Proceedings of Machine Learning Research vol 75:1–34, 2018 31st Annual Conference on Learning Theory

Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints

Wenlong Mou WMOU@EECS.BERKELEY.EDU
Department of EECS, University of California, Berkeley

Liwei Wang WANGLW@CIS.PKU.EDU.CN
Key Laboratory of Machine Perception, MOE, School of EECS, Peking University
Center for Data Science, Peking University, Beijing Institute of Big Data Research

Xiyu Zhai XZ380@CAM.AC.UK
University of Cambridge

Kai Zheng ZHENGK92@PKU.EDU.CN
Key Laboratory of Machine Perception, MOE, School of EECS, Peking University

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We study the generalization errors of *non-convex* regularized ERM procedures using Stochastic Gradient Langevin Dynamics (SGLD). Two theories are proposed with non-asymptotic discrete-time analysis, using stability and PAC-Bayesian theory respectively. The stability-based theory obtains a bound of $O\left(\frac{1}{n}L\sqrt{\beta T_N}\right)$, where L is Lipschitz parameter, β is inverse temperature, and T_N is the sum of step sizes. For PAC-Bayesian theory, though the bound has a slower $O(1/\sqrt{n})$ rate, the contribution of each step decays exponentially through time, and the uniform Lipschitz constant is also replaced by actual norms of gradients along the optimization trajectory. Our bounds have reasonable dependence on aggregated step sizes, and do not explicitly depend on dimensions, norms or other capacity measures of the parameter. The bounds characterize how the noises in the algorithm itself controls the statistical learning behavior in non-convex problems, without uniform convergence in the hypothesis space, which sheds light on the effect of training algorithms on the generalization error for deep neural networks.

Keywords: algorithm-dependent generalization bound; stochastic gradient Langevin dynamics; stability; PAC-Bayesian theory; non-convex learning

1. Introduction

One of the central topics of modern statistical learning theory is to derive algorithm-dependent and data-dependent generalization bounds for learning algorithms and models. A learning algorithm may use a large hypothesis space, but its randomized way of exploring the space controls actual capacity in a data-dependent manner. As a result, algorithm-dependent bounds usually go beyond classical notions of model capacities, such as VC dimensions and Rademacher complexities. For stochastic gradient methods (SGM) in particular, the number of iterations and step sizes serve as implicit regularization and restrict the growth of model capacity. Algorithm-dependent generalization bounds and statistical properties have been intensively studied for SGM under convex settings (Hardt et al., 2015; Lin and Rosasco, 2016; Lin et al., 2016; Wei et al., 2017; Chen et al., 2016), but very few are known for the non-convex case. Nevertheless, practitioners believe the latter to hold true in a regime far beyond existing theories. The prevailing success of stochastic gradient methods in non-convex learning problems is attributed not only to computational speed, but also to their merits on generalization error. The most important arena for algorithm-dependent generalization bound is perhaps deep learning, where model capacity is usually larger than number of data points, but good test error is achieved in practice.

The goal of this paper is to understand the effect of stochastic gradient methods on generalization performance with non-convex risk minimization. We would also like to emphasize that algorithm-dependent bounds for multi-pass non-convex optimization algorithms play a much more non-trivial role than their convex counterparts: single pass of SGD for convex objectives already achieves optimality in stochastic optimization; but in non-convex settings, the computational aspects naturally requires going through training data for much more than one pass. We consider the (regularized) empirical risk minimization procedure, where $R(\cdot)$ is a regularization term independent of data.

$$\underset{\boldsymbol{w}}{\text{minimize}} \left\{ F_n(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{w}, z_i) + R(\boldsymbol{w}) \right\}. \tag{1}$$

The excess risk of a learning algorithm is the sum of its optimization error and generalization error. While a lot of existing works have studied the first part, we focus on the second aspect. We consider the generalization error, i.e., the gap between training loss and population loss, by taking expectation with respect to the randomized algorithm $\mathcal{A}$. (We slightly abuse the notation: $\mathrm{err}_{gen}(\boldsymbol{w})$ is actually a function of the distribution of $\boldsymbol{w}$)

$$\mathrm{err}_{gen}(\boldsymbol{w}) \triangleq \mathbb{E}_{\mathcal{A}} \Big(\mathbb{E}_z \ell(\boldsymbol{w}; z) - \hat{\mathbb{E}}_n \ell(\boldsymbol{w}; z) \Big) \tag{2}$$

For flexibility and convenience, we do not assume any relationship between the loss function $f(\cdot; z_i)$ for optimization algorithm and $\ell(\cdot, z_i)$: they can be the same, or surrogate loss may be used. For example, in classification problems, f_i is usually hinge loss or logistic loss, while ℓ_i is $0-1$ loss.

We study the Stochastic Gradient Langevin Dynamics(SGLD) algorithm, which adds isotropic Gaussian noise to each stochastic gradient step, i.e.,

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta_k \tilde{\boldsymbol{g}}_k(\boldsymbol{w}) + \sqrt{\frac{2\eta_k}{\beta}} \mathcal{N}(0, I_d), \tag{3}$$

where $\tilde{\boldsymbol{g}}_k = \boldsymbol{g}_k(\boldsymbol{w}) + \boldsymbol{\nabla} R(\boldsymbol{w})$ is the stochastic gradient for regularized objective, and $\boldsymbol{g}_k = |B_k|^{-1}\sum_{j\in B_k} \boldsymbol{\nabla} f(\boldsymbol{w}, z_j)$ is the gradient evaluated on current batch B_k. We assume the algorithm is initialized with $\boldsymbol{w}_0 \sim \pi_0 = \mathcal{N}(0, \sigma_0^2 I_d)$, which is commonly used in practice.

The SGLD algorithm exhibits several nice properties even for non-convex functions, and has been used for sampling (Bubeck et al., 2015; Nagapetyan et al., 2017; Dalalyan, 2017; Cheng et al., 2017; Cheng and Bartlett, 2017) and non-convex optimization (Raginsky et al., 2017; Zhang et al., 2017b). The noise helps the algorithm to escape from saddle points and even shallow local minima, and hit a good local minimum in polynomial time. In deep learning practice, SGLD and other noise injection methods have also been shown to be helpful (Neelakantan et al., 2015; Chaudhari et al., 2016; Ye et al., 2017; Zhang et al., 2017a).

The effect of stochastic gradient methods on statistical learning has attracted lots of interests in existing literature: For least square regression in RKHS, (Lin and Rosasco, 2016; Lin et al., 2016) analyze multi-pass stochastic gradient methods, leading to optimal population risks; more general cases are studied via uniform stability of parameters under ℓ_2 norm (Hardt et al., 2015; London, 2016). Most of them require objective functions to be convex. While Hardt et al. (2015) considered non-convex smooth objective functions, their results depend exponentially on aggregated step sizes and smoothness parameter. Raginsky et al. (2017) proved strong excess risk bounds for SGLD under different assumptions, and their results are based on convergence to stationary distributions, which usually has exponential dependence on dimension. Recently, Pensia et al. (2018) proposed another algorithm-dependent generalization bounds for non-convex learning, based on the method of mutual information (Xu and Raginsky, 2017). Their bound works for more general iterative algorithms with noise injection, but the rate for SGLD is not as sharp as ours.

1.1. Contributions

We adopt two theoretical tools: uniform stability (Elisseeff et al., 2005; Rakhlin et al., 2005) and PAC-Bayesian theory (McAllester, 2003; Germain et al., 2016) to obtain data-dependent and algorithm-dependent bounds. These two approaches not only make it convenient to analyze generalization properties along optimization trajectory, but also provide different viewpoints towards the effect of SGLD on generalization: stability only depends on relative location between parameters trained with neighboring datasets, and $O(1/n)$ fast rates are usually available; on the other hand, PAC-Bayes bounds can benefit from norm-based regularization, and it also gives instance-dependent results, instead of taking worst-case upper bounds.

The main contributions of this paper are thus two-fold. The two generalization bounds obtained by the two methods reveal different aspects in which SGLD controls model complexity. It is important to note that the bounds have no explicit dependence on dimension of parameter space, nor do they explicitly depend on norm of parameters. By assuming only the Lipschitz condition on the objective function, the generalization bounds are controlled by aggregated step sizes.

Stability-based Bounds

We use the well-known connection between uniform stability and expected generalization error of randomized learning algorithms (Elisseeff et al., 2005). To derive upper bound for the hypothesis stability $\sup_z\{\ell(\boldsymbol{w}_N, z) - \ell(\boldsymbol{w}'_N, z)\}$, we choose to exploit the squared Hellinger distance $D_H(p_N||p'_N)$ between the distributions of parameter trained on adjacent datasets, instead of Euclidean distance in the parameter space, which is commonly use in previous works. This key difference makes it possible to prove non-trivial bounds with the presence of fence-sitting situation (Illustrated in Appendix B), in which the iterations in $\boldsymbol{w}_k$ are sensitive to perturbations.

By bounding the uniform stability of SGLD algorithm, we get the following result:

Theorem 1 *(Informal version of Theorem 13) Consider N rounds of SGLD with parameters β, $\{\eta_i\}_{i=1}^N$ and batch size 1. Suppose that the loss function $l(\boldsymbol{w}; z)$ is uniformly bounded by C, and each $f(\cdot, z)$ is L-Lipschitz. Assuming $\eta_i \leq \frac{\ln 2}{\beta L^2}, \forall i$, we have:*

$$\mathbb{E}[\mathrm{err}_{gen}(\boldsymbol{w}_N)] \leq \frac{2LC}{n}\left(\beta\sum_{i=1}^{N}\eta_i\right)^{1/2} \tag{4}$$

The theorem works without assuming any decay of step sizes. Nor do we assume any properties about convexity or second order smoothness. We can also deal with a few larger step sizes, as discussed in the complete version (Theorem 13).

The bound achieves an $O(1/n)$ fast rate, and only has square root dependence on the aggregated step sizes. Regarding Lipschitz and temperature parameters as constants, good generalization performance is guaranteed as long as $T_N = \sum_{i=1}^{N} \eta_i$ is much smaller than $o(n^2)$.

PAC-Bayesian Bounds

By bounding the KL divergence between output distribution of the algorithm and Gaussian priors, we obtain the following generalization guarantee via PAC-Bayesian theory.

Theorem 2 *(Informal version of Theorem 16) Let the ℓ_2 regularization term be $R(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{w}\|^2$. Under sub-Gaussian assumptions on $\ell(\cdot,\cdot)$, with suitable choice of initialization variance, we have the following with high probability:*

$$\mathrm{err}_{gen}(\boldsymbol{w}_N) \leq O\left(\sqrt{\frac{\beta}{n}\sum_{k=1}^{N} \eta_k e^{-\frac{\lambda}{3}(T_N - T_k)} \mathbb{E}\left[\|\boldsymbol{g}_k(\boldsymbol{w}_k)\|^2\right]}\right). \tag{5}$$

Though the bound can only achieve a slower $O(1/\sqrt{n})$ rate, it can benefit largely from the exponentially decaying factor. And the uniform Lipschitz constant is also replaced by a data-dependent gradient norm. Therefore, even if the gradients may be large at the beginning, their contribution to the generalization bound will be diminishing exponentially as time elapses. As long as the last few gradient steps are not very large, the generalization error will be controlled nicely. This phenomenon makes it possible for this bound to be even better. Besides, the assumption on loss function class is also weaker than stability bound and satisfied by many natural problems. For example, if the loss function grows linearly at infinity, Gaussian initialization ensures subGaussian properties of $\ell(\cdot,\cdot)$.

Previous analyses of the Gaussian noise in stochastic gradient methods mainly focus on its benefit for optimization aspect. The question naturally comes whether it also helps generalization a lot. Our paper gives an affirmative answer. Intuitively, the Gaussian noise makes the distribution smooth and stable, which restricts the average ability of over-fitting for the predictor. In the Appendix B, we present a graphical illustration of this phenomenon.

Our Techniques

Let's first consider Langevin diffusion $d\boldsymbol{w}(t) = -\boldsymbol{\nabla} F_n(\boldsymbol{w}(t))dt + \sqrt{2\beta^{-1}}d\boldsymbol{B}(t)$. The distribution π_t of $\boldsymbol{w}_t$ satisfies the Fokker-Planck equation $\frac{\partial \pi}{\partial t} = \frac{1}{\beta}\Delta\pi + \boldsymbol{\nabla}\cdot(\pi\boldsymbol{\nabla} F_n)$. We can conveniently take time derivatives for the quantities of our interests, and estimate their upper bounds. In Section 3, we illustrate this idea by analyzing generalization error bounds for the continuous time limits.

Intuitively, the exponential decaying factor in PAC-Bayes bound is because the amount of influence on final distribution by a single step is being weakened by the interplay between Gaussian noise and ℓ_2 penalty. Technically, it comes from the logarithmic Sobolev inequality, which relates Fisher information to KL divergence in our derivative bounds. In contrast to the convergence analysis, we are computing these quantities with respect to an isotropic Gaussian. So we do not suffer from the exponentially small constant for log-Sobolev in multi-modal stationary distributions.

Going from continuous to discrete is highly nontrivial. Note that almost all existing discretization techniques for Langevin Dynamics induce errors polynomial in dimension (Raginsky et al., 2017; Bubeck et al., 2015). If we directly estimate discretization gap, all the previous efforts will go in vain. Fortunately, since our results do not rely on convergence to the stationary, we can avoid discretization gap by creating a different equation for each step, so that the continuous process coincides exactly with the discrete update. Brownian motions with constant drifts and Ornstein-Uhlenbeck processes are exploited in the two continuous-time constructions, respectively.

2. Preliminaries

Notation: We assume data points $\boldsymbol{z}_i \in \mathcal{Z}(\forall i \in \{1, 2, \ldots, n\})$ are i.i.d. samples from an unknown distribution. Model parameter is $\boldsymbol{w} \in \mathbb{R}^d$. A pair of neighboring datasets $S, S' \in \mathcal{Z}^n$ means that S and S' differ on exactly one data point. For a continuous time stochastic differential equation (SDE) run on dataset S, the iteration point at time t is denoted as $\boldsymbol{w}_t$, and corresponding density function is denoted as $\pi_t(\boldsymbol{w})$. For discrete time SGLD run on S, the iteration point and its density function at round k are written as $\boldsymbol{w}_k, p_k(\boldsymbol{w})$ respectively. All above notations are also suitable for S' with an additional prime. We sometimes omit the subscript t for π_t, π'_t without confusion. η_k is the step size of discrete SGLD at iteration k, and $T_k \triangleq \sum_{j=1}^{k} \eta_j$. Let $\boldsymbol{g}_k(\cdot)$ be the stochastic gradient operator at round k without regularization, and let $\tilde{\boldsymbol{g}}_k(\boldsymbol{w}) = \boldsymbol{g}_k(\boldsymbol{w}) + \boldsymbol{\nabla} R(\boldsymbol{w})$. $D_H(p||q)$ represents the squared Hellinger distance between density function p and q, i.e., $D_H(p||q) \triangleq \frac{1}{2}\int_{\mathbb{R}^d}\left(\sqrt{p} - \sqrt{q}\right)^2 dw$.

Now we define an important property of the update operators which will be frequently used:

Definition 3 (non-expansive) *Suppose $\boldsymbol{w}$ and $\boldsymbol{w}'$ are two random points in $\mathbb{R}^d$, and their distributions are denoted as $\mathcal{P}, \mathcal{P}'$. We say a bivariate functional $D(\cdot||\cdot)$ defined on two density functions, is non-expansive, if for any (possibly randomized) measurable mapping $\psi : \mathbb{R}^d \to \mathbb{R}^d$, we have*

$$D(\psi(\mathcal{P})||\psi(\mathcal{P}')) \leqslant D(\mathcal{P}||\mathcal{P}'), \tag{6}$$

where $\psi(\mathcal{Q})$ is defined as the probability distribution of $\psi(\boldsymbol{X})$ where $\boldsymbol{X} \sim \mathcal{Q}$.

It is well known that all f-divergences (including KL divergence, squared Hellinger distance, and total variation distance) are non-expansive and jointly convex (Csiszár et al., 2004).

2.1. Stability and generalization

Stability of the algorithm has a close relation with its generalization performance, which dates back to Bousquet and Elisseeff (2002). Intuitively, the more stable an algorithm is, the better its generalization performance will be. Here, we adopt the notion of uniform stability of a randomized algorithm (Elisseeff et al., 2005; Hardt et al., 2015), and use it to bound generalization performance.

Definition 4 (Uniform Stability) *A randomized algorithm $\mathcal{A}$ is ϵ_n-uniformly stable w.r.t the loss ℓ, if for all neighboring sets $S, S' \in \mathcal{Z}^n$, it holds that $\sup_{\boldsymbol{z}} |\mathbb{E}_{\mathcal{A}}[\ell(\boldsymbol{w}_S; \boldsymbol{z})] - \mathbb{E}_{\mathcal{A}}[\ell(\boldsymbol{w}_{S'}; \boldsymbol{z})]| \leqslant \epsilon_n$, where $\boldsymbol{w}_S, \boldsymbol{w}_{S'}$ are outputs of $\mathcal{A}$ on S and S' respectively.*

Theorem 5 (Generalization in expectation) *(Elisseeff et al., 2005; Hardt et al., 2015) Suppose a randomized algorithm A is ϵ_n-uniformly stable, then there is $|\mathbb{E}[\mathrm{err}_{gen}(\boldsymbol{w}_S)]| \leqslant \epsilon_n$.*

Under suitable assumptions, it is straightforward to extend our results to high-probability guarantees with respect to random draw of training data with an additional $O(\sqrt{\frac{\log 1/\delta}{n}})$ term, using McDiarmid Inequality. For simplicity, we restrict our attention to expected generalization bounds.

2.2. PAC-Bayesian theory

Different with the uniform stability theory, which requires considering the worst case neighboring datasets, the generalization bounds implied by PAC-Bayesian theory are completely algorithmic and data dependent. However, most of the generalization bounds via PAC-Bayesian theory assume bounded loss function, or work under specific contexts (Dalalyan and Tsybakov, 2012). Germain et al. (2016) extended previous results to s-subGaussian losses, but their result introduced an extra additive error term $\frac{1}{2}s^2$. To get rid of this additive term and facilitate our later analysis, we first improve the PAC-Bayesian result in Germain et al. (2016) as follows:

Theorem 6 *For loss function $\{\ell(\boldsymbol{w};\boldsymbol{x})\}$ and data distribution $\mathcal{D}$. Given any prior distribution $\mathcal{P}$ over Ω. If the loss class is s-subGaussian w.r.t $\mathcal{D}\times\mathcal{P}$, i.e $\mathbb{E}e^{\lambda(\ell(\boldsymbol{w};\boldsymbol{x})-\mathbb{E}\ell(\boldsymbol{w};\boldsymbol{x}))}\le e^{\frac{1}{2}\lambda^2 s^2}$ $(\forall\lambda)$, and let Ξ be a class of distributions over Ω, with $\sup_{\mathcal{Q}\in\Xi} D_{KL}(\mathcal{Q}||\mathcal{P})\le M$, then with probability $1-\delta$:*

$$\forall\mathcal{Q}\in\Xi,\quad \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\mathcal{Q}}\ell(\boldsymbol{w};\boldsymbol{x})\le\hat{\mathbb{E}}_n\mathbb{E}_{\mathcal{Q}}\ell(\boldsymbol{w};\boldsymbol{x})+O\left(s\sqrt{\frac{D_{KL}(\mathcal{Q}||\mathcal{P})\vee 1+\log\frac{1}{\delta}+\log\log M}{n}}\right)$$

3. Continuous Time Limit: Generalization Bounds for Langevin Equation

Intuitively, SGLD can be seen as a discretization of Langevin Equation. Understanding generalization performance of the ideal continuous-time algorithm provides important insights into more technically involved analysis for discrete-time algorithm. In this section, we will present two generalization error bounds for continuous time Langevin equation, using stability and PAC-Bayesian theory respectively. We elaborate on the techniques used in our analysis, which give a high-level view of how generalization bound for discrete-time SGLD can be possibly obtained.

Consider the following continuous-time Langevin Equation, where F_n is the (regularized) empirical objective function.

$$d\boldsymbol{w}(t)=-\boldsymbol{\nabla}F_n(\boldsymbol{w}(t))dt+\sqrt{2\beta^{-1}}d\boldsymbol{B}(t),\quad t\ge 0 \tag{7}$$

where $\{\boldsymbol{B}(t)\}_{t\ge 0}$ is the standard Brownian motion in $\mathbb{R}^d$.

Let π_t be the density of $\boldsymbol{w}_t$, which satisfies the following Fokker-Planck equation (see Appendix C for background):

$$\frac{\partial\pi}{\partial t}=\frac{1}{\beta}\Delta\pi+\boldsymbol{\nabla}\cdot(\pi\boldsymbol{\nabla}F_n) \tag{8}$$

3.1. Uniform Stability

We are going to bound uniform stability at arbitrary time T with respect to loss function, which directly controls generalization in expectation. In this part, the function $R(\cdot)$ doesn't affect our analysis. It can be 0 or any regularization functions, hence we omit it.

For uniform stability, we assume that $f(\boldsymbol{w};\boldsymbol{z})$ satisfies the following condition which is slightly weaker than uniform Lipschitz w.r.t w for any z. Note that the generalization performance is defined in terms of loss function ℓ, which may not be continuous, but the Lipschitz assumption is imposed on objective f of our algorithm, which can be a surrogate function for ℓ.

$$\forall z,z',\quad \|\nabla f(\boldsymbol{w};z)-\nabla f(\boldsymbol{w};z')\|\le L \tag{9}$$

As a result, we have $\|\boldsymbol{\nabla}F_n(\boldsymbol{w})-\boldsymbol{\nabla}F_n'(\boldsymbol{w})\|\le\frac{L}{n}$ for any neighboring datasets S,S'.

First, we can use squared Hellinger distance between outputs of the algorithm over neighboring datasets to bound uniform stability ϵ_n.

Lemma 7 *Assuming ℓ is uniformly bounded by C, and denote π (π') as the pdf of outputs of any algorithm run over dataset S (S'). Then the uniform stability ε_n satisfies: $\epsilon_n \leqslant \sup_{S,S'} 2C\sqrt{D_H(\pi_T||\pi_T')}$*

Compared with Hardt et al. (2015), the bound based on f-divergence can better characterize stability with non-convex objective: through one step of iteration, the ℓ_2 distance $\mathbb{E}\|\boldsymbol{w}_k - \boldsymbol{w}_k'\|^2$ between parameters can expand a lot due to shape of non-convex surface, f-divergences are non-expansive under the same transformation, and will decrease by convolution with Gaussian noise.

Proposition 8 *Under above assumptions, the expected generalization error for continuous-time Langevin Equation is bounded by:*

$$\mathbb{E}[\mathrm{err}_{gen}(\boldsymbol{w}_T)] \leq \frac{LC\sqrt{\beta T}}{2n} \tag{10}$$

Proof (Sketch)

$$\begin{aligned}\frac{d}{dt}D_H(\pi_t||\pi_t') &= -\int_{\mathbb{R}^d}\frac{\partial}{\partial t}\sqrt{\pi\pi'}dw = -\frac{1}{4}\int_{\mathbb{R}^d}\sqrt{\pi\pi'}\left(\frac{1}{\beta}\|\boldsymbol{\nabla}\log\frac{\pi'}{\pi}\|^2 + \boldsymbol{\nabla}\log\frac{\pi}{\pi'}\cdot(\boldsymbol{\nabla}F_n - \boldsymbol{\nabla}F_n')\right)dw\\ &\leqslant \frac{1}{4}\int_{\mathbb{R}^d}\frac{\beta}{4}\sqrt{\pi\pi'}\|\boldsymbol{\nabla}F_n - \boldsymbol{\nabla}F_n'\|^2 dw \qquad \text{(Cauchy-Schwartz inequality)}\\ &\leqslant \frac{\beta L^2}{16n^2} \qquad (\text{as } \|\boldsymbol{\nabla}F_n(\boldsymbol{w}) - \boldsymbol{\nabla}F_n'(\boldsymbol{w})\| \leqslant \frac{L}{n})\end{aligned}$$

The first equality uses Fokker-Planck equation (8) and integral by parts. We integrate through time to get upper bound on $D_H(\pi||\pi')$. Lemma 7 and Theorem 5 then lead to the conclusion. ■

3.2. PAC-Bayesian Bounds

In this subsection, we consider the regularized ERM problem with regularization term $R(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{w}\|^2$. Assume the initial distribution γ of parameter $\boldsymbol{w}$ as $\mathcal{N}(0, \sigma_0^2 I)$, then we set $\lambda = \frac{1}{\beta\sigma_0^2}$ for technical reasons. The choice of λ makes $\lambda\boldsymbol{w}$ cancels out with $\frac{1}{\beta}\boldsymbol{\nabla}\log\gamma$ term exactly. Using similar techniques as the above subsection, we get:

Proposition 9 *Assume that $\ell(\boldsymbol{w}; z)$ is s-subGaussian with respect to $\gamma \times \mathcal{D}$. Suppose $M > 0$ satisfies $D_{KL}(\pi_T||\gamma) \leq M$ uniformly for worst-case data, then the following holds for Langevin equation with probability $1-\delta$:*

$$\mathrm{err}_{gen}(\boldsymbol{w}_T) \leq s\left(\frac{\beta}{2n}\int_0^T e^{-\frac{\lambda}{2}(T-t)}\mathbb{E}\left\|\boldsymbol{\nabla}\hat{\mathbb{E}}_n f(\boldsymbol{w}_t)\right\|^2 dt + \frac{\log 1/\delta + \log\log M}{n}\right)^{\frac{1}{2}} \tag{11}$$

Proof (Sketch)

$$\frac{d}{dt}D_{KL}(\pi_t||\gamma) = \int_{\mathbb{R}^d}\frac{\partial\pi}{\partial t}(\log\pi + 1 - \log\gamma)dw$$

$$= -\frac{1}{\beta}\int_{\mathbb{R}^d}\pi\|\boldsymbol{\nabla}\log\pi - \boldsymbol{\nabla}\log\gamma\|^2 dw - \int_{\mathbb{R}^d}\pi\langle\boldsymbol{\nabla}\hat{\mathbb{E}}_n f(\boldsymbol{w}), \boldsymbol{\nabla}\log\pi - \boldsymbol{\nabla}\log\gamma\rangle dw$$

$$\leq -\left(\frac{1}{\beta} - \frac{1}{2\beta}\right)\int_{\mathbb{R}^d}\pi\|\boldsymbol{\nabla}\log\pi - \boldsymbol{\nabla}\log\gamma\|^2 dw + \frac{\beta}{2}\int_{\mathbb{R}^d}\pi\|\boldsymbol{\nabla}\hat{\mathbb{E}}_n f(\boldsymbol{w})\|^2 dw \tag{12}$$

$$\leq -\frac{1}{2\beta\sigma_0^2}D_{KL}(\pi_t||\gamma) + \frac{\beta}{2}\int_{\mathbb{R}^d}\pi_t\|\boldsymbol{\nabla}\hat{\mathbb{E}}_n f(\boldsymbol{w})\|^2 dw \tag{13}$$

Inequalities (12) and (13) are obtained through Cauchy-Schwartz inequality and logarithmic Sobolev inequality (Gross, 1975) respectively. Then integrating w.r.t time T and combining with Theorem 6 lead to the conclusion. ■

4. Stability of Discrete-Time SGLD

Though the ideal continuous-time Langevin Equation attains small generalization error, that does not directly imply bounds for discrete-time SGLD algorithms. To relate discrete-time analyses with continuous-time ones, most previous works estimate the discretization gap, which usually results in at least linear dependence on d (Raginsky et al., 2017). In our analyses, we directly construct different SDEs that are similar to original Langevin Equation, for each discrete-time updates. This technique makes it possible to circumvent the potentially large gaps between discrete and continuous time algorithms, as we can see from this and the next section. The techniques are closely related to the continuous-time interpolation and KL calculation based on Girsanov theorem in (Dalalyan, 2017; Raginsky et al., 2017). However, directly applying their methods in the path space will lead to weaker $O(1/\sqrt{n})$ rate. Instead, we are comparing one-time marginal distributions. This offers flexibility for handling randomness from the random draw of stochastic gradients.

In this section, we will consider the stability of SGLD algorithm for non-convex objectives. For simplicity, we restrict our attention to the common choice of stochastic gradient, where one data point is used for each iteration, i.e., $\boldsymbol{g}_k(\boldsymbol{w}) = \boldsymbol{\nabla} f(\boldsymbol{w}; z_{i_k})$, where i_k is the index of randomly drawn training example. Our method also extends to other variants such as full gradients or mini-batch, which will be elaborated in the Appendix F.1. Assuming step sizes are not too large, we can achieve essentially the same rate as in the continuous-time case, without any additional dependence on dimension or norms. We also propose a method for dealing with large step sizes and get the main theorem for arbitrary choice of algorithmic parameters.

4.1. Estimating the Squared Hellinger Distance

Our proof strategy is induction on steps. For each step, we consider the conditional distributions of $\boldsymbol{w}_{k+1}, \boldsymbol{w}'_{k+1}$ conditioned on chosen index. If $i_k \neq i^*$, the gradient update is non-expansive, and the Gaussian noise does not increase squared Hellinger distance, either; If $i_k = i^*$, the amount of increase can be simply controlled by a constant from Lipschitz assumption. We can put them together to get an $O(\sqrt{T_N/n})$ upper bound, as shown in Appendix F.2. However, this does not achieve the $O(1/n)$ rate as in the continuous-time case, and is actually loose for small step sizes. This is because the Gaussian convolution step in the $i_k \neq i^*$ case makes $D_H(p_k||p'_k)$ decrease by a certain amount, which can align well with minus information-type term in $i_k = i^*$ case, and compensate the positive term to attain the fast rate. In Lemma 10, we give an upper bound for the squared Hellinger distance that achieves the fast rate.

Lemma 10 *Suppose for* $\forall k, \eta_k \leqslant \frac{\ln 2}{\beta L^2}$, *then there is*

$$\sqrt{D_H(p_N||p'_N)} \leq \frac{L}{n}\left(\beta \sum_k \eta_k\right)^{1/2} \tag{14}$$

Proof (Sketch)

It is easy to see the k-th update of SGLD is equivalent to the following step:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - (1-X)\eta_k \boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{j_k}) - X\eta_k \boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{i_*}) + \mathcal{N}(0, \frac{2\eta_k}{\beta} I_d),$$
$$\boldsymbol{w}_k \sim p_k, \quad j_k \sim \mathcal{U}(\{1, 2, \cdots, n\} \setminus \{i_*\})$$

where $\boldsymbol{w}_k, j_k, X$ are independent and $\mathcal{P}(X=1) = \frac{1}{n}, \mathcal{P}(X=0) = \frac{n-1}{n}$.

Now, we consider a family of random variables $\boldsymbol{\theta}_t$ $(0 \leq t \leq \eta_k)$ defined by

$$\boldsymbol{\theta}_t = \boldsymbol{w}_k - \eta_k \boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{j_k}) - Xt(\boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{i_*}) - \boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{j_k})) + \mathcal{N}(0, \frac{2t}{\beta} I_d) \tag{15}$$

Denote the p.d.f of $\boldsymbol{\theta}_t$ as $\pi(\boldsymbol{x}, t)$. Similarly, we also define $\boldsymbol{\theta}'_t, \pi'(\boldsymbol{x}, t)$. We can check $\pi(\boldsymbol{x}, t)$ satisfies the following PDE with $\hat{\boldsymbol{g}}(\boldsymbol{w}) = \mathbb{E}_{X, \boldsymbol{w}_k, j_k}[X(\boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{i_*}) - \boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{j_k}))|\boldsymbol{\theta}_t = \boldsymbol{w}]$:

$$\frac{\partial \pi}{\partial t} = \frac{1}{\beta} \triangle \pi + \boldsymbol{\nabla} \cdot (\pi \hat{\boldsymbol{g}}) \tag{16}$$

Similarly, we also have $\hat{\boldsymbol{g}}'$ for π'.

Although $\hat{\boldsymbol{g}} - \hat{\boldsymbol{g}}'$ is not necessarily pointwise bounded by $O(\frac{1}{n})$ as in the continuous-time case, we can prove a bound of order $O(\frac{1}{n})$ for weighted average:

$$\int_{\mathbb{R}^d} \sqrt{\pi\pi'} \|\hat{\boldsymbol{g}} - \hat{\boldsymbol{g}}'\|^2 dw \leq \frac{4\sqrt{2}L^2}{(n-1)^2} \tag{17}$$

Then as in previous analysis, we compute the time derivative of squared Hellinger distance:

$$\begin{aligned} \frac{d}{dt} D_H(\pi_t || \pi'_t) &= -\frac{1}{4} \int_{\mathbb{R}^d} \sqrt{\pi\pi'} \left(\frac{1}{\beta} \|\boldsymbol{\nabla} \log \frac{\pi'}{\pi}\|^2 + \boldsymbol{\nabla} \log \frac{\pi}{\pi'} \cdot (\hat{\boldsymbol{g}}_t - \hat{\boldsymbol{g}}'_t) \right) dw \\ &\leq \frac{\beta}{8} \int_{\mathbb{R}^d} \sqrt{\pi\pi'} \|\hat{\boldsymbol{g}} - \hat{\boldsymbol{g}}'\|^2 dw < \frac{\beta L^2}{n^2} \end{aligned}$$

So we have

$$D_H(p_{k+1} || p'_{k+1}) = D_H(\pi_{\eta_k} || \pi'_{\eta_k}) \leq D_H(\pi_0 || \pi'_0) + \frac{\beta L^2}{n^2} \eta_k \leq D_H(p_k || p'_k) + \frac{\beta L^2}{n^2} \eta_k \tag{18}$$

Then one arrives at the statement by induction. ■

4.2. Dealing with Large Step Sizes

Lemma 10 requires an upper bound $\frac{\ln 2}{\beta L^2}$ on the step sizes. Though this is a mild requirement, it is sometimes not satisfied for the first few steps, when we are using decreasing step sizes. In this situation, however, a large step will make $\beta L^2 \eta_k = \Omega(1)$ contribution to the sum, which is also undesirable. On the other hand, a stochastic gradient step can change a distribution within at most $O(1/n)$ scale in terms of L^1 distance. So for larger steps, a rough estimate based on L^1 distance can be better. As step size changes, the best method of estimation may be different.

In this subsection, we describe a general framework for concatenating two stability bounds together, as well as a simple bound that tames the large steps.

Theorem 11 *For two biconvex and non-expansive bivariate-functionals $D_A(\cdot||\cdot)$ and $D_B(\cdot||\cdot)$ that controls the stability, i.e., $\varepsilon_n \leq A_\ell D_A(p_N||p'_N) \wedge B_\ell D_B(p_N||p'_N)$, for constants A_ℓ, B_ℓ depending only on ℓ. For SGLD with step sizes $\eta_1, \cdots, \eta_N$, if we can estimate D_A and D_B by $D_A(p_N||p'_N) \leq h_A(\eta_1, \cdots, \eta_N), \quad D_B(p_N||p'_N) \leq h_B(\eta_1, \cdots, \eta_N)$, then we have:*

$$\epsilon_n \leq A_\ell h_A(\eta_1, \cdots, \eta_k) + B_\ell h_B(\eta_{k+1}, \cdots, \eta_N), \quad \forall k \in \{1, 2 \cdots, N-1\}. \tag{19}$$

For large step sizes, we can easily obtain a stability bound based on L^1 distance, assuming the loss function $\ell(\cdot, \cdot)$ is bounded.

$$\epsilon_n = \sup_z \left| \int \ell(\boldsymbol{w}; z)(p_N - p'_N) dw \right| \leqslant \sup \|\ell\|_{L^\infty} \int |p_N - p'_N| dw \tag{20}$$

The L^1 distance can be further upper bounded by number of iterations:

Lemma 12 *For an SGLD algorithm running k_0 iterations, there is $d_{TV}(p_{k_0}||p'_{k_0}) \leq \frac{k_0}{n}$.*

Applying the framework in Theorem 11 with $k_0 \triangleq \sup\{k : \eta_k > \frac{\ln 2}{\beta L^2}\}$, we obtain the final result.

Theorem 13 *Consider N rounds of SGLD with parameters β and $\{\eta_i\}$. Suppose the loss function $\ell(\boldsymbol{w}; z)$ is uniformly bounded by C, and $\forall z, z'$, there is $\|\boldsymbol{\nabla} f(\boldsymbol{w}; z) - \boldsymbol{\nabla} f(\boldsymbol{w}; z')\| \leq L$. Let $k_0 = \sup\{k : \eta_k > \frac{\ln 2}{\beta L^2}\}$, we have the following generalization bound in expectation*

$$\mathbb{E}[\mathrm{err}_{gen}(\boldsymbol{w}_N)] \leq \frac{2k_0}{n} + \frac{2LC}{n} \left(\beta \sum_{i=k_0+1}^{N} \eta_i \right)^{1/2} \tag{21}$$

5. PAC-Bayesian Theory for Discrete-Time SGLD

In this section, we present a non-asymptotic analysis for the generalization error of SGLD using PAC-Bayesian theory. We use an ℓ_2 regularization term $R(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{w}\|^2$, so that $\tilde{\boldsymbol{g}}_k(\boldsymbol{w}) = \boldsymbol{g}_k(\boldsymbol{w}) + \lambda \boldsymbol{w}$, which has been shown to be helpful in the continuous time case. As in the previous section, we directly construct stochastic processes and corresponding PDEs based on the discrete-time updates, instead of estimating the discretization gap. However, the uniform way of interpolating the stochastic process will lead to a conditional expectation term $\mathbb{E}\left[\boldsymbol{\theta}_0 | \boldsymbol{\theta}_t = \boldsymbol{w}\right]$, which cannot cancel perfectly with $\boldsymbol{w}$. Therefore, we construct the stochastic process in a non-uniform way, using Ornstein-Uhlenbeck process. We also need the prior γ_k to vary with k in a data-independent way for technical reasons. In this section, we allow $\boldsymbol{g}_k$ to be any estimator for the gradient, since our proof essentially relies upon the norm for each stochastic gradient, instead of how it is calculated.

5.1. Constructing the PDEs

The following theorem relates discrete-time updates with a PDE:

Theorem 14 *Starting from $\boldsymbol{\theta}_0 \sim \pi_0$, for fixed mapping $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$ and $\forall t \in [0, \tau_k]$, let*

$$\boldsymbol{\theta}_t = e^{-\lambda t}\boldsymbol{\theta}_0 - \frac{1 - e^{-\lambda t}}{\lambda}\boldsymbol{g}(\boldsymbol{\theta_0}) + \mathcal{N}\left(0, \frac{1 - e^{-2\lambda t}}{\beta'_k \lambda} I_d\right). \tag{22}$$

The pdf π_t of $\boldsymbol{\theta}_t$ satisfies the following PDE:

$$\frac{\partial \pi}{\partial t}(\boldsymbol{w}) = \frac{1}{\beta'_k}\Delta\pi(\boldsymbol{w}) + \boldsymbol{\nabla}\cdot(\lambda\pi(\boldsymbol{w})\boldsymbol{w}) + \boldsymbol{\nabla}\cdot\left(\pi(\boldsymbol{w})\mathbb{E}\left[\boldsymbol{g}(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_t = \boldsymbol{w}\right]\right) \tag{23}$$

With the presence of ℓ_2 regularization term, we use an Ornstein-Uhlenbeck process and integrate with respect to the initial distribution for each discrete step, which is different from the differential equation construction using Brownian motion with constant drifts in previous sections.

The gradient update in Theorem 14 can be related to standard SGLD step as follows:

$$\begin{cases} \eta_k & = \frac{1-e^{-\lambda\tau_k}}{\lambda} \\ \sqrt{\frac{2\eta_k}{\beta}} & = \sqrt{\frac{1-e^{-2\lambda\tau_k}}{\beta'_k\lambda}} \end{cases} \implies \begin{cases} \tau_k = & -\frac{1}{\lambda}\ln(1-\eta_k\lambda) \\ \beta'_k = & \left(1 - \frac{\lambda\eta_k}{2}\right)\beta \end{cases} \tag{24}$$

Using this transformation of parameters, conditioned on the choice of $\boldsymbol{g}_k(\cdot)$, the final distribution π_{τ_k} in Theorem 14 is exactly the same with output distribution of SGLD update

$$\boldsymbol{w}_{k+1} = (1-\lambda\eta_k)\boldsymbol{w}_k - \eta_k\boldsymbol{g}_k(\boldsymbol{w}_k) + \sqrt{\frac{2\eta_k}{\beta}}\mathcal{N}(0, I_d) \tag{25}$$

In Section 3.2, we require the regularization parameter λ to be exactly equal to $\frac{1}{\beta\sigma_0^2}$. However, in our construction, β'_k can vary according to η_k, making it impossible to fit with fixed parameter λ. In order to handle this technical issue, we allow the prior distribution to change in a data-independent way during iterations, and let prior at k-th round be γ_k. (Note that PAC-Bayes theorem is still valid, since the prior is fixed and data-independent for any fixed k) To exactly cancel out the difference induced by mismatch between regularization parameter and β'_k, we construct a continuous time prior $\tilde{\gamma}$ satisfying the following PDE: (in our notation, we use $\tilde{\gamma}_t$ and $\tilde{\sigma}_t^2$ to denote the prior in continuous time process and its variance, while γ_k and σ_k^2 denote discrete time steps).

$$\frac{\partial\tilde{\gamma}}{\partial t} = \frac{1}{\beta'_k}\Delta\tilde{\gamma} + \boldsymbol{\nabla}\cdot(\lambda\tilde{\gamma}\boldsymbol{w}), \quad t \in [0, \tau_k] \tag{26}$$

It is easy to prove by induction that $\tilde{\gamma}$ is isotropic Gaussian. Let $\tilde{\gamma}_t = \mathcal{N}(0, \tilde{\sigma}_t^2 I_d)$, we have:

$$\tilde{\sigma}_t^2 = \begin{cases} e^{-2\lambda t}\tilde{\sigma}_0^2 + \frac{1-e^{-2\lambda t}}{\beta'_k\lambda}, & \lambda > 0 \\ \tilde{\sigma}_0^2 + \frac{t}{\beta'_k}, & \lambda = 0 \end{cases} \xRightarrow{t=\tau_k} \quad \sigma_{k+1}^2 = \begin{cases} e^{-2\lambda\tau_k}\sigma_k^2 + \frac{1-e^{-2\lambda\tau_k}}{\beta'_k\lambda}, & \lambda > 0 \\ \sigma_k^2 + \frac{\tau_k}{\beta'_k}, & \lambda = 0 \end{cases} \tag{27}$$

Putting them together, we are ready to cancel out the $\boldsymbol{w}$ term in upper bound for KL divergence.

5.2. Estimating the KL Divergence

In this section, we present an upper bound on the KL divergence $D_{KL}(p_k||\gamma_k)$ based on the interpolation in previous section, which leads to the final generalization bound. We first give the following estimate for one-step SGLD update.

Lemma 15 *Consider an SGLD update for regularized ERM with transformed parameters (τ_k, β'_k), and let prior $\tilde{\sigma}_t$ be defined above. We have the following inequality:*

$$D_{KL}\left(p_{k+1}\middle|\middle|\gamma_{k+1}\right) \leq e^{-\frac{\tau_k}{2b_k}}D_{KL}\left(p_k\middle|\middle|\gamma_k\right) + \frac{\beta'_k\tau_k}{2}\mathbb{E}\|\boldsymbol{g}_k(\boldsymbol{w}_k)\|^2 \tag{28}$$

where $b_k = \max\left(\sigma_{k-1}^2\beta'_k, \frac{1}{\lambda}\right)$ for $\lambda > 0$, and $b_k = \sigma_{k-1}^2\beta'_k + \tau_k$ for $\lambda = 0$.

Using Lemma 15 iteratively, we can obtain the KL divergence upper bound for the whole SGLD algorithm, which is stated in the following theorem: (We actually state a special case, with the most general version postponed to the Appendix)

Theorem 16 *Given algorithmic parameters $N, \{\eta_k\}, \beta, \sigma_0, \lambda$ fixed. Assume that $\sigma_0^2 \leq \frac{3}{2\beta\lambda}$, and loss function $\ell(w; z)$ is s-subGaussian with respect to distribution $\mathcal{N}(0, \sigma^2 I_d) \times \mathcal{D}$ for any $\sigma^2 \in \left(0, \frac{3}{2\beta\lambda}\right)$. Assume that $f(\boldsymbol{w}; z_i)$ is uniformly L-Lipschitz with respect to $\boldsymbol{w}$. Assume that $\eta_k\lambda < \frac{1}{2}, \forall k$. The following inequality uniformly holds for SGLD with probability $1 - \delta$: (with respect to random draw of training data)*

$$\mathrm{err}_{gen}(\boldsymbol{w}_N) \leq 2s \left(\frac{\beta}{n} \sum_{k=1}^{N} \eta_k e^{-\frac{\lambda}{3}(T_N - T_k)} \mathbb{E}\left[\|\boldsymbol{g}_k(\boldsymbol{w}_k)\|^2 \right] + \frac{\log N/\delta + \log\log NL}{n} \right)^{\frac{1}{2}} \tag{29}$$

Though having a slower $O(1/\sqrt{n})$ rate compared with stability results, Theorem 16 makes milder tail assumptions and achieves high-probability bounds. More importantly, the bound itself has several advantages, which could be helpful for large model classes such as deep neural networks:

- The uniform Lipschitz constant is replaced with norms of actual gradients $\mathbb{E}\|\boldsymbol{g}_k(\boldsymbol{w}_k)\|^2$ along optimization trajectory (the expectation is taken only with the randomized algorithm but not with data). The bound only has doubly logarithmic dependence on L. As L usually depends on range of data and parameters, it can be large. However, the gradient themselves should not be large, or the optimization trajectories will be unreliable.

- The time-decaying factor $e^{-\frac{\lambda}{3}(T_n - T_k)}$ eliminates effect of earlier gradients, which could be much larger than the last few ones. Furthermore, when ℓ_2 regularization is imposed on a Lipschitz function, the bound will be finite when $T \to \infty$, as SGLD will not go too far away with the presence of ℓ_2 regularization.

6. Conclusion

In this paper, we study the problem of non-convex (regularized) ERM with Stochastic Gradient Langevin Dynamics, from the perspective of statistical learning theory. Algorithm-dependent generalization bounds are established using uniform stability and PAC-Bayesian theory, respectively. For stability-based results, we get a generalization error bound of $O\left(\frac{1}{n}(k_0 + L\sqrt{\beta \sum \eta_i})\right)$, where k_0 is the largest index k with $\eta_k \beta L^2 > \ln 2$. This bound attains $O(1/n)$ fast rate and only depends on Lipschitz constant L and aggregated step sizes. For PAC-Bayesian theory with $\frac{\lambda}{2}\|w\|^2$ regularization, we get a generalization bound of $O\left(\sqrt{\frac{\beta}{n} \sum \eta_k \mathbb{E}\|\boldsymbol{g}_k\|^2 \exp(-\frac{\lambda}{3}(T_N - T_k))}\right)$, in which the contribution of each step decays exponentially. In addition to time-decaying effect, this bound also replaces the uniform Lipschitz constant with expected gradient norms along trajectory. Our bounds have no explicit dependence on dimension or norms. This is the first algorithm-dependent generalization bound for non-convex ERM with polynomial dependence on aggregated step sizes and smoothness properties of objective function. Our theoretical results provide potential explanations for generalization performance of large non-convex models such as deep neural networks, and emphasizes the merits of Gaussian noise for non-convex learning problems.

Acknowledgments

We thank Zhou Lu, Feicheng Wang and Xiang Wang for helpful discussions. This work was partially supported by National Basic Research Program of China (973 Program) (grant no. 2015CB352502), NSFC (61573026) and Center for Data Science, Beijing Institute of Big Data Research in Peking University. This work was done when X.Z. was visiting Peking University.

References

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

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *arXiv preprint arXiv:1507.02564*, 2015.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, and Yann LeCun. Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016.

Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*, 2016.

Xiang Cheng and Peter Bartlett. Convergence of langevin mcmc in kl-divergence. *arXiv preprint arXiv:1705.09048*, 2017.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

Imre Csiszár, Paul C Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 651–676, 2017.

Arnak S Dalalyan and Alexandre B Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. *Journal of Computer and System Sciences*, 78(5):1423–1443, 2012.

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

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. Pac-bayesian theory meets bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016.

Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

István Gyöngy. Mimicking the one-dimensional marginal distributions of processes having an itô differential. *Probability theory and related fields*, 71(4):501–516, 1986.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

Chris Junchi Li, Lei Li, Junyang Qian, and Jian-Guo Liu. Batch size matters: A diffusion approximation framework on nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

Qianxiao Li, Cheng Tai, and E Weinan. Dynamics of stochastic gradient algorithms. *arXiv preprint arXiv:1511.06251*, 2015.

Junhong Lin and Lorenzo Rosasco. Optimal learning for multi-pass stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pages 4556–4564, 2016.

Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization for multiple passes sgm. In *International Conference on Machine Learning*, pages 2340–2348, 2016.

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

Peter A Markowich and Cédric Villani. On the trend to equilibrium for the fokker-planck equation: an interplay between physics and functional analysis. *Mat. Contemp*, 19:1–29, 2000.

David A McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

Tigran Nagapetyan, Andrew B Duncan, Leonard Hasenclever, Sebastian J Vollmer, Lukasz Szpruch, and Konstantinos Zygalakis. The true cost of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1706.02692*, 2017.

Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. *arXiv preprint arXiv:1801.04295*, 2018.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

Alexander Rakhlin, Sayan Mukherjee, and Tomaso Poggio. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

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

Yuting Wei, Fanny Yang, and Martin J Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. In *Advances in Neural Information Processing Systems*, pages 6067–6077, 2017.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2521–2530, 2017.

Nanyang Ye, Zhanxing Zhu, and Rafal K Mantiuk. Langevin dynamics with continuous tempering for training deep neural networks. *arXiv preprint arXiv:1703.04379*, 2017.

Chiyuan Zhang, Qianli Liao, Alexander Rakhlin, Karthik Sridharan, Brando Miranda, Noah Golowich, and Tomaso Poggio. Theory of deep learning iii: Generalization properties of sgd. Technical report, Center for Brains, Minds and Machines (CBMM), 2017a.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1702.05575*, 2017b.

Appendix A. Additional Related Works

Deliberate injection of Gaussian noise has become a rising star in the literature of non-convex optimization. Ge et al. (2015); Jin et al. (2017) show that Gaussian noise helps SGD escape 2nd order saddle points efficiently. Stochastic Gradient Langevin Dynamics, proposed as discrete version of Langevin Equation $dw_t = -\nabla F(\boldsymbol{w}_t)dt + \sqrt{\frac{2}{\beta}}d\boldsymbol{B}_t$, also plays an important role in optimization and sampling. It is well-known that Langevin Equation asymptotically converges to equilibrium distribution $p(\boldsymbol{w}) \propto e^{-\beta F(\boldsymbol{w})}$, see e.g. (Markowich and Villani, 2000). This property has been utilized for posterior sampling, known as Langevin Monte Carlo. The discretization error and mixing time are intensively studied by Bubeck et al. (2015); Nagapetyan et al. (2017), for log-concave distributions. Dalalyan and Tsybakov (2012) also used Langevin MC to approximate Exponential Weighted Aggregate, and proved PAC-Bayesian bounds for regression learning with sparsity prior. For non-convex learning and optimization, Raginsky et al. (2017) makes the first attempt towards excess risks by non-convex SGLD, combining algorithmic convergence and generalization error. But their results are based on convergence to equilibrium, which relies upon constants in Poincaré Inequality, leading to inevitably exponential dependence on dimension. Though the mixing time can be prohibitive in non-convex case, Zhang et al. (2017b) recently show that hitting time of SGLD for small-loss region can be much better, and the Gaussian noise in SGLD helps to escape shallow local minima. Their results also emphasize the importance of generalization guarantees for discrete-time non-asymptotic SGLD in non-convex settings.

Besides, several recent works also studied the connection between SGD and stochastic differential equations, such as SME (Li et al., 2015, 2017). Though our results for SGLD cannot directly extend to their SDEs with data-dependent diffusion term, our methods are potentially applicable for generalization error bounds in their settings.

Appendix B. Illustration about Why Gaussian Noise Helps

In this section, we will first illustrate why prior analyses on stability can be very large for non-convex objective function, and how this can be overcome by adding Gaussian noise. This important observation motivates our analysis based on KL-Divergence and Hellinger distances, which highlights the effect of smooth distributions on generalization error bounds.

Stability-based analysis for gradient algorithms on non-convex losses will suffer from a ”fence-sitting” situation, as illustrated in Figure 1. Consider a non-convex empirical loss surface with two local minima, which is divided into two regions by a ridge. If $\boldsymbol{w}_k$ lies on one side of this ridge, a noiseless first-order method will lead to the local minimum on this side. However, if $\boldsymbol{w}_k$ comes close to the ridge in its trajectory, small shift on the loss surface caused by changing one point will lead it to a completely different local minimum, as we can see from the figure.

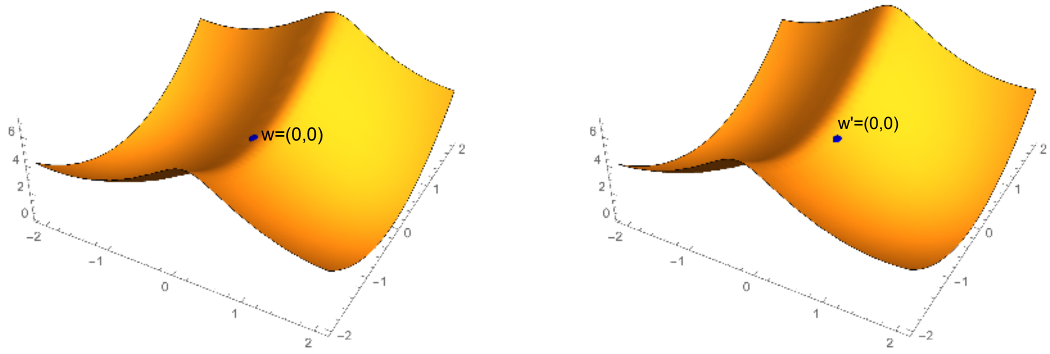

Figure 1: Illustration of ”Fence-Sitting” Situation for Stability of Non-convex Optimization

To guarantee stability, we need $\boldsymbol{w}_k$ to randomly decide which side to go when it comes close to the ridge. The noise needs to be isotropic and smooth enough in order to cross this ridge, as the direction of variation can be quite arbitrary. SGLD successfully tackles the fence-sitting problem by smoothing the probability of going either side, and adding noise to subsequent steps to avoid unstable shallow local minima. The bounds for SGD in Hardt et al. (2015) also exploits randomness of choosing i_k, but the noise is not smooth enough. So their bound requires the subsequent steps to be very small, to keep w_k not far from the ridge.

Appendix C. Backgrounds on Fokker-Planck Equation

It is known that that the movement of a particle in the d-dimensional space influenced by its current state and random forces (here we only consider a simple case), can be characterized by the following stochastic differential equation (SDE):

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t, t)dt + \sqrt{2\beta^{-1}}d\boldsymbol{B}_t \tag{30}$$

where $\boldsymbol{X}_t$ is the random position of the particle at time t, $\boldsymbol{\mu}(\boldsymbol{X}_t, t)$ is the d-dimensional random drift vector, and B_t is the d dimensional Brownian motion. Denote the density function of $\boldsymbol{X}_t$ as $p(\boldsymbol{x}, t)$, then Fokker-Planck equation describes the evolution of $p(\boldsymbol{x}, t)$:

$$\frac{\partial p(\boldsymbol{x}, t)}{\partial t} = \frac{1}{\beta}\Delta p(\boldsymbol{x}, t) - \boldsymbol{\nabla} \cdot (p(\boldsymbol{x}, t)\boldsymbol{\mu}(\boldsymbol{x}, t)) \tag{31}$$

where Δ is the Laplace operator.

For Gaussian distribution, we have the following log-Sobolev inequality, which relates Fisher information and KL divergence.

Theorem 17 *For $\gamma = \mathcal{N}(0, \sigma_0^2 I)$ and any distribution π which has absolute continuous density, we have:*

$$\mathbb{E}_\pi \left(\log \frac{\pi}{\gamma} \right) \leq \sigma_0^2 \mathbb{E}_\pi \left\| \boldsymbol{\nabla} \log \frac{\pi}{\gamma} \right\|^2 \tag{32}$$

A special case of Langevin equation is Ornstein-Uhlenbeck process, which plays a critical role in our discretization construction.

Proposition 18 *An Ornstein-Uhlenbeck process is solution to the following SDE:*

$$d\boldsymbol{X}_t = \lambda(\boldsymbol{b} - \boldsymbol{X}_t)dt + \sqrt{\frac{2}{\beta}} d\boldsymbol{B}_t, \tag{33}$$

for some constant vector $\boldsymbol{b} \in \mathbb{R}^d$ and constant $\lambda > 0$. Its Fokker-Planck equation is:

$$\frac{\partial \pi}{\partial t} = \frac{1}{\beta} \Delta \pi + \lambda \boldsymbol{\nabla} \cdot ((\boldsymbol{w} - \boldsymbol{b})\pi), \tag{34}$$

and the solution can be directly written as:

$$\boldsymbol{X}_t = e^{-\lambda t} \boldsymbol{X}_0 + \left(1 - e^{-\lambda t}\right) \boldsymbol{b} + \sqrt{\frac{2}{\beta}} \int_0^t e^{-\lambda(t-s)} d\boldsymbol{B}_s \tag{35}$$

The proof can be found in any standard textbook about Fokker-Planck equations, see, e.g. (Risken, 1996)

Appendix D. Omitted Proofs in Section 2

Proof of theorem 6

Proof For simplicity, we replace $\ell(\boldsymbol{w}, z)$ with $\ell(\boldsymbol{w}, z) - \mathbb{E}_{\mathcal{D} \times \mathcal{P}} \ell(\boldsymbol{w}, z)$, and assume the distribution of loss function is centered under data distribution and the prior. It is easy to check that this modification does not affect following analysis.

We use the Donsker-Varadhan change of measure inequality: for any pair of distributions $\mathcal{P}$ and $\mathcal{Q}$ and functional ϕ, we have

$$\mathbb{E}_{\mathcal{Q}}(\phi(\ell)) \leq D_{KL}(\mathcal{Q} || \mathcal{P}) + \ln \mathbb{E}_{\mathcal{P}} \left(e^{\phi(\ell)} \right) \tag{36}$$

Consider functions $\phi(\ell)$ in the form of $\phi(w) = \lambda \left(\mathbb{E}\ell(w; z) - \hat{\mathbb{E}}_n \ell(w; z) \right)$ (function class of $\ell(w, \cdot)$ indexed by w), while the values of λ will be determined later. (The notation $\hat{\mathbb{E}}_n$ denotes empirical expectation, i.e., $\hat{\mathbb{E}}_n h(x) = \frac{1}{n} \sum_{i=1}^n h(x_i)$)

For any fixed $\lambda > 0, \delta' > 0$, by Markov inequality we have the following with probability $1 - \delta'$

$$\mathbb{E}_{\mathcal{P}} \left(e^{\lambda \left(\mathbb{E}\ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z) \right)} \right) \leq \frac{1}{\delta'} \mathbb{E}_S \mathbb{E}_{\mathcal{P}} \left(e^{\lambda \left(\mathbb{E}\ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z) \right)} \right) \tag{37}$$

For any finite set $\Lambda \subseteq \mathbb{R}^+, \delta > 0$, let $\delta' = \frac{\delta}{|\Lambda|}$, we have:

$$\forall \lambda \in \Lambda, \quad \mathbb{P}\left(\mathbb{E}_{\mathcal{P}}\left(e^{\lambda\left(\mathbb{E}\ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right)}\right) > \frac{|\Lambda|}{\delta}\mathbb{E}_S \mathbb{E}_{\mathcal{P}}\left(e^{\lambda\left(\mathbb{E}\ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right)}\right)\right) \leq \frac{\delta}{|\Lambda|} \tag{38}$$

and by union bound,

$$\mathbb{P}\left(\exists \lambda \in \Lambda, \mathbb{E}_{\mathcal{P}}\left(e^{\lambda\left(\mathbb{E}\ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right)}\right) > \frac{|\Lambda|}{\delta}\mathbb{E}_S \mathbb{E}_{\mathcal{P}}\left(e^{\lambda\left(\mathbb{E}\ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right)}\right)\right) \leq |\Lambda|\delta' = \delta \tag{39}$$

Let $S' = \{x'_1, x'_2, \cdots, x'_n\}$ be an independent copy of n samples, and let $\hat{\mathbb{E}}'_n$ denotes empirical expectation with respect to S', i.e., $\hat{\mathbb{E}}'_n h(x) = \frac{1}{n}\sum_{i=1}^n h(x'_i)$. We have

$$\mathbb{E}_S\left(e^{\lambda\left(\mathbb{E}\ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right)}\right) = \mathbb{E}_S\left(e^{\mathbb{E}_{S'}\lambda\left(\hat{\mathbb{E}}'_n \ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right)}\right) \leq \mathbb{E}_{S,S'}\left(e^{\lambda\left(\hat{\mathbb{E}}'_n \ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right)}\right) \tag{40}$$

The last inequality is due to convexity of exponential function and Jensen's Inequality.

Given λ fixed, we can expand the right hand side based on independence, and each term is upper bounded by $\exp\left(\frac{\lambda^2 s^2}{n^2}\right)$ by subGaussian assumption. Putting them together, we have:

$$\mathbb{E}_{S,S',\mathcal{P}}\left(e^{\lambda\left(\hat{\mathbb{E}}'_n \ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right)}\right) = \prod_{i=1}^n \mathbb{E}\left(e^{\frac{\lambda}{n}\left(\ell(w;z'_i) - \ell(w;z_i)\right)}\right) \leq e^{\frac{\lambda^2 s^2}{n}} \tag{41}$$

Combining two inequalities above, we have the following bound:

$$\mathbb{E}_{S,\mathcal{P}}\left(e^{\lambda\left(\mathbb{E}\ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right)}\right) \leq e^{\frac{\lambda^2 s^2}{n}} \tag{42}$$

Using Equation (38), we have the following with probability $1 - \delta$:

$$\forall \lambda \in \Lambda, \quad \mathbb{E}_{\mathcal{P}}\left(e^{\lambda\left(\mathbb{E}\ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right)}\right) \leq \frac{|\Lambda|}{\delta} e^{\frac{\lambda^2 s^2}{n}} \tag{43}$$

Combined with Equation (36) by letting $\phi_\lambda(w) = \lambda\left(\mathbb{E}\ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right)$ for different values of $\lambda \in \Lambda$, for any posterior $\mathcal{Q}$, we have the following result with probability $1 - \delta$:

$$\forall \lambda \in \Lambda, \quad \mathbb{E}_{\mathcal{Q}}\left(\mathbb{E}\ell(w;z) - \hat{\mathbb{E}}_n \ell(w;z)\right) \leq \frac{1}{\lambda}\left(D_{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{|\Lambda|}{\delta}\right) + \frac{\lambda s^2}{n} \tag{44}$$

Take $\Lambda = \left\{\frac{1}{s}\sqrt{n\left(2^i + \log\frac{1}{\delta} + \log\log M\right)}\right\}_{i=1}^{\lceil \log M \rceil}$ with $|\Lambda| = \lceil \log M \rceil$. For any posterior $\mathcal{Q} \in \Xi$, $D_{KL}(\mathcal{Q}||\mathcal{P}) \leq M$. Choose the index $i \in \{1, 2, \cdots, \lceil \log M \rceil\}$ such that $2^i \leq D_{KL}(\mathcal{Q}||\mathcal{P}) < 2^{i+1}$ (if $D_{KL}(\mathcal{Q}||\mathcal{P}) < 2$, let $i = 1$) and plug the corresponding value of λ into Equation (44), we can easily get the following upper bound for the right hand side:

$$\frac{1}{\lambda}\left(D_{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{|\Lambda|}{\delta}\right) + \frac{\lambda s^2}{n} \leq 2s\sqrt{\frac{D_{KL}(\mathcal{Q}||\mathcal{P}) \vee 1 + \log\frac{1}{\delta} + \log\log M}{n}}, \quad \forall \mathcal{Q} \in \Xi \tag{45}$$

So the theorem is proven. ■

Remark: if we choose a single value of λ fixed, the proof becomes the same as Germain et al. (2016), which is based on Donsker-Varadhan change-of-measure inequality. But their bound does not give optimal dependence on KL divergence. In order to overcome this difficulty, we use a set of values for λ and union bound to obtain the $\sqrt{\frac{D_{KL}}{n}}$ bound, at a price of double logarithmic term.

Appendix E. Omitted Proofs in Section 3

Proof of Lemma 7

Proof

$$\begin{aligned}\epsilon_n =& \sup_{z,S,S'} \left| \int_{\mathbb{R}^d} \ell(\boldsymbol{w}; z)\pi(\boldsymbol{w})dw - \int_{\mathbb{R}^d} \ell(\boldsymbol{w}; z)\pi'(\boldsymbol{w})dw \right| \\ =& \sup_{z,S,S'} \left| \int_{\mathbb{R}^d} \ell(\boldsymbol{w}; z) \left(\sqrt{\pi} + \sqrt{\pi'}\right) \left(\sqrt{\pi} - \sqrt{\pi'}\right) dw \right| \\ \leq& \sup \left\{ \left(\int_{\mathbb{R}^d} \ell(\boldsymbol{w}; z)^2 \left(\sqrt{\pi} + \sqrt{\pi'}\right)^2 dw \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \left(\sqrt{\pi} - \sqrt{\pi'}\right)^2 dw \right)^{\frac{1}{2}} \right\} \\ =& 2 \sup_{\pi} \|\ell\|_{L^2(\pi)} \sqrt{D_H(\pi||\pi')} \\ \leq& 2C\sqrt{D_H(\pi||\pi')}\end{aligned} \tag{46}$$

■

Proof of Proposition 8

Proof According to the analysis above, we only need to bound $D_H(\pi||\pi')$ from above.

Apparently, at time $t = 0$, $D_H(\pi||\pi') = 0$. We then estimate $\frac{d}{dt}D_H(\pi_t||\pi'_t)$:

$$\begin{aligned}\frac{d}{dt}D_H(\pi_t||\pi'_t) =& -\int_{\mathbb{R}^d} \frac{\partial}{\partial t}\sqrt{\pi\pi'}dw \\ =& -\int_{\mathbb{R}^d} \frac{\sqrt{\pi'}}{2\sqrt{\pi}}\frac{\partial \pi}{\partial t}dw - \int_{\mathbb{R}^d} \frac{\sqrt{\pi}}{2\sqrt{\pi'}}\frac{\partial \pi'}{\partial t}dw \\ =& -\int_{\mathbb{R}^d} \frac{\sqrt{\pi'}}{2\sqrt{\pi}}\left(\frac{1}{\beta}\Delta\pi + \boldsymbol{\nabla}\cdot(\pi\boldsymbol{\nabla}F_n)\right)dw - \int_{\mathbb{R}^d} \frac{\sqrt{\pi}}{2\sqrt{\pi'}}\left(\frac{1}{\beta}\Delta\pi' + \boldsymbol{\nabla}\cdot(\pi'\boldsymbol{\nabla}F'_n)\right)dw \\ =& \frac{1}{2}\int_{\mathbb{R}^d} \boldsymbol{\nabla}\frac{\sqrt{\pi'}}{\sqrt{\pi}}\cdot\left(\frac{1}{\beta}\nabla\pi + \pi\boldsymbol{\nabla}F_n\right)dw + \frac{1}{2}\int_{\mathbb{R}^d} \boldsymbol{\nabla}\frac{\sqrt{\pi}}{\sqrt{\pi'}}\left(\frac{1}{\beta}\boldsymbol{\nabla}\pi' + \pi'\boldsymbol{\nabla}F'_n)\right)dw\end{aligned} \tag{47}$$

The last equality is due to integration by parts. Technical conditions such as uniform decaying tails of π and π' can be found in (Risken, 1996). We then proceed to calculate the part induced by gradient update (with coefficient 1) and those induced by Gaussian convolution (with coefficient $\frac{1}{\beta}$)

individually, which can be described as follows:

$$\begin{aligned}\frac{d}{dt}D_H(\pi_t||\pi_t') &= \frac{1}{2}\int_{\mathbb{R}^d}\boldsymbol{\nabla}\frac{\sqrt{\pi'}}{\sqrt{\pi}}\cdot\left(\frac{1}{\beta}\nabla\pi+\pi\boldsymbol{\nabla}F_n\right)dw+\frac{1}{2}\int_{\mathbb{R}^d}\boldsymbol{\nabla}\frac{\sqrt{\pi}}{\sqrt{\pi'}}\left(\frac{1}{\beta}\boldsymbol{\nabla}\pi'+\pi'\boldsymbol{\nabla}F_n')\right)dw\\ &= \frac{1}{4}\int_{\mathbb{R}^d}\sqrt{\pi\pi'}\boldsymbol{\nabla}\log\frac{\pi'}{\pi}\cdot\left(\frac{1}{\beta}\boldsymbol{\nabla}\log\pi+\boldsymbol{\nabla}F_n\right)dw+\frac{1}{4}\int_{\mathbb{R}^d}\sqrt{\pi\pi'}\boldsymbol{\nabla}\log\frac{\pi}{\pi'}\cdot\left(\frac{1}{\beta}\boldsymbol{\nabla}\log\pi'+\boldsymbol{\nabla}F_n')\right)dw\\ &= -\frac{1}{4}\int_{\mathbb{R}^d}\sqrt{\pi\pi'}\left(\frac{1}{\beta}\|\boldsymbol{\nabla}\log\frac{\pi'}{\pi}\|^2+\boldsymbol{\nabla}\log\frac{\pi}{\pi'}\cdot(\boldsymbol{\nabla}F_n-\boldsymbol{\nabla}F_n')\right)dw\\ &\le \frac{1}{4}\int_{\mathbb{R}^d}\frac{\beta}{4}\sqrt{\pi\pi'}\|\boldsymbol{\nabla}F_n-\boldsymbol{\nabla}F_n'\|^2dw\\ &\le \frac{\beta L^2}{16n^2}\end{aligned} \tag{48}$$

Integrating through time and plugging into the estimate above, we have:

$$\epsilon_n \le 2C\sqrt{D_H(\pi_T||\pi_T')}\le\frac{LC\sqrt{\beta T}}{2n} \tag{49}$$

■

Proof of Proposition 9

Proof We only need to bound the KL divergence to prior distribution γ.

$$\begin{aligned}\frac{d}{dt}D_{KL}(\pi_t||\gamma) &= \int_{\mathbb{R}^d}\frac{\partial\pi}{\partial t}(\log\pi+1-\log\gamma)dw\\ &= -\frac{1}{\beta}\int_{\mathbb{R}^d}\pi\|\boldsymbol{\nabla}\log\pi-\boldsymbol{\nabla}\log\gamma\|^2dw-\int_{\mathbb{R}^d}\pi\langle\boldsymbol{\nabla}\hat{\mathbb{E}}_nf(\boldsymbol{w})+\lambda\boldsymbol{w}+\frac{1}{\beta}\boldsymbol{\nabla}\log\gamma,\boldsymbol{\nabla}\log\pi-\boldsymbol{\nabla}\log\gamma\rangle dw\\ &\le -\left(\frac{1}{\beta}-\frac{1}{2C}\right)\int_{\mathbb{R}^d}\pi\|\boldsymbol{\nabla}\log\pi-\boldsymbol{\nabla}\log\gamma\|^2dw+\frac{C}{2}\int_{\mathbb{R}^d}\pi\|\boldsymbol{\nabla}\hat{\mathbb{E}}_nf(\boldsymbol{w})+\lambda\boldsymbol{w}+\frac{1}{\beta}\boldsymbol{\nabla}\log\gamma\|^2dw\end{aligned} \tag{50}$$

We use Cauchy-Schwartz inequality in the second step, and the constant C will be determined later. The first term is minus Fisher information $I(\pi||\gamma)$, which can be upper bounded by $-D_{KL}(\pi||\gamma)$ itself using logarithmic Sobolev inequality (Gross, 1975; Markowich and Villani, 2000):

$$D_{KL}(\pi||\gamma)\le\sigma_0^2I(\pi||\gamma),\quad \text{for } \gamma=\mathcal{N}(0,\sigma_0^2I) \tag{51}$$

Let $C=\beta$ and plug into the log Sobolev inequality, we get:

$$\frac{d}{dt}D_{KL}(\pi_t||\gamma)\le-\frac{1}{2\beta\sigma_0^2}D_{KL}(\pi_t||\gamma)+\frac{\beta}{2}\int_{\mathbb{R}^d}\pi_t\|\boldsymbol{\nabla}\hat{\mathbb{E}}_nf(\boldsymbol{w})+\lambda\boldsymbol{w}+\frac{1}{\beta}\boldsymbol{\nabla}\log\gamma\|^2dw \tag{52}$$

Solving for D_{KL} with initial value $D_{KL}(\pi_0||\gamma)=0$, we get:

$$D_{KL}(\pi_T||\gamma)\le\frac{\beta}{2}\int_0^T e^{\frac{-(T-t)}{2\beta\sigma_0^2}}\mathbb{E}_{\pi_t}\left\|\boldsymbol{\nabla}\hat{\mathbb{E}}_nf(\boldsymbol{w})+\lambda\boldsymbol{w}+\frac{1}{\beta}\boldsymbol{\nabla}\log\gamma\right\|^2dt \tag{53}$$

Since we use Gaussian prior, the second term in the expectation can be directly calculated as $\frac{1}{\beta}\boldsymbol{\nabla}\log\gamma = -\frac{1}{\beta\sigma_0^2}\boldsymbol{w}$, which exactly cancel out with the $\lambda\boldsymbol{w}$ term. So we have:

$$D_{KL}(\pi_T||\gamma) \leq \frac{\beta}{2}\int_0^T e^{\frac{-\lambda(T-t)}{2}}\mathbb{E}_{\pi_t}\|\boldsymbol{\nabla}\hat{\mathbb{E}}_n f(\boldsymbol{w})\|^2 dt \tag{54}$$

■

Remark: if we do not add the ℓ_2 regularization term, the bound will become

$$D_{KL}(\pi_T||\gamma) \leq \frac{\beta}{2}\int_0^T e^{\frac{-(T-t)}{2\beta\sigma_0^2}}\mathbb{E}_{\pi_t}\left\|\boldsymbol{\nabla}F_n + \frac{1}{\beta}\boldsymbol{\nabla}\log\gamma\right\|^2 dt, \tag{55}$$

which directly depends on norm of the parameter. This is undesirable in high dimensions, since the diffusion term will make the norm at least $\Omega(d)$. Therefore, the use of ℓ_2 regularization is critical to our analysis.

Appendix F. Omitted Proofs in Section 4

F.1. Stability of Langevin Monte Carlo

We consider the following LMC algorithm, which uses full gradients in each update.

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{\eta_k}{n}\sum_{i=1}^n \boldsymbol{\nabla}f(\boldsymbol{w}_k; z_i) + \sqrt{\frac{2\eta_k}{\beta}}\mathcal{N}(0, I_d) \tag{56}$$

Suppose two neighboring datasets S, S' differing only in the i_*-th data. Then one can divide each iteration into two parts: the first part just update $\boldsymbol{w}_k$ and $\boldsymbol{w}'_k$ with gradients over $n-1$ same data and z_{i_*}, i.e.

$$\boldsymbol{w}_k^{(1)} := \boldsymbol{w}_k - \frac{\eta_k}{n}\sum_{i\neq i_*}\boldsymbol{\nabla}f(\boldsymbol{w}_k; z_i) - \frac{\eta_k}{n}\boldsymbol{\nabla}f(\boldsymbol{w}_k; z_{i_*}) \tag{57}$$

,

$$\boldsymbol{w}_k^{(1)\prime} := \boldsymbol{w}'_k - \frac{\eta_k}{n}\sum_{i\neq i_*}\boldsymbol{\nabla}f(\boldsymbol{w}'_k; z_i) - \frac{\eta_k}{n}\boldsymbol{\nabla}f(\boldsymbol{w}'_k; z_{i_*}) \tag{58}$$

and then we obtain $\boldsymbol{w}_{k+1}$ and $\boldsymbol{w}'_{k+1}$ by adding Gaussian noise and replacing the gradient of sample z_{i_*} in $\boldsymbol{w}_k^{(1)\prime}$ by the gradient of sample z'_{i_*}, i.e. $\boldsymbol{w}'_{k+1} = \boldsymbol{w}_k^{(1)\prime} - \frac{\eta_k}{n}\boldsymbol{\nabla}(f(\boldsymbol{w}_k; z'_{i_*}) - f(\boldsymbol{w}_k; z_{i_*})) + \sqrt{\frac{2\eta_k}{\beta}}\mathcal{N}(0, I_d)$. In the first step, squared Hellinger distance does not increase because of the non-expansive property. For the second step, one can view them as consecutive SDEs with drift term $\boldsymbol{g}, \boldsymbol{g}'$ of order $O(\frac{1}{n})$. Hence we can prove the increments of $D_H(\pi||\pi')$ after one iteration is of order $O(\frac{1}{n^2})$, which leads to the following generalization bound.

Theorem 19 (Generalization Error of LMC) *Assuming that*

$$\forall z, z', \forall \boldsymbol{w}, \|\boldsymbol{\nabla}f(\boldsymbol{w}; z) - \boldsymbol{\nabla}f(\boldsymbol{w}; z')\| \leq L$$

Let $\boldsymbol{w}_N$ be result of LMC at N-th round. If loss function $\ell(\cdot,\cdot)$ is uniformly bounded by constant, then the following inequality holds:

$$\mathbb{E}[\mathrm{err}_{gen}(\boldsymbol{w}_T)] \leq O\left(\frac{L\sqrt{\beta\sum_{k=1}^{N}\eta_k}}{n}\right) \tag{59}$$

where the expectation is taken over the randomness of training data.

Proof Here we bound uniform stability of full gradient SGLD by estimating squared Hellinger distance.

We shall assume $\|\boldsymbol{\nabla} f_i\| \leq L$ (which can actually be relaxed to $\|\boldsymbol{\nabla}(f_i - f_j)\| \leq 2L$). (In the proof we will use $f_i(\boldsymbol{w})$ for abbreviation of $f(\boldsymbol{w}, z_i)$, and $f'_i(w)$ for $f(\boldsymbol{w}, z'_i)$. The prime notation on f does not stand for derivative, which is always denoted using $\boldsymbol{\nabla}$ operator.)

Suppose at step k, the starting parameters are W_{k-1} and W'_{k-1} resp. The ending parameters are given by

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \frac{\eta_k}{n}\sum_{i=1}^{n}\boldsymbol{\nabla} f_i(\boldsymbol{w}_k) + \sqrt{\frac{2\eta_k}{\beta}}\boldsymbol{B}_k \tag{60}$$

$$\boldsymbol{w}'_{k+1} = \boldsymbol{w}'_k - \frac{\eta_k}{n}\left(\boldsymbol{\nabla} f'_{i_*}(\boldsymbol{w}'_k) + \sum_{i=1, i\neq i_*}^{n}\boldsymbol{\nabla} f_i(\boldsymbol{w}'_k)\right) + \sqrt{\frac{2\eta_k}{\beta}}\boldsymbol{B}'_k \tag{61}$$

where $\boldsymbol{B}_k, \boldsymbol{B}'_k \sim \mathcal{N}(0, I_d)$.

We consider a family of random variable $\boldsymbol{\theta}_t, \boldsymbol{\theta}'_t (0 \leq t \leq \eta_k)$ defined by

$$\boldsymbol{\theta}_t = \boldsymbol{w}_k - \frac{\eta_k}{n}\sum_{i=1}^{n}\boldsymbol{\nabla} f_i(\boldsymbol{w}_k) + \sqrt{\frac{2t}{\beta}}\boldsymbol{B}_k \tag{62}$$

$$\boldsymbol{\theta}'_t = \boldsymbol{w}'_k - \frac{\eta_k}{n}\sum_{i=1}^{n}\boldsymbol{\nabla} f_i(\boldsymbol{w}'_k) - \frac{t}{n}\Big(\boldsymbol{\nabla} f'_{i_*}(\boldsymbol{w}'_k) - \boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}'_k)\Big) + \sqrt{\frac{2t}{\beta}}\boldsymbol{B}'_k \tag{63}$$

Till now, we only consider the one-time distribution of $\boldsymbol{\theta}_t, \boldsymbol{\theta}'_t$, and their dependence on $\boldsymbol{w}_k, \boldsymbol{w}'_k$, without taking the inter-dependence of whole process into consideration, so we use a simple way of expanding the Gaussian noise. In the actual construction of the SDE, it will be expanded via Brownian motion.

Let the pdf of $\boldsymbol{\theta}_t, \boldsymbol{\theta}'_t$ be π_t, π'_t. We can see that

- $\boldsymbol{\theta}_0 = \boldsymbol{w}_k - \frac{\eta_k}{n}\sum\limits_{i=1}^{n}\boldsymbol{\nabla} f_i(\boldsymbol{w}_k), \boldsymbol{\theta}'_0 = \boldsymbol{w}'_k - \frac{\eta_k}{n}\sum\limits_{i=1}^{n}\boldsymbol{\nabla} f_i(\boldsymbol{w}'_k)$, so that

$$D_H(\pi_0||\pi'_0) \leq D_H(p_k||p'_k) \tag{64}$$

- the explicit formulae for π_t and π'_t are given by

$$\pi_t(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{w}_k}\left(\frac{\beta}{4\pi t}\right)^{d/2}\exp\Big(-\frac{\beta}{4t}\|\boldsymbol{w} - \boldsymbol{w}_k + \frac{\eta}{n}\sum_{i=1}^{n}\boldsymbol{\nabla} f_i(\boldsymbol{w}_k)\|^2\Big) \tag{65}$$

and

$$\pi_t'(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{w}_k'} \left(\frac{\beta}{4\pi t}\right)^{d/2} \exp\left(-\frac{\beta}{4t}\|\boldsymbol{w}-\boldsymbol{w}_k'+\frac{\eta}{n}\sum_{i=1}^{n}\boldsymbol{\nabla} f_i(\boldsymbol{w}_k')+\frac{t}{n}\left(\boldsymbol{\nabla} f_{i_*}'(\boldsymbol{w}_k') - \boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k')\right)\|^2\right) \tag{66}$$

Although formidable at first glance, π_t and π_t' are nothing but superposition of Gaussian density functions w.r.t $\boldsymbol{w}$.

Define $\boldsymbol{g}_t(\boldsymbol{w})$ to be $\mathbf{0}$ and define $\boldsymbol{g}_t'(\boldsymbol{w})$ by

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{w}_k'}[\frac{1}{n}\left(\boldsymbol{\nabla} f_{i_*}'(\boldsymbol{w}_k') - \boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k')\right)|\boldsymbol{\theta}_t' = \boldsymbol{w}]\\ =&\frac{1}{n\pi_t'(\boldsymbol{w})}\mathbb{E}_{\boldsymbol{w}_k'}\left(\left(\boldsymbol{\nabla} f_{i_*}'(\boldsymbol{w}_k') - \boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k')\right)\left(\frac{\beta}{4\pi t}\right)^{d/2} e^{-\frac{\beta}{4t}\|\boldsymbol{w}-\boldsymbol{w}_k'+\frac{\eta}{n}\sum_{i=1}^{n}\boldsymbol{\nabla} f_i(\boldsymbol{w}_k')+\frac{t}{n}\left(\boldsymbol{\nabla} f_{i_*}'(\boldsymbol{w}_k')-\boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k')\right)\|^2}\right)\end{aligned} \tag{67}$$

Then by taking derivatives w.r.t to $\boldsymbol{w}$ and t, we can obtain the following equations, which has the same one-time marginal distribution as $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_t'$ (though they are not the same process):

$$\frac{\partial \pi_t}{\partial t} = \frac{1}{\beta}\Delta\pi_t + \boldsymbol{\nabla}\cdot(\pi_t \boldsymbol{g}_t) \tag{68}$$

$$\frac{\partial \pi_t'}{\partial t} = \frac{1}{\beta}\Delta\pi_t' + \boldsymbol{\nabla}\cdot\left(\pi_t' \boldsymbol{g}_t'\right) \tag{69}$$

From definition and the assumption $\forall z, z', \|\nabla f(\boldsymbol{w}; z) - \nabla f(\boldsymbol{w}; z')\| \leq L$, we have

$$\forall \boldsymbol{w}, \|\boldsymbol{g}_t(\boldsymbol{w}) - \boldsymbol{g}_t'(\boldsymbol{w})\| \leq \frac{L}{n} \tag{70}$$

$$\begin{aligned}\frac{d}{dt}D_H(\pi_t||\pi_t') &= -\frac{1}{2}\int_{\mathbb{R}^d}\left(\frac{1}{\beta}\sqrt{\pi\pi'}\|\boldsymbol{\nabla}\log\frac{\pi}{\pi'}\|^2 + \sqrt{\pi\pi'}\boldsymbol{\nabla}\log\frac{\pi}{\pi'}\cdot(\boldsymbol{g}_t - \boldsymbol{g}_t')\right)dw\\ &\leq \frac{\beta}{8}\int_{\mathbb{R}^d}\sqrt{\pi\pi'}\|\boldsymbol{g}_t - \boldsymbol{g}_t'\|^2 dw\\ &= \frac{\beta L^2}{8n^2}\end{aligned}$$

As a result, we can estimate the change of squared Hellinger distance in this step:

$$\begin{aligned}D_H(\pi_{k+1}||\pi_{k+1}') &= D_H(\pi_{\eta_k}||\pi_{\eta_k}')\\ &= D_H(\pi_0||\pi_0') + \int_0^{\eta_k}\frac{d}{dt}D_H(\pi_t||\pi_t')dt\\ &\leq D_H(p_0||p_0') + \int_0^{\eta_k}\frac{\beta L^2}{8n^2}dt\\ &= D_H(p_0||p_0') + \frac{\beta L^2}{8n^2}\eta_k\end{aligned}$$

Then by induction we shall have a final bound for $D_{KL}(\pi||\pi')$ of the form $\frac{\beta L^2}{8n^2}\sum\limits_{k=1}^{N}\eta_k$.

Then the bound for uniform stability is given by

$$\epsilon_n \le O\left(\frac{L\sqrt{\beta\sum_{k=1}^{N}\eta_k}}{n}\right) \tag{71}$$

■

Combining the techniques from the proof of Theorem 19 and Lemma 10, we can easily get the general result for the case of mini-batch.

F.2. Stability of SGLD - A Succinct Analysis

As random draw of a training example is more popular in practice, it is desirable to analyze generalization properties of SGLD. In the rest part of this section, we will assume $\boldsymbol{g}_k = \boldsymbol{\nabla} f_{i_k}(\boldsymbol{w})$, where i_k is the index of randomly drawn training example. We will first present a simple analysis for stability of SGLD. Though the resulting bound is not optimal, the analysis illustrates important principles for understanding how SGLD helps stability. In the following, we will derive upper bounds for $\delta_k \triangleq D_H(p_k||p_k')$ recursively. There are two possible cases for i_k:

- If $i_k \neq i_*$, then SGLD implemented over S or S' will use the same gradient mapping, i.e. $\psi_k : \boldsymbol{w} \mapsto \boldsymbol{w} - \eta_k \nabla f(\boldsymbol{w}; z_{i_k})$, then we have

  $$D_H(\mathcal{P}(\psi_k(\boldsymbol{w}_k)|i_k)||\mathcal{P}(\psi_k(\boldsymbol{w}_k')|i_k)) \le D_H(p_k||p_k') = \delta_k \tag{72}$$

  Furthermore let $\mathcal{G}_k = \mathcal{N}(0, \frac{\eta_k}{\beta} I_d)$, by the convexity of squared Hellinger distance (which is implied by joint convexity of f-divergence), there is

  $$\begin{aligned} D_H(\mathcal{P}(\boldsymbol{w}_{k+1}|i_k)||\mathcal{P}(\boldsymbol{w}_{k+1}'|i_k)) &= D_H(\mathcal{G}_k * \mathcal{P}(\psi_k(\boldsymbol{w}_k)|i_k)||\mathcal{G}_k * \mathcal{P}(\psi_k(\boldsymbol{w}_k')|i_k)) \\ &\le D_H(\mathcal{P}(\psi_k(\boldsymbol{w}_k)|i_k)||\mathcal{P}(\psi_k(\boldsymbol{w}_k')|i_k)) \\ &\le \delta_k \end{aligned}$$

  So in this case, the SGLD update is non-expansive with respect to δ_k.

- If $i_k = i_*$, we have nothing but limited step size in hand. The increase of f-divergence can be bounded through norm-based shifts in parameter space only under smoothness conditions, which is helped by Gaussian noise. Therefore, we expand the discrete-time update into a stochastic process, where the effect of gradient flow is smoothed by Gaussian at each time t.

Concretely, for $i_k = i_*$, the update can be interpolated as:

$$\forall t \in [0, \eta_k], \quad \boldsymbol{\theta}_t = \boldsymbol{\theta}_0 - \int_0^t \boldsymbol{\nabla} f_{i_k}(\boldsymbol{\theta}_0) ds + \sqrt{\frac{2}{\beta}}\int_0^t d\boldsymbol{B}_s, \quad \boldsymbol{\theta}_0 = \boldsymbol{w}_k \tag{73}$$

However, $\boldsymbol{\theta}_t$ is not a Markov process, as it always involves the initial random point $\boldsymbol{\theta}_0$. Using the same technique as in Raginsky et al. (2017), we define $\boldsymbol{g}_t(\boldsymbol{v}) \triangleq \mathbb{E}\left(\boldsymbol{\nabla} f_{i_k}(\boldsymbol{\theta}_0)\middle|\boldsymbol{\theta}_t = \boldsymbol{v}\right)$. Mimicking distribution results (Gyöngy, 1986) guarantees solution to the following SDE has the same one-time marginal as $\boldsymbol{\theta}_t$.

$$d\boldsymbol{v}_t = \boldsymbol{g}_s(\boldsymbol{v}_s)ds + \sqrt{\frac{2}{\beta}}d\boldsymbol{B}_s, \quad v_0 \sim p_k \tag{74}$$

The corresponding Fokker-Planck equation for above process is:

$$\frac{\partial \pi}{\partial t} = \boldsymbol{\nabla} \cdot \left(\frac{1}{\beta} \boldsymbol{\nabla} \pi + \pi \boldsymbol{g}_t \right) \tag{75}$$

We also have counterparts for the neighboring dataset, denoted as π_t'. With the help of these PDEs, we can bound the variation of squared Hellinger distance.

As in the ideal case, we can compute that

$$\begin{aligned} \frac{d}{dt} D_H(\pi_t || \pi_t') &= -\frac{1}{4} \int_{\mathbb{R}^d} \sqrt{\pi\pi'} \left(\frac{1}{\beta} \| \boldsymbol{\nabla} \log \frac{\pi'}{\pi} \|^2 + \boldsymbol{\nabla} \log \frac{\pi}{\pi'} \cdot (\boldsymbol{g}_t - \boldsymbol{g}_t') \right) dw \\ &\leq \frac{\beta}{16} \int_{\mathbb{R}^d} \sqrt{\pi\pi'} \| \boldsymbol{g}_t - \boldsymbol{g}_t' \|^2 dw \\ &\leq \frac{\beta L^2}{16} \end{aligned} \tag{76}$$

For $i_k = i_*$, we have

$$D_H(\mathcal{P}(\boldsymbol{w}_{k+1}|i_*) || \mathcal{P}(\boldsymbol{w}_{k+1}'|i_*)) \leqslant \delta_k + \frac{\beta L^2}{16} \eta_k \tag{77}$$

Combining above two cases and using the convexity of squared Hellinger distance, we obtain

$$\delta_{k+1} \leq \frac{n-1}{n} \delta_k + \frac{1}{n} (\delta_k + \frac{\beta L^2}{8} \eta_k) = \delta_k + \frac{\beta L^2}{8n} \eta_k. \tag{78}$$

Putting them together, we get following guarantees for SGLD:

Theorem 20 *Consider N rounds of SGLD with parameters β and $\{\eta_i\}$. If we assume*

1. *the loss function $\ell(\boldsymbol{w}; z)$ is uniformly bounded by C;*
2. *$\forall z, z'$, the gradients of objective function satisfy $\|\boldsymbol{\nabla} f(\boldsymbol{w}; z) - \boldsymbol{\nabla} f(\boldsymbol{w}; z')\| \leq L$*

Then we have the following generalization bound in expectation

$$\mathbb{E}[\text{err}(\boldsymbol{w}_N)] \leq \frac{LC}{2} \left(\frac{\beta}{n} \sum_{i=1}^{k} \eta_i \right)^{1/2} \tag{79}$$

F.3. Proof of Lemma 10

Proof Consider the following SGLD update step:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta_k \nabla f(\boldsymbol{w}_k; z_{i_k}) + \sqrt{\frac{2\eta_k}{\beta}} \boldsymbol{B}_k, \quad \boldsymbol{w}_0 \sim \mathcal{N}(0, \sigma_0^2 I_d), \quad \boldsymbol{B}_k \sim \mathcal{N}(0, I_d), \quad i_k \sim \mathcal{U}\{1, 2, \cdots, n\} \tag{80}$$

where $\boldsymbol{w}_0, \boldsymbol{B}_k, i_k$ are independent. Apparently it is equivalent to the following one:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - (1 - X)\eta_k \boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{j_k}) - X\eta_k \boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{i_*}) + \sqrt{\frac{2\eta_k}{\beta}} \boldsymbol{B}_k,$$
$$\boldsymbol{w}_0 \sim \mathcal{N}(0, \sigma_0^2 I_d), \quad \boldsymbol{B}_k \sim \mathcal{N}(0, I_d), \quad j_k \sim \mathcal{U}(\{1, 2, \cdots, n\} \setminus \{i_*\})$$

where $\boldsymbol{w}_0, \boldsymbol{B}_k, i_k, X$ are independent and $\mathcal{P}(X = 1) = \frac{1}{n}, \mathcal{P}(X = 0) = \frac{n-1}{n}$.

As in the case of LMC, we are going to construct a pair of random variable sequences indexed by t, and then construct an SDE with the same one-time marginals.

We consider a family of random variables $\boldsymbol{\theta}_t$ $(0 \leq t \leq \eta_k)$ defined by

$$\boldsymbol{\theta}_t = \boldsymbol{w}_k - \eta_k \boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{j_k}) - Xt(\boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{i_*}) - \boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{j_k})) + \sqrt{\frac{2t}{\beta}} \boldsymbol{B}_k \tag{81}$$

Denote pdf of $\boldsymbol{\theta}_t$ by π_t. For neighboring datasets, we also have $\boldsymbol{\theta}'_t$ and π'_t. We can see that

- $\boldsymbol{\theta}_0 = \boldsymbol{w}_k - \eta_k \boldsymbol{\nabla} f(\boldsymbol{w}_k; z_{j_k}), \boldsymbol{\theta}'_0 = \boldsymbol{w}'_k - \eta_k \boldsymbol{\nabla} f(\boldsymbol{w}'_k; z_{j_k})$, so by non-expansiveness,

$$D_H(\pi_0 || \pi'_0) \leq D_H(p_k || p'_k) \tag{82}$$

- $\boldsymbol{\theta}_{\eta_k} = \boldsymbol{w}_{k+1}$ and $\boldsymbol{\theta}'_{\eta_k} = \boldsymbol{w}'_{k+1}$
- For $0 \leq t \leq \eta_k$, π_t and π'_t are given by

$$\pi_t(\boldsymbol{w}) = \mathbb{E}_{X, j_k, \boldsymbol{w}_k} \left(\frac{\beta}{4\pi t}\right)^{d/2} \exp(-\beta \|\boldsymbol{w} - \boldsymbol{w}_k + \eta_k \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k) + Xt(\boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k) - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k))\|^2 / (4t)) \tag{83}$$

and

$$\pi'_t(\boldsymbol{w}) = \mathbb{E}_{X, j_k, \boldsymbol{w}'_k} \left(\frac{\beta}{4\pi t}\right)^{d/2} \exp(-\beta \|\boldsymbol{w} - \boldsymbol{w}'_k + \eta_k \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}'_k) + Xt(\boldsymbol{\nabla} f'_{i_*}(\boldsymbol{w}'_k) - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}'_k))\|^2 / (4t)) \tag{84}$$

(As in the LMC case, in this proof we use $f_i(\boldsymbol{w})$ for abbreviation of $f(\boldsymbol{w}, z_i)$, and $f'_i(w)$ for $f(\boldsymbol{w}, z'_i)$.)

Although formidable at first glance, π_t and π'_t are nothing but superposition of Gaussian density functions w.r.t $\boldsymbol{w}$. Here $f_i(\boldsymbol{w}_k) = f(\boldsymbol{y}; z_i), f'_i(\boldsymbol{y}) = f(\boldsymbol{y}; z'_i)$.

Define $\hat{\boldsymbol{g}}$ by

$$\begin{aligned} &\mathbb{E}_{X, j_k, \boldsymbol{w}_k}[X(\boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k) - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k)) | \boldsymbol{\theta}_t = \boldsymbol{w}] \\ =& \frac{1}{\pi_t(\boldsymbol{w})} \mathbb{E}_{X, j_k, \boldsymbol{w}_k} X(\boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k) - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k)) \cdot \left(\frac{\beta}{4\pi t}\right)^{d/2} e^{-\beta \|\boldsymbol{w} - \boldsymbol{w}_k + \eta_k \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k) + Xt(\boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k) - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k))\|^2 / (4t)} \end{aligned} \tag{85}$$

and $\hat{\boldsymbol{g}}'$ by

$$\begin{aligned}&\mathbb{E}_{X,j_k,\boldsymbol{w}_k'}[X(\boldsymbol{\nabla} f_{i_*}'(\boldsymbol{w}_k') - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k'))|\boldsymbol{\theta}_t' = \boldsymbol{w}]\\ =&\frac{1}{\pi_t'(\boldsymbol{w})}\mathbb{E}_{X,j_k,\boldsymbol{w}_k'}X(\boldsymbol{\nabla} f_{i_*}'(\boldsymbol{w}_k') - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k'))\cdot\left(\frac{\beta}{4\pi t}\right)^{d/2} e^{-\beta\|\boldsymbol{w}-\boldsymbol{w}_k'+\eta_k\boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k')+Xt(\boldsymbol{\nabla} f_{i_*}'(\boldsymbol{w}_k')-\boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k'))\|^2/(4t)}\end{aligned} \tag{86}$$

Then it can be easily verified by calculating derivatives w.r.t $\boldsymbol{w}$ and t that:

$$\frac{\partial \pi}{\partial t} = \frac{1}{\beta}\triangle\pi + \boldsymbol{\nabla}\cdot(\pi\hat{\boldsymbol{g}}) \tag{87}$$

and

$$\frac{\partial \pi'}{\partial t} = \frac{1}{\beta}\triangle\pi' + \boldsymbol{\nabla}\cdot(\pi'\hat{\boldsymbol{g}}') \tag{88}$$

With the Lemma 21 below and using similar analysis as before, then we compute the time derivative of squared Hellinger distance to be

$$\begin{aligned}\frac{d}{dt}D_H(\pi_t||\pi_t') &= -\frac{1}{4}\int_{\mathbb{R}^d}\sqrt{\pi\pi'}\left(\frac{1}{\beta}\|\boldsymbol{\nabla}\log\frac{\pi'}{\pi}\|^2 + \boldsymbol{\nabla}\log\frac{\pi}{\pi'}\cdot(\hat{\boldsymbol{g}}_t - \hat{\boldsymbol{g}}_t')\right)dw\\ &\leq \frac{\beta}{16}\int\sqrt{\pi\pi'}\|\hat{\boldsymbol{g}} - \hat{\boldsymbol{g}}'\|^2 dw\\ &< \frac{\beta L^2}{n^2}\end{aligned}$$

So we have

$$D_H(p_{k+1}||p_{k+1}') = D_H(\pi_{\eta_k}||\pi_{\eta_k}') \leq D_H(\pi_0||\pi_0') + \frac{\beta L^2}{n^2}\eta_k \leq D_H(p_k||p_k') + \frac{\beta L^2}{n^2}\eta_k \tag{89}$$

Then one arrives at the statement by induction. ■

Lemma 21 *Under the same assumptions with Lemma 10, there is*

$$\int\sqrt{\pi\pi'}\|\boldsymbol{g}_t - \boldsymbol{g}_t'\|^2 dw \leq \frac{4\sqrt{2}L^2}{(n-1)^2} \tag{90}$$

Proof Let u_t, u_t' denote the pdfs of θ_t, θ_t' conditioned on $X = 1$ respectively, and let v_t, v_t' denote the pdfs of θ_t, θ_t' conditioned on $X = 0$ respectively.

Then it's easily seen from equation 85 and equation 86 that

$$\hat{\boldsymbol{g}}_t(\boldsymbol{w}) = \frac{u_t(\boldsymbol{w})}{n\pi_t(\boldsymbol{w})}\mathbb{E}(\boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k) - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k)|\boldsymbol{\theta}_t = \boldsymbol{w}) \tag{91}$$

and

$$\hat{\boldsymbol{g}}_t'(\boldsymbol{w}) = \frac{u_t'(\boldsymbol{w})}{n\pi_t'(\boldsymbol{w})}\mathbb{E}(\boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k') - \boldsymbol{\nabla} f_{j_k}'(\boldsymbol{w}_k')|\boldsymbol{\theta}_t' = \boldsymbol{w}) \tag{92}$$

So we have bounds:

$$\|\hat{\boldsymbol{g}}_t(\boldsymbol{w})\| \leq \frac{u_t(\boldsymbol{w})L}{n\pi_t(\boldsymbol{w})} \tag{93}$$

and

$$\|\hat{\boldsymbol{g}}'_t(\boldsymbol{w})\| \leq \frac{u'_t(\boldsymbol{w})L}{n\pi'_t(\boldsymbol{w})} \tag{94}$$

Then we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \sqrt{\pi_t\pi'_t}\|\hat{\boldsymbol{g}}_t - \hat{\boldsymbol{g}}'_t\|^2 dw &\leq 2\int_{\mathbb{R}^d} \sqrt{\pi_t\pi'_t}\|\hat{\boldsymbol{g}}\|^2 dw + 2\int_{\mathbb{R}^d} \sqrt{\pi_t\pi'_t}\|\hat{\boldsymbol{g}}'\|^2 dw \\
&\leq 2\sqrt{\int \pi_t\|\hat{\boldsymbol{g}}\|^4 dw \int \pi'_t dw} + 2\sqrt{\int \pi'_t\|\hat{\boldsymbol{g}}'\|^4 dw \int \pi_t dw} \\
&= 2\sqrt{\int \pi_t\|\hat{\boldsymbol{g}}\|^4 dw} + 2\sqrt{\int \pi'_t\|\hat{\boldsymbol{g}}'\|^4 dw} \\
&\leq 2\sqrt{\int \pi_t\left(\frac{u_t L}{n\pi_t}\right)^4 dw} + 2\sqrt{\int \pi'_t\left(\frac{u'_t L}{n\pi'_t}\right)^4 dw} \\
&\leq 2L^2\sqrt{\int \frac{u_t^4}{n((n-1)v_t + u_t)^3} dw} + 2L^2\sqrt{\int \frac{u_t'^4}{n((n-1)v'_t + u'_t)^3} dw} \\
&\leq \frac{2L^2}{(n-1)^2}\sqrt{\int \frac{u_t^4}{v_t^3} dw} + \frac{2L^2}{(n-1)^2}\sqrt{\int \frac{u_t'^4}{v_t'^3} dw}
\end{aligned}$$

To proceed, we shall first seek to find the PDEs satisfied by u_t, v_t, u'_t, v'_t.

By definition, the explicit expressions for u_t, v_t are

$$u_t(\boldsymbol{w}) = \mathbb{E}_{j_k,\boldsymbol{w}_k}\left(\frac{\beta}{4\pi t}\right)^{d/2} \exp(-\beta\|\boldsymbol{w} - \boldsymbol{w}_k + \eta_k \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k) + t(\boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k) - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k))\|^2/(4t)) \tag{95}$$

and

$$v_t(\boldsymbol{w}) = \mathbb{E}_{j_k,\boldsymbol{w}_k}\left(\frac{\beta}{4\pi t}\right)^{d/2} \exp(-\beta\|\boldsymbol{w} - \boldsymbol{w}_k + \eta_k \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k)\|^2/(4t)) \tag{96}$$

Define $\boldsymbol{g}_t(\boldsymbol{w})$ by

$$\begin{aligned}
&\mathbb{E}_{j_k,\boldsymbol{w}_k}\left[\boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k) - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k)\middle| X = 1, \boldsymbol{\theta}_t = \boldsymbol{w}\right] \\
=&\frac{1}{u_t(\boldsymbol{w})}\mathbb{E}_{j_k,\boldsymbol{w}_k}\Big(\boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k) - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k)\Big) \cdot \left(\frac{\beta}{4\pi t}\right)^{d/2} e^{-\beta\|\boldsymbol{w} - \boldsymbol{w}_k + \eta_k \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k) + t(\boldsymbol{\nabla} f_{i_*}(\boldsymbol{w}_k) - \boldsymbol{\nabla} f_{j_k}(\boldsymbol{w}_k))\|^2/(4t)}
\end{aligned} \tag{97}$$

Then the following equality holds:

$$\frac{\partial u_t}{\partial t} = \frac{1}{\beta}\Delta u_t + \boldsymbol{\nabla} \cdot (u\boldsymbol{g}_t) \tag{98}$$

And for v_t, the following equality holds:

$$\frac{\partial v_t}{\partial t} = \frac{1}{\beta}\Delta v_t \tag{99}$$

Using the Lemma 22 below, it follows that for $t \le \eta_k \le \frac{\ln 2}{\beta L^2}$

$$\int \frac{u_t^4}{v_t^3} dw \le 8 \tag{100}$$

Similarly we have

$$\int \frac{u_t'^4}{v_t'^3} dw \le 8 \tag{101}$$

As a result,

$$\int \sqrt{\pi\pi'} \|\boldsymbol{g}_t - \boldsymbol{g}_t'\|^2 dw \le \frac{4\sqrt{2}L^2}{(n-1)^2} \tag{102}$$

■

Lemma 22 *Let $u, v \in C^\infty([0, +\infty) \times \mathbb{R}^d)$ satisfying respectively:*

- $\frac{\partial u}{\partial t} = \frac{1}{\beta}\triangle u + \boldsymbol{\nabla} \cdot (u\boldsymbol{g}_t)$
- $\frac{\partial v}{\partial t} = \frac{1}{\beta}\triangle v + \boldsymbol{\nabla} \cdot (v\boldsymbol{g}_t')$

and $u_0 = v_0$.

Assume that $\|\boldsymbol{g}_t - \boldsymbol{g}_t'\| \le L$

Then for $t \le \frac{\ln 2}{\beta L^2}$, we have

$$\int \frac{u_t^4}{v_t^3} dw \le 8 \tag{103}$$

Proof

$$\begin{aligned}
\frac{d}{dt}\int_{\mathbb{R}^d} \frac{u_t^4}{v_t^3} dw &= \int 4\frac{\partial u}{\partial t}\frac{u^3}{v^3} dw - 3\frac{\partial v}{\partial t}\frac{u^4}{v^4} dw \\
&= \int \left(-4(\frac{1}{\beta}\boldsymbol{\nabla} u + u\boldsymbol{g}) \cdot \boldsymbol{\nabla}\frac{u^3}{v^3} + 3(\frac{1}{\beta}\boldsymbol{\nabla} v + v\boldsymbol{g}') \cdot \boldsymbol{\nabla}\frac{u^4}{v^4} \right) dw \\
&= \int \frac{u^4}{v^3} \left\{ -4(\frac{1}{\beta}\boldsymbol{\nabla}\log u + \boldsymbol{g}) \cdot \boldsymbol{\nabla}\log\frac{u^3}{v^3} + 3(\frac{1}{\beta}\boldsymbol{\nabla}\log v + \boldsymbol{g}') \cdot \boldsymbol{\nabla}\log\frac{u^4}{v^4} \right\} dw \\
&= \int \frac{12u^4}{v^3} \left\{ -(\frac{1}{\beta}\boldsymbol{\nabla}\log u + \boldsymbol{g}) \cdot \boldsymbol{\nabla}\log\frac{u}{v} + (\frac{1}{\beta}\boldsymbol{\nabla}\log v + \boldsymbol{g}') \cdot \boldsymbol{\nabla}\log\frac{u}{v} \right\} dw \\
&= \int \frac{12u^4}{v^3} \left\{ -\frac{1}{\beta}\|\boldsymbol{\nabla}\log\frac{v}{u}\|^2 - (\boldsymbol{g} - \boldsymbol{g}') \cdot \boldsymbol{\nabla}\log\frac{u}{v} \right\} dw \\
&\le \int \frac{3\beta u^4}{v^3}\|\boldsymbol{g} - \boldsymbol{g}'\|^2 dw \\
&\le 3\beta L^2 \int_{\mathbb{R}^d} \frac{u_t^4}{v_t^3} dw
\end{aligned}$$

Then

$$\frac{d}{dt}\ln\int\frac{u_t^4}{v_t^3}dw \leq 3\beta L^2 \tag{104}$$

For $t \leq \frac{\ln 2}{\beta L^2}$, we have

$$\ln\int\frac{u_t^4}{v_t^3}dw \leq \frac{\ln 2}{\beta L^2}\cdot 3\beta L^2 = 3\ln 2 \tag{105}$$

i.e.

$$\int\frac{u_t^4}{v_t^3}dw \leq 8 \tag{106}$$

■

F.4. Dealing with Large Step Sizes

Proof of Theorem 11

Proof Consider the concatenated procedure $\mathcal{A}''$ that use samples S for the first k steps and samples S' for the rest steps. We denote the corresponding parameters and densities by $\boldsymbol{w}_k''$ and p_k''.

$$\begin{aligned}\epsilon_n &= \sup_z\left|\int\ell(\boldsymbol{w};z)(p_N'(\boldsymbol{w})-p_N(\boldsymbol{w}))dw\right|\\ &\leq \sup_z\left|\int\ell(\boldsymbol{w};z)(p_N'(\boldsymbol{w})-p_N''(\boldsymbol{w}))dw\right|+\sup_z\left|\int\ell(\boldsymbol{w};z)(p_N(\boldsymbol{w})-p_N''(\boldsymbol{w}))dw\right|\\ &\leq A_\ell D_A(p_N'||p_N'')+B_\ell D_B(p_N||p_N'')\end{aligned}$$

For step $k+1,\cdots,N$ the concatenated procedure $\mathcal{A}''$ uses sample set S'. So the transformation from p_k' to p_N' is the same as the transformation from p_k'' to p_N''. By non-expansiveness, we have $D_A(p_N'||p_N'')\leq D_A(p_k'||p_k'')\leq h_A(\eta_1,\cdots,\eta_k)$.

Note that $p_l = p_l''$ for $l=1,\cdots,k$, so we have $D_B(p_N||p_N'')\leq h_B(\eta_{k+1},\cdots,\eta_N)$.

Therefore, we obtain

$$\epsilon_n \leq A_\ell h_A(\eta_1,\cdots,\eta_k)+B_\ell h_B(\eta_{k+1},\cdots,\eta_N) \tag{107}$$

■

Proof of Lemma 12

Proof For $k=0$, both p_k and p_k' are equal to the prior distribution so that

$$\int|p_0-p_0'|dw = 0 \tag{108}$$

Assume the distributions before the k th step is p_k and p'_k, and denote the distribution density functions for $\boldsymbol{w}_k, \boldsymbol{w}'_k$ after k steps conditioned on $i_k = i$ by $p_k^{(i)}, p_k^{(i)\prime}$ respectively, then

$$\begin{aligned}\int |p_{k+1} - p'_{k+1}| dw &= \int \left| \frac{1}{n} \sum_{i=1}^{n} p_k^{(i)} - \frac{1}{n} \sum_{i=1}^{n} p_k^{(i)\prime} \right| dw \\ &\leq \frac{1}{n} \sum_{i=1}^{n} \int \left| p_k^{(i)} - p_k^{(i)\prime} \right| dw\end{aligned}$$

For $i \neq i_*$, $\int |p_k^{(i)} - p_k^{(i)\prime}| dw \leq \int |p_k - p'_k| dw$ since they undergo the same gradient step and Gaussian convolution.

For $i = i_*$, $\int |p_k^{(i)} - p_k^{(i)\prime}| dw \leq 2$.

As a result, we have

$$\int |p_{k+1} - p'_{k+1}| dw \leq \int |p_k - p'_k| dw + \frac{2}{n} \tag{109}$$

By induction, after k_0 steps,

$$\int |p_{k_0} - p'_{k_0}| dw \leq \frac{2k_0}{n} \tag{110}$$

■

Appendix G. Omitted Proofs in Section 5

Proof of Theorem 14

Proof Given $\boldsymbol{\theta}_0 = \boldsymbol{y}$ fixed, the conditional density of $\boldsymbol{\theta}_t$ given by the assumption is a Gaussian pdf, which satisfies Ornstein-Uhlenbeck equation with $\boldsymbol{b} = -\frac{1}{\lambda}\boldsymbol{g}(\boldsymbol{y})$ and parameter λ, according to Proposition 18.

So the conditional density $\pi(\cdot|\boldsymbol{\theta}_0 = \boldsymbol{y})$ satisfies Fokker-Planck Equation:

$$\frac{\partial \pi(\boldsymbol{w}|\boldsymbol{\theta}_0 = \boldsymbol{y})}{\partial t} = \frac{1}{\beta} \Delta \pi(\boldsymbol{w}|\boldsymbol{\theta}_0 = \boldsymbol{y}) + \boldsymbol{\nabla} \cdot (\lambda \pi(\boldsymbol{w}|\boldsymbol{\theta}_0 = \boldsymbol{y}) \boldsymbol{w}) + \boldsymbol{\nabla} \cdot (\pi(\boldsymbol{w}|\boldsymbol{\theta}_0 = \boldsymbol{y}) \boldsymbol{g}(\boldsymbol{y})) \tag{111}$$

Let $\boldsymbol{\theta}_0 = \boldsymbol{y} \sim \pi_0$, and take expectations for both sides. By construction π_t is smooth enough to justify exchange of order of integration and differentiation. So, for any $\boldsymbol{w} \in \mathbb{R}^d$, we have:

$$\begin{cases} \mathbb{E}\left(\dfrac{\partial \pi(\boldsymbol{w}|\boldsymbol{\theta}_0 = \boldsymbol{y})}{\partial t}\right) = \dfrac{\partial \pi}{\partial t} \\ \mathbb{E}\left(\Delta \pi(\boldsymbol{w}|\boldsymbol{\theta}_0 = \boldsymbol{y})\right) = \Delta \pi \\ \mathbb{E}\left(\boldsymbol{\nabla} \cdot (\lambda \pi(\boldsymbol{w}|\boldsymbol{\theta}_0 = \boldsymbol{y}) \boldsymbol{w})\right) = \boldsymbol{\nabla} \cdot (\lambda \pi \boldsymbol{w}) \\ \mathbb{E}\left(\boldsymbol{\nabla} \cdot (\pi(\boldsymbol{w}|\boldsymbol{\theta}_0 = \boldsymbol{y}) \boldsymbol{g}(\boldsymbol{y}))\right) = \boldsymbol{\nabla} \cdot \displaystyle\int_{\mathbb{R}^d} \boldsymbol{g}(\boldsymbol{y}) \pi_{0,t}(\boldsymbol{y}, \boldsymbol{w}) dy = \boldsymbol{\nabla} \cdot (\pi(\boldsymbol{w}) \mathbb{E}(\boldsymbol{g}(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_t = \boldsymbol{w})) \end{cases}, \tag{112}$$

where $\pi_{0,t}(\cdot,\cdot)$ is the joint density of θ_0 and θ_t. Putting them together we get the PDE as desired. ■

Proof of Lemma 15

Proof Consider the partial differential equations constructed in Section 5.1. We take p_k as the initial distribution. The randomness of this SGLD update comes from two sources: random choice of stochastic gradient operator $\boldsymbol{g}_k(\cdot)$ and the Gaussian noise. The first one is by uniform draw of data points, and is independent with the rest part of the algorithm. (Though $\boldsymbol{g}_k(\boldsymbol{w}_k)$ depends on previous trajectory through $\boldsymbol{w}_k$, $\boldsymbol{g}_k(\cdot)$ as a random function is independent.) So we first condition on the choice of $\boldsymbol{g}_k(\cdot)$, and let the conditional distribution of $\boldsymbol{w}_{k+1}$ be $p_{k+1}|_{\boldsymbol{g}_k}$. By convexity of KL divergence, any upper bound for the conditional distribution is a valid upper bound for p_k.

For the PDE, we take derivative of KL divergence between time-varying posterior and time-varying prior. In the following, we denote $\mathbb{E}\left[\boldsymbol{\nabla} f_{i_k}(\boldsymbol{\theta}_0)|\boldsymbol{\theta}_t=\boldsymbol{w}\right]$ by $\boldsymbol{h}_t(\boldsymbol{w})$ for convenience.

$$\begin{aligned}
\frac{d}{dt}D_{KL}(\pi_t||\tilde{\gamma}_t) &= \int_{\mathbb{R}^d}\frac{\partial\pi}{\partial t}(\log\pi+1-\log\tilde{\gamma})dw-\int_{\mathbb{R}^d}\frac{\pi}{\tilde{\gamma}}\frac{\partial\tilde{\gamma}}{\partial t}dw\\
&=\int_{\mathbb{R}^d}\pi\langle\boldsymbol{h}_t(\boldsymbol{w})+\lambda\boldsymbol{w}+\frac{1}{\beta'_k}\boldsymbol{\nabla}\log\pi,\boldsymbol{\nabla}\log\pi-\boldsymbol{\nabla}\log\tilde{\gamma}\rangle dw\\
&\quad-\int_{\mathbb{R}^d}\pi\langle\lambda\boldsymