof of Theorem 11. By setting $P = \mathbb{E}_{\bar{S}}[Q_{(\bar{S};0)}]$. Suppose $S = (\bar{S}; z)$ and $S^0 = (\bar{S}; 0)$ are fixed, let W and W^0 denote their training process, respectively. Considering the following 3 cases:

1. $W_{t;0} \leftarrow W_{t-1;K+1}$: In this case, for a fixed $w_{(t-1;K+1)}$, we have
$$KL\left(W_{t;0}|W_{(t-1;K+1)} = w_{(t-1;K+1)}; W^0_{t;0}|W^0_{(t-1;K+1)} = w_{(t-1;K+1)}\right) = 0.$$
2. $W_{t;k+1} \leftarrow W_{t;k} - \eta'\nabla_w g(W_{t;k}; S_{B_{t;k}}) + \eta'\gamma(W_{t-1;K+1} - W_{t;k}) + \sqrt{\eta'}\varepsilon, \varepsilon \sim N(0, \frac{1}{2}I_d)$: In this case, fix a $w_{(t;k)}$, applying Lemma 21 gives
$$KL\left(W_{t;k+1}|W_{(t;k)} = w_{(t;k)}; W^0_{t;k+1}|W^0_{(t;k)} = w_{(t;k)}\right) \le \frac{8\cdot 23\ \eta'\|\nabla F(w_{t;k}; z)\|_2^2}{\varepsilon^2 n^2}.$$
3. $W_{t;K+1} \leftarrow W_{t;K} - \eta(W_t; K - \mu_{t;K})$: In this case, for a fixed $w_{(t;K)}$, we have
$$KL\left(W_{t;K+1}|W_{(t;K)} = w_{(t;K)}; W^0_{t;K+1}|W^0_{(t;K)} = w_{(t;K)}\right) = 0.$$

By applying Lemma 10, we have

$$KL(W_{T;K+1}; W^0_{T;K+1}) \le \frac{8\cdot 23\ \eta'}{\varepsilon^2 n^2}\sum_{t=1}^{T}\sum_{k=0}^{K-1} g_e(t;k);$$

and Where $g_e(t;k)$ is the empirical squared gradient norm of the k-th SGLD iteration in the t-th SGD iteration, respectively. The rest of the proof is the same as the proof of Theorem 11. ■

B PROOFS IN SECTION 4

B.1 MARKOV SEMIGROUP AND LOG-SOBOLEV INEQUALITY

The continuous version of the noisy gradient descent method is the Langevin dynamics, described by the following stochastic differential equation:

$$dW_t = -\nabla F(W_t)\,dt + \sqrt{2^{-1}}\,dB_t, \quad W_0 \sim \mu_0, \tag{20}$$

where B_t is the standard Brownian motion. To analyze the above Langevin dynamics, we need some preliminary knowledge about Log-Sobolev inequalities.

Let $p_t(w,y)$ denote the probability density function (i.e., probability kernel) describing the distribution of W_t starting from w. For a given SDE such as (20), we can define the associated diffusion semigroup P:

Definition 28 (Diffusion Semigroup) (see e.g., (Bakry et al., 2013, p. 39)) Given a stochastic differential equation (SDE), the associated diffusion semigroup $P = (P_t)_{t\ge 0}$ is a family of operators that satisfy for every $t \ge 0$, P_t is a linear operator sending any real-valued bounded measurable function f on $\mathbb{R}^d$ to

$$P_t f(w) = \mathbb{E}[f(W_t) \mid W_0 = w] = \int_{\mathbb{R}^d} f(y)\,p_t(w,dy).$$

The semigroup property $P_{t+s} = P_t \circ P_s$ holds for every $t, s \ge 0$. Another useful property of P_t is that it maps a nonnegative function to a nonnegative function. The carré du champ operator of this diffusion semigroup (w.r.t (20)) is (Bakry et al., 2013, p. 42)

$$\Gamma(f,g) = \tfrac{1}{2^{-1}}\langle \nabla f, \nabla g\rangle.$$

We use the shorthand notation $\Gamma(f) = \Gamma(f,f) = \frac{1}{2^{-1}}\|\nabla f\|_2^2$, and define (with the convention that $0 \log 0 = 0$)

$$\mathrm{Ent}_\mu(f) = \int_{\mathbb{R}^d} f \log f \,d\mu - \int_{\mathbb{R}^d} f\,d\mu \log \int_{\mathbb{R}^d} f\,d\mu.$$

Definition 29 (Logarithmic Sobolev Inequality) (see e.g., (Bakry et al., 2013, p. 237)) A probability measure μ is said to satisfy a logarithmic Sobolev inequality $\mathrm{LSI}(\alpha)$ (with respect to Γ), if for all functions $f : \mathbb{R}^d \to \mathbb{R}^+$ in the Dirichlet domain $\mathcal{D}(E)$,

$$\mathrm{Ent}_\mu(f) \le \frac{1}{2\alpha}\int_{\mathbb{R}^d} \frac{\Gamma(f)}{f}\,d\mu.$$

$\mathcal{D}(E)$ is the set of functions $f \in L^2(\mu)$ for which the quantity $\frac{1}{t}\int_{\mathbb{R}^d} f(f - P_t f)\,d\mu$ has a finite (decreasing) limit as t decreases to 0.

A well-known Logarithmic Sobolev Inequality is the following result for Gaussian measures.

Lemma 30 (Logarithmic Sobolev Inequality for Gaussian measure) (Bakry et al., 2013, p. 258) Let μ be the centered Gaussian measure on $\mathbb{R}^d$ with covariance matrix $\sigma^2 I_d$. Then μ satisfies the following LSI:

$$\mathrm{Ent}_\mu(f) \le \frac{\sigma^2}{2}\int_{\mathbb{R}^d} \frac{\|\nabla f\|_2^2}{f}\,d\mu$$

Lemma 30 states that the centered Gaussian measure with covariance matrix $\sigma^2 I_d$ satisfies $\mathrm{LS}(\sigma^2)$ (with respect to Γ), where $\Gamma = \frac{1}{2^{-1}}\langle \nabla f, \nabla g\rangle$ is the carré du champ operator of the diffusion semigroup defined above.

Before proving our results, we need some known results from Markov diffusion process. It is well known that the invariant measure (Bakry et al., 2013, p. 10) of the above CLD is the Gibbs measure $d\mu = \frac{1}{Z}\exp(-F(w))\,dw$ (Menz et al., 2014, (1.3)). In other words, μ satisfies $\int_{\mathbb{R}^d} P_t f\,d\mu = \int_{\mathbb{R}^d} f\,d\mu$ for every bounded positive measurable function f, where P_t is the Markov semigroup in Definition 28. The following lemma by Holley and Stroock Holley & Stroock (1987) (see also (Bakry et al., 2013, p. 240)) allows us to determine the Logarithmic Sobolev constant of the invariant measure μ.

Lemma 31 (Bounded perturbation) Assume that the probability measure satisfies LS( ) (with respect to). Let be a probability measure such that $\frac{1}{b}\,d \le d \le b\,d$ for some constant $b > 1$. Then satisfies LS(b^2) (with respect to).

In fact, Lemma 31 is a simple consequence of the following variational formula in the special case that $(x) = x \log x$, which we will also need in our proof:

Lemma 32 (Variational formula) (see .g., (Bakry et al., 2013, p. 240)) Let $I \to \mathbb{R}$ on some open interval $I \subset \mathbb{R}$ be convex of class C^2. For every (bounded or suitably integrable) measurable function $f : \mathbb{R}^d \to \mathbb{R}$ with values in I,

$$\int_{\mathbb{R}^d} (f)\,d \;-\; \left(\int_{\mathbb{R}^d} f\,d\right) = \inf_{r \in I} \int_{\mathbb{R}^d} [\,(f) - (r) - {}^{0}(r)(f - r)]\,d\,: \tag{21}$$

It is worth noting the integrand of the right-hand side is nonnegative due to the convexity of .

B.2 Logarithmic Sobolev Inequality for CLD

Recall that $F_S(w) = F(w;S) := F_0(w;S) + \|w\|_2^2/2$ is the sum of the empirical original objective $F_0(w;S)$ and ℓ_2 regularization. Let $d = \frac{1}{Z}\exp(-F_S(w))\,dw$ be the invariant (Gibbs) measure of CLD, and is the centered Gaussian measure $= \frac{1}{Z}\exp(-\|w\|_2^2/2)\,dw$. Invoking Lemma 30 with $^2 = \frac{1}{}$ shows that satisfies LS($1/$) (with respect to). Consider the density $h(w) = \frac{d}{d} = \frac{Z}{Z}\exp(-F_0(w;S))$. If the original objective function F_0 is C-bounded, we have $\exp(-2C)\ \le\ h(w)\ \le \exp(2C)$. By applying Lemma 31 with $b = \exp(2C)$, we have the following lemma.

Lemma 33. Under Assumption 14, let $(f;g) = \frac{1}{2}\langle\nabla f, \nabla g\rangle$ be the carré du champ operator of the diffusion semigroup associated to CLD, and be the invariant measure of the SDE. Then, satisfies LS($e^{4C}/$) with respect to .

Let $_t$ be the probability measure of W_t. By definition of P_t, for any real-valued bounded measurable function f on $\mathbb{R}^d$ and any $s, t \ge 0$,

$$\mathbb{E}_{w\,\sim\,_{t+s}}[f(w)] = \mathbb{E}_{w\,\sim\,_s}[P_t f(w)]. \tag{22}$$

In particular, if the invariant measure $= {}_1$ exists, we have

$$\mathbb{E}_{w\,\sim\,}[f(w)] = \mathbb{E}_{w\,\sim\,_1}[P_t f(w)] = \mathbb{E}_{w\,\sim\,_{t+1}}[f(w)] = \mathbb{E}_{w\,\sim\,}[P_t f(w)]. \tag{23}$$

The following lemma is crucial for establishing the first generalization bound for CLD. In fact, we establish a Log-Sobolev inequality for , the parameter distribution at time , for any time $t > 0$. Note that our choice of the initial distribution $_0$ is important for the proof.[6]

Lemma 34. Under Assumption 14, let $_t$ be the probability measure of W_t in (CLD) with initial probability measure $d_0 = \frac{1}{Z}e^{-\frac{\|w\|^2}{2}}\,dw$. Let be the carré du champ operator of diffusion semigroup associated to (CLD). Then, for any $f : \mathbb{R}^d \to \mathbb{R}^+$ in D(E):

$$\mathrm{Ent}_{\,t}(f) \;\le\; \frac{e^{8C}}{2}\int_{\mathbb{R}^d} \frac{(f)}{f}\,d\,_t$$

Proof Let be the invariant measure of CLD. By Lemma 33 and Definition 29,

$$\mathrm{Ent}\,(f) \;\le\; \frac{e^{4C}}{2}\int_{\mathbb{R}^d} \frac{\|\nabla f\|_2^2}{f}\,d\,: \tag{24}$$

[6] For arbitrary initial distribution, it is impossible to prove similar inequality for any 0 (unless the loss is strongly convex).

By applying Lemma 32 with $\Phi(x) = x \log x$, we rewrite the left-hand side as

$$\begin{aligned}\mathrm{Ent}_{\gamma}(f) &:= \int_{\mathbb{R}^d} f \log f \, d\gamma - \int_{\mathbb{R}^d} f \, d\gamma \log \int_{\mathbb{R}^d} f \, d\gamma \\ &= \inf_{r \in \mathbb{R}} \int_{\mathbb{R}^d} [\Phi(f) - \Phi(r) - \Phi'(r)(f - r)] \, d\gamma \\ &= \inf_{r \in \mathbb{R}} \int_{\mathbb{R}^d} [P_t(\Phi(f) - \Phi(r) - \Phi'(r)(f - r))] \, d\gamma,\end{aligned}$$

where the last equation holds by the definition of invariant measure $\int P_t f \, d\gamma = \int f \, d\gamma$. Thus, we have

$$\inf_{r \in \mathbb{R}} \int_{\mathbb{R}^d} [P_t(\Phi(f) - \Phi(r) - \Phi'(r)(f - r))] \, d\gamma = \mathrm{Ent}_{\gamma}(f) - \frac{e^{4C}}{2} \int_{\mathbb{R}^d} \frac{k \nabla_v f k_2^2}{f} \, d\gamma; \tag{25}$$

Let γ_t be the probability measure of W_t. Lemma 32 and (22) together imply that

$$\begin{aligned}\mathrm{Ent}_{\gamma_t}(f) &= \inf_{r \in \mathbb{R}} \int_{\mathbb{R}^d} [\Phi(f) - \Phi(r) - \Phi'(r)(f - r)] \, d\gamma_t \\ &= \inf_{r \in \mathbb{R}} \int_{\mathbb{R}^d} [P_t(\Phi(f) - \Phi(r) - \Phi'(r)(f - r))] \, d\gamma_0\end{aligned} \tag{26}$$

Since $P_t(\Phi(f) - \Phi(r) - \Phi'(r)(f - r)) \geq 0$[7] and $\frac{d\gamma_0}{d\gamma} \leq \exp(2C)$, we have

$$\begin{aligned}\mathrm{Ent}_{\gamma_t}(f) &= \inf_{r \in \mathbb{R}} \int_{\mathbb{R}^d} [P_t(\Phi(f) - \Phi(r) - \Phi'(r)(f - r))] \frac{d\gamma_0}{d\gamma} \, d\gamma \\ &\leq \exp(2C) \mathrm{Ent}_{\gamma}(f) - \frac{e^{6C}}{2} \int_{\mathbb{R}^d} \frac{k \nabla_v f k_2^2}{f} \, d\gamma.\end{aligned} \tag{27}$$

Since $\frac{d\gamma}{d\gamma_0} \leq \exp(2C)$ and γ is the invariant measure, we conclude that

$$\begin{aligned}\mathrm{Ent}_{\gamma_t}(f) - \frac{e^{6C}}{2} \int_{\mathbb{R}^d} \frac{k \nabla_v f k_2^2}{f} \, d\gamma &= \frac{e^{6C}}{2} \int_{\mathbb{R}^d} P_t \left(\frac{k \nabla_v f k_2^2}{f} \right) d\gamma \\ &= \frac{e^{6C}}{2} \int_{\mathbb{R}^d} P_t \left(\frac{k \nabla_v f k_2^2}{f} \frac{d\gamma}{d\gamma_0} \right) d\gamma_0 \\ &\leq \frac{e^{8C}}{2} \int_{\mathbb{R}^d} P_t \left(\frac{k \nabla_v f k_2^2}{f} \right) d\gamma_0 \\ &= \frac{e^{8C}}{2} \int_{\mathbb{R}^d} \frac{k \nabla_v f k_2^2}{f} \, d\gamma_t = \frac{e^{8C}}{2} \int_{\mathbb{R}^d} \frac{I(f)}{f} \, d\gamma_t\end{aligned} \tag{28}$$

■

Lemma 16. Under Assumption 14, let γ_t be the probability measure of W_t in CLD (with $d\gamma_0 = \frac{1}{Z} e^{-\frac{k w k^2}{2}} dw$). Let γ be a probability measure that is absolutely continuous with respect to γ_t. Suppose $d\gamma_t = \gamma_t(w) \, dw$ and $d\gamma = \gamma(w) \, dw$. Then it holds that:

$$\mathrm{KL}(\gamma \,;\, \gamma_t) \leq \frac{\exp(8C)}{2} \int_{\mathbb{R}^d} r \log \frac{\gamma(w)}{\gamma_t(w)}_2^2 \gamma(w) \, dw. \tag{29}$$

Proof Let $f(w) = \gamma(w) / \gamma_t(w)$, by Lemma 34 and $\int_{\mathbb{R}^d} f \, d\gamma_t = 1$, we have

$$\int_{\mathbb{R}^d} f \log f \, d\gamma_t \leq \frac{e^{8C}}{2} \int_{\mathbb{R}^d} \frac{k \nabla_v f k_2^2}{f} \, d\gamma_t \tag{30}$$

[7] This is because Φ is convex and P_t is a positive operator.

We can see that the left-hand side is equal to $KL(\ ;\ _t)$ [8], and the right-hand side is equal to

$$\frac{e^{8\ C}}{2}\int_{R^d} r\ \frac{(w)}{_t(w)}\ ^2_2\ _t(w)\ dw = \frac{e^{8\ C}}{2}\int_{R^d} r\ \log\frac{(w)}{_t(w)}\ ^2_2\ (w)\ dw:$$

This concludes the proof. ■

B.3 The Discretization Lemma from Raginsky et al. (2017)

Let $h(w;z) = F_0(w;z) + \frac{kwk_2^2}{2}$. We can rewrite $F_S(w) = \frac{1}{n}\sum_{i=1}^n h(w;z_i)$. Define $_{S;k}$ and $_{S;t}$ as the probability measure of W_k (in GLD) and W_t (in CLD), respectively. Raginsky et al. (2017) provided a bound of $KL(\ _{S;k}\ ;\ _{S;K}\ )$ under Assumption 35. This bound enables us to derive a generalization error bound for the discrete GLD from the bound for the continuous CLD. We use the assumption from Raginsky et al. (2017). Their work considers the following SGLD:

$$W_{k+1} = W_k \quad g_S(W_k) + \sqrt{\frac{2}{}}\ _k:$$

Where $g_S(w_k)$ is a conditionally unbiased estimate of the gradient $rF_S(w_k)$. In our GLD setting, $g_S(W_k)$ is equal to $r\ F_S(w_k)$.

Assumption 35. Let $F_S(w) = \frac{1}{n}\sum_{i=1}^n h(w;z_i) = F_0(w;S) + \frac{}{2}kwk_2^2$.

1. The function h takes non-negative real values, and there exist constants $A, B \quad 0$, such that
$$jh(0;z)j \quad A \qquad \text{and} \qquad kr\ h(0;z)k_2 \quad B \qquad 8z \ 2\ Z :$$
2. For each $z \ 2\ Z$, the function $h(\ ;z)$ is M-smooth: for some $M > 0$,
$$kr\ h(w;z)\ \ r\ \ h(v;z)k_2 \quad M\ kw \quad vk_2\ ; \qquad 8w;v \ 2\ R^d:$$
3. For each $z \ 2\ Z$, the function $h(\ ;z)$ is $(m;b)$-dissipative: for some $m > 0$ and $b \quad 0$,
$$hw;r\ h(w;z)i \quad m\ kwk_2^2 \quad b; \qquad 8w \ 2\ R^d:$$
4. There exists a constant $2\ [0;1)$, such that, for each $S \ 2\ Z^n$,
$$E[kg_S(w) \quad r \quad F_S(w)k_2^2] \quad 2 \quad M^2\ kwk_2^2 + B^2 \ ; \qquad 8w \ 2\ R^d:$$
5. The probability law $_0$ of the initial hypothesis W_0 has a bounded and strictly positive density p_0 with respect to the Lebesgue measure on R^d, and
$$_0 := \log\int_{R^d} e^{kwk_2^2} p_0(w)\ dw < 1 :$$

Lemma 36. (Raginsky et al., 2017, Lemma 7) Suppose that Assumption 35 holds and set $_{S;0} = _0$. Then, for any $k \ 2\ N$ and any $\ 2\ (0;1 \wedge \frac{m}{4M^2})$, the following inequality holds

$$KL(\ _{S;k}\ ;\ _{S;k}\ ) \quad (C_0 \quad + C_1\ )k\ ;$$

where C_0 and C_1 are constants that only depend on M, $_0$, m, b, , B and d.

B.4 Proofs for Main Theorems

Theorem 15. Under Assumption 14, CLD (with initial probability measure $d_0 = \frac{1}{Z}e^{\frac{kwk^2}{2}}\ dw$) has the following expected generalization error bound:

$$err_{gen} \quad \frac{2e^{4\ C}\ CL}{n}\ \sqrt{\ \ 1 \quad \exp\ \frac{T}{e^{8\ C}}} \tag{31}$$

[8] Indeed, $\int_{R^d} f\ \log f\ d\ _t = \int_{R^d} \frac{}{_t}\log(\frac{}{_t})\ _t\ dw = KL(\ ;\ _t)$

In addition, if F_0 is also M-smooth and non-negative, by setting $\gamma > 2$, $\xi > 0$ and $\beta \in [0, 1 \wedge \frac{\xi}{8M^2}]$, the GLD (running K iterations with the same η_0 as CLD) has the expected generalization error bound:

$$\mathrm{err}_{\mathrm{gen}} \leq 2C\sqrt{\frac{2KC_1^2}{n} + \frac{2CLe^{4C}}{n}\left(1 - \exp\left(\frac{K}{e^{8C}}\right)\right)}; \tag{32}$$

where C_1 is a constant that only depends on m, γ, ξ, b, L and d.

Proof of Theorem 15 We apply the uniform stability framework. Suppose S and S^0 are two neighboring datasets that differ on exactly one data point. Let $(W_t)_{t\geq 0}$ and $(W_t^0)_{t\geq 0}$ be the process of CLD running on S and S^0, respectively. Let ρ_t and ρ_t be the pdf of W_t^0 and W_t. We have

$$\begin{aligned}\frac{d}{dt}\mathrm{KL}(\rho_t;\rho_t) &= \frac{d}{dt}\int_{\mathbb{R}^d}\rho_t \log\frac{\rho_t}{\rho_t}\,dw \\ &= \int_{\mathbb{R}^d}\frac{d\rho_t}{dt}\log\frac{\rho_t}{\rho_t} + \rho_t\,\frac{\rho_t}{\rho_t}\,\frac{\frac{d\rho_t}{dt}\rho_t - \rho_t\frac{d\rho_t}{dt}}{\rho_t^2}\,dw \\ &= \int_{\mathbb{R}^d}\frac{d\rho_t}{dt}\log\frac{\rho_t}{\rho_t}\,dw - \int_{\mathbb{R}^d}\frac{\rho_t}{\rho_t}\frac{d\rho_t}{dt}\,dw\end{aligned} \tag{33}$$

According to Fokker-Planck equation (see Risken (1996)) for CLD, we know that

$$\frac{\partial\rho_t}{\partial t} = \frac{1}{\xi}\nabla\rho_t + \nabla\cdot(\rho_t\nabla F_{S^0}), \qquad \frac{\partial\rho_t}{\partial t} = \frac{1}{\xi}\nabla\rho_t + \nabla\cdot(\rho_t\nabla F_S).$$

It follows that

$$\begin{aligned}I &:= \int_{\mathbb{R}^d}\frac{d\rho_t}{dt}\log\frac{\rho_t}{\rho_t}\,dw \\ &= \int_{\mathbb{R}^d}\frac{1}{\xi}\nabla\rho_t + \nabla\cdot(\rho_t\nabla F_{S^0})\log\frac{\rho_t}{\rho_t}\,dw \\ &= -\frac{1}{\xi}\int_{\mathbb{R}^d}\langle\nabla\log\frac{\rho_t}{\rho_t};\nabla\rho_t\rangle\,dw - \int_{\mathbb{R}^d}\langle\nabla\log\frac{\rho_t}{\rho_t};\rho_t\nabla F_{S^0}\rangle\,dw; \qquad \text{(integration by parts)}\end{aligned}$$

and

$$\begin{aligned}J &:= \int_{\mathbb{R}^d}\frac{\rho_t}{\rho_t}\frac{d\rho_t}{dt}\,dw \\ &= \int_{\mathbb{R}^d}\frac{\rho_t}{\rho_t}\frac{1}{\xi}\nabla\rho_t + \nabla\cdot(\rho_t\nabla F_S)\,dw \\ &= -\frac{1}{\xi}\int_{\mathbb{R}^d}\langle\nabla\frac{\rho_t}{\rho_t};\nabla\rho_t\rangle\,dw - \int_{\mathbb{R}^d}\langle\nabla\frac{\rho_t}{\rho_t};\rho_t\nabla F_S\rangle\,dw. \qquad \text{(integration by parts)}\end{aligned}$$

Together with (33), we have

$$\begin{aligned}\frac{d}{dt}\mathrm{KL}(\rho_t;\rho_t) &= I - J \\ &= -\frac{1}{\xi}\int_{\mathbb{R}^d}\Big\langle\frac{\nabla\rho_t}{\rho_t} - \frac{\nabla\rho_t}{\rho_t};\nabla\rho_t\Big\rangle\Big\langle\frac{\nabla\rho_t}{\rho_t} - \frac{\rho_t\nabla\rho_t}{\rho_t^2};\nabla\rho_t\Big\rangle\,dw \\ &\quad - \int_{\mathbb{R}^d}\langle\nabla\log\frac{\rho_t}{\rho_t};\rho_t\nabla F_{S^0}\rangle - \frac{\rho_t}{\rho_t}\langle\nabla\log\frac{\rho_t}{\rho_t};\rho_t\nabla F_S\rangle\,dw \\ &= -\frac{1}{\xi}\int_{\mathbb{R}^d}\rho_t\Big\|\nabla\log\frac{\rho_t}{\rho_t}\Big\|_2^2\,dw + \int_{\mathbb{R}^d}\rho_t\langle\nabla\log\frac{\rho_t}{\rho_t};\nabla F_S - \nabla F_{S^0}\rangle\,dw \\ &\leq -\frac{1}{2\xi}\int_{\mathbb{R}^d}\rho_t\Big\|\nabla\log\frac{\rho_t}{\rho_t}\Big\|_2^2\,dw + \frac{\xi}{2}\int_{\mathbb{R}^d}\rho_t\|\nabla F_S - \nabla F_{S^0}\|_2^2\,dw.\end{aligned}$$

The last step holds because $a = \sqrt{\alpha}$, $b = \sqrt{\beta - \alpha}$ and $\frac{\|a\|_2^2}{2} + \frac{\|b\|_2^2}{2} \le \frac{4L^2}{n^2}$. Since $\|\nabla F_S - \nabla F_{S'}\|_2^2 \le \frac{4L^2}{n^2}$, by Lemma 16, we have

$$KL(\mu_t \| \nu_t) \le \frac{e^{8C}}{2} \int_{\mathbb{R}^d} \mu_t \log \frac{\mu_t}{\nu_t}^2 dw,$$

which implies

$$\frac{1}{e^{8C}} KL(\mu_t \| \nu_t) \le \frac{1}{2} \int_{\mathbb{R}^d} \mu_t \log \frac{\mu_t}{\nu_t}^2 dw.$$

Hence,

$$\frac{d}{dt} KL(\mu_t \| \nu_t) \le \frac{1}{e^{8C}} KL(\mu_t \| \nu_t) + \frac{2L^2}{n^2}, \quad \text{with } KL(\mu_0 \| \nu_0) = 0. \tag{34}$$

Solving this differential inequality gives

$$KL(\mu_t \| \nu_t) \le \frac{2L^2 e^{8C}(1 - e^{-t/e^{8C}})}{n^2}. \tag{35}$$

By Pinsker's inequality, we can finally see that

$$\sup_z |\mathbb{E}_A[l(W_T^0; z) - l(W_T; z)]| \le 2C \sqrt{\frac{1}{2} KL(\mu_T \| \nu_T)} \le \frac{2e^{4C} CL}{n} \sqrt{1 - e^{-\frac{T}{e^{8C}}}}.$$

By Lemma 4, the generalization error of CLD is bounded by the right-hand side of the above inequality.

Now, we prove the second part of the theorem. Let $(W_k)_{k=0}$ and $(W_k^0)_{k=0}$ be the (discrete) GLD processes training on S and S^0, respectively. Then for any $z \in Z$:

$$\begin{aligned} &|\mathbb{E}[l(W_K; z)] - \mathbb{E}[l(W_K^0; z)]| \\ &\le 2C \cdot TV(\mu_{S,K}, \mu_{S^0,K}) \qquad (C\text{-boundedness}) \\ &\le 2C \cdot (TV(\mu_{S,K}, \mu_{S,K}) + TV(\mu_{S,K}, \mu_{S^0,K}) + TV(\mu_{S^0,K}, \mu_{S^0,K})). \end{aligned}$$

Since $\alpha > 2$ and $\beta > \frac{1}{2}$, Assumption 35 holds with $A = C$, $B = L$, $m = \frac{\alpha}{2}$, $b = \frac{L^2}{2}$, $\mu = 0$ and $\mu_0 = \frac{d}{2} \log\left(1 + \frac{\alpha^2}{2}\right)$. By applying Pinskers inequality and Lemma 36, we have

$$TV(\mu_{S,K}, \mu_{S,K}) \le \sqrt{\frac{1}{2} KL(\mu_{S,K} \| \mu_{S,K})} \le \sqrt{\frac{1}{2} K C_1^2} \tag{36}$$

and

$$TV(\mu_{S^0,K}, \mu_{S^0,K}) \le \sqrt{\frac{1}{2} KL(\mu_{S^0,K} \| \mu_{S^0,K})} \le \sqrt{\frac{1}{2} K C_1^2}. \tag{37}$$

From (35), we have

$$TV(\mu_{S,K}, \mu_{S^0,K}) \le \sqrt{\frac{1}{2} KL(\mu_{S,K} \| \mu_{S^0,K})} \le \sqrt{\frac{L^2 e^{8C}\left(1 - e^{-\frac{K}{e^{8C}}}\right)}{n^2}} \tag{38}$$

Combining (36), (37) and (38), we have

$$|\mathbb{E}[l(W_K; z)] - \mathbb{E}[l(W_K^0; z)]| \le 2C \left(\sqrt{2KC_1^2} + \frac{2CLe^{4C}}{n} \sqrt{1 - e^{-\frac{K}{e^{8C}}}}\right) := \varepsilon_n.$$

By Definition 3, GLD is ε_n-uniformly stable. Applying Lemma 4 gives the generalization bound of GLD. ■

Lemma 37 (Exponential decay in entropy) (Bakry et al., 2013, Theorem 5.2.1) The logarithmic Sobolev inequality $LS(\rho)$ for the probability measure μ is equivalent to saying that for every positive function f in $L^1(\mu)$ (with finite entropy),

$$Ent_\mu(P_t f) \le e^{-2t/\rho} Ent_\mu(f)$$

for every $t \ge 0$.

The following Lemma shows that $P_t(\frac{d\ 0}{d}) = \ _t$ in our diffusion process.

Lemma 38. Let P denote the diffusion semigroup of CLD. Let d denote the invariant measure of P and let $_t$ denote the probability measure of W_t. Then $P_t(\frac{d\ 0}{d}) = \ _t$.

Proof Let $d \ = \ (x)\,dx$ and $d\ _t = \ _t(x)\,dx$. As shown in (Pavliotis, 2014, page 118), our diffusion process (Smoluchowski dynamics) is reversible, which means $(x)p_t(x;y) = \ (y)p_t(y;x)$. Thus for any $g(x)$, we have

$$\begin{aligned}
\mathop{E}_{x\ P_t(\frac{d\ 0}{d})}[g(x)] &= \int g(x)\ (x)(P_t(d\ _0=d\ ))(x)dx \\
&= \int g(x)\ (x)dx \int\ _0(y)p_t(x;y)=\ (y)dy \\
&= \int g(x)\ (x)dx \int\ _0(y)p_t(y;x)=\ (x)dy \\
&= \int g(x)\ (x)\ _t(x)=\ (x)dx = \mathop{E}_{x\ \ _t}[g(x)]:
\end{aligned}$$

Since g is arbitrary, $P_t(\frac{d\ 0}{d})$ and $_t$ must be the same. ■

Theorem 39. Suppose that $t > \ 8\,C$. Under Assumption 14, CLD (with initial distribution $d\ _0 = \frac{1}{Z}e^{-\frac{k\,w\,k^2}{2}}\,dw$) has the following expected generalization error bound:

$$\mathrm{err}_{gen} \quad \frac{8\,C^{\ 2}}{n} + 4\,C\exp\left(\frac{T}{e^{4\ C}}\right)^{p\ \overline{C}}:$$

In addition, if F_0 is also M-smooth and non-negative, by setting $> \ 2$, $> \frac{1}{2}$ and $\ 2\ [0;1\ ^\ \frac{2}{8M^{\ 2}}\ \frac{1}{}]$, the GLD process (running K iterations with the same $_0$ as CLD) has the expected generalization error bound:

$$\mathrm{err}_{gen} \quad 2C^{p\ \overline{2KC_1\ ^2}} + \frac{8\,C^{\ 2}}{n} + 4\,C\exp\left(\frac{K}{e^{4\ C}}\right)^{p\ \overline{C}};$$

where C_1 is a constant that only depends on $h, \ , \ , b, L$ and d.

Proof of Theorem 39 Suppose S and S^0 are two datasets that differ on exactly one data point. Let $(W_t)_{t\ \ 0}$ and $(W_t^0)_{t\ \ 0}$ be their processes, respectively. Let $d\ _t = \ _t(w)\,dw$ and $d\ _t^0 = \ _t^0(w)\,dw$ be the probability measure of W_t and W_t^0, respectively. The invariant measure of CLD S and S^0 are denoted as $$ and ^{0}, respectively. Recall that

$$d \ = \frac{1}{Z}e^{\ F_S(w)}\,dw; \qquad d\ ^0 = \frac{1}{Z\ _0}e^{\ F_{S^0}(w)}\,dw:$$

The total variation distance of $$ and ^{0} is

$$\begin{aligned}
\mathrm{TV}(\ ;\ ^0) &= \frac{1}{2}\int_{R^d} 1 \quad \frac{d\ ^0}{d}\ d \\
&= \frac{1}{2}\int_{R^d} 1 \quad \frac{Z}{Z\ _0}\exp(\quad (F_{S^0}(w) \quad F_S(w)))\ \frac{1}{Z}e^{\ F_S(w)}\,dw:
\end{aligned} \tag{39}$$

Since $\frac{Z}{Z\ _0}\exp(\quad (F_{S^0}(w) \quad F_S(w))) \ 2\ \left[e^{\ \frac{4\ C}{n}}; e^{\frac{4\ C}{n}}\right]$ and $\frac{4\ C}{n} < 1=2$, we have

$$\mathrm{TV}(\ ;\ ^0) \quad \max\left(\frac{1}{2}\ 1\ \ e^{\ \frac{4\ C}{n}}\ ; \frac{1}{2}\ e^{\frac{4\ C}{n}}\ \ 1\right) \quad \frac{4\ C}{n}: \tag{40}$$

Since $$ and ^{0} satisfy $\mathrm{LS}(e^{4\ C=\ })$ (Lemma 33), applying Lemma 37 with $= \frac{d\ 0}{d}$ and $^0 = \frac{d\ ^0}{d\ ^0}$ and Lemma 38 yields:

$$\mathrm{KL}(\ _t;\ ) \quad \exp\left(\frac{2\,t}{e^{4\ C}}\right)\ \mathrm{KL}(\ _0;\ ); \qquad \mathrm{KL}(\ ^0_t;\ ^0) \quad \exp\left(\frac{2\,t}{e^{4\ C}}\right)\ \mathrm{KL}(\ ^0_0;\ ^0): \tag{41}$$

Since $KL(\ _0;\ )$ and $KL(\ ^0_0;\ ^0)$ are upper bounded by $\frac{1}{2}C$, Pinsker's inequality implies that $TV(\ _t;\ )$ and $TV(\ ^0_t;\ ^0)$ are upper bounded by $\sqrt{\frac{2t}{e^{4C}}\ C}$. Combining with (40) and note that $TV(\ _t;\ ^0_t)\ \ TV(\ _t;\ ) + TV(\ ;\ ^0) + TV(\ ^0_t;\ ^0)$, we have

$$\sup_z \left| \mathop{E}_A[L(W_T; z)\ \ L(W^0_T; z)] \right| \ \ 2C\ \ TV(\ _t;\ ^0_t)\ \ 4C\ \ \sqrt{\frac{2t}{e^{4C}}\ \ C + \frac{8C^2}{n}}.$$

By Lemma 4, the generalization error of CLD is bounded by the right-hand side.

The proof for GLD proceeds in the same way as the second part of the proof of Theorem 15.

C Experiment Details

We first present the general setup of our experiments:

Dataset: We use MNIST (LeCun et al., 1998) and CIFAR10 (Krizhevsky & Hinton, 2009) in our experiments.

Neural network: In our experiments, we test two different neural networks: a smaller version of AlexNet (Krizhevsky et al., 2012) and MLP. The structures of the networks are similar to what are used in Zhang et al. (2017a).

> Small AlexNet: k is the kernel size, d is the depth of a convolution layer, $fc(\cdot)$ is the fully-connected layer that has $\cdot$ neurons. The ReLU activation are used in the first 6 layers.

1	2	3	4	5	6	7
conv(k:5,d:64)	pool(k:3)	conv(k:5,d:192)	pool(k:3)	fc(384)	fc(192)	fc(10)

> MLP: The MLP used in our experiment has 3 hidden layers, each having width 512. We also use ReLU as the activation function in MLP.

Objective function: For a data point $z = (x; y)$ in MNIST, the objective function is

$$F(W; z) = \ \ln(\text{softmax}(\text{net}_W(x))[y]);$$

where $\text{softmax}(a)[i] = \frac{e^{a[i]}}{\sum_{j=1}^{10} e^{a[j]}}$, and $\text{net}_W(x)$ is the output of the neural network (10 dimensional vector). Note that the objective function F is exactly the cross-entropy loss.

0/1 loss : The 0-1 loss L^{01} is defined as:

$$L^{01}(W; (x; y)) = \begin{cases} 1 & (\arg\max_i \text{net}_W(x)[i]) \neq y; \\ 0 & \text{otherwise} \end{cases} \tag{42}$$

Random labels: Suppose the dataset contains n datapoint, and the corruption portion is p. We randomly select $n\ p$ data points, and replace their labels with random labels, as in Zhang et al. (2017a).

C.1 Experimental results for GLD

The result of this experiment (see Figure 1) is discussed in Section 3.1. Here we present our implementation details.

We repeat our experiment 5 times. At every individual run, we first randomly sample 10000 data points from the complete MNIST training data. The initial learning rate $= 0:003$. It decays 0.995 after every 60 steps, and it stops decaying when it is lower than 0.0005. During the training, we keep $_t = 0:2\sqrt{\ _t}$. Recall that the empirical squared gradient norm $g_{et}(t)$ in our bound (Theorem 9)

is $\mathbb{E}_{W_{t-1}}[\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(W_{t-1},z)\|^2]$. Since it is time-consuming to compute the exact $g_e(t)$, in our experiment, we use an unbiased estimation instead. At every step, we randomly sample a mini-batch B with batch size 200 from the training data, and use $\frac{1}{200}\sum_{i\in B}\|\nabla f(W_{t-1},z_i)\|^2$ as $g_e(t)$ to compute our bound in Figure 1. The estimation of $g_e(t)$ at every step t is shown in Figure 1(d). Since $g_e(t)$ is not very stable, in our figure, we plot its moving average over a window of size 100 to make the curve smoother (i.e. $g_{avg}(t) = \frac{1}{100}\sum_{\tau=t}^{t+100} g_e(\tau)$).

C.2 Experimental Results for SGLD

In this subsection, we present some experiment results for running SGLD on both MNIST and CIFAR10 datasets, to demonstrate that our bound (see Theorem 11), in particular the sum of the empirical squared gradient norms along the training path, can distinguish normal dataset from dataset that contains random labels. As shown in Figure 3, the curves of our bounds look quite similar to the generalization curves. Due to the sub-optimal constants in our bound, the bound is currently greater than 1, and hence we omit the numbers on the y-axis.

We note that in our experiments presented in Figure 3, the learning rate that we choose is larger than that required by the second condition of Theorem 11. This is because the global Lipschitz constant L is hard to estimate and the model is not able to fit training data under a very large noise. As discussed in Section 3.1, we can relax $(20L)^{-1}_t$ to $(2\max_{i\in[n]}\|\nabla F(W_{t-1},z_i)\|)^{-1}_t$. By applying gradient clipping trick, we can further relax this condition to

$$\eta_t \le \min\{C_L, (2\max_{i\in[n]}\|\nabla F(W_{t-1},z_i)\|)^{-1}\}\eta_t, \tag{43}$$

where C_L is defined in Section 3.1. In order to show that our observation ("random normal") still holds when the step size satisfies the requirement of our theory, we run an experiment that using gradient clipping trick with $C_L = 1$. The model is trained on a small subset of MNIST as fitting the original data set with random labels under such a large Gaussian noise is extremely slow. As shown in Figure 4, the experimental results remain unchanged when all the conditions of our bound are met.

These experiments indicate that the sum of squared empirical gradient norms is highly related to the generalization performance, and we believe by further optimizing the constants in our bound, it is possible to achieve a generalization bound that is much closer to the real generalization error.

Figure 2: Training MLP with GLD ($\eta_t = 0.2\,\eta_t$) on the full MNIST dataset without label corruption. Learning rate $\eta_t = 0.01\,(0.95)^{\lfloor t/60c\rfloor}$. Note that in the early stage, the testing accuracy is even higher than the training accuracy, thus we plot the absolute value of generalization error. As shown in this figure, even when the training accuracy approaches 90%, our bound is still relatively small.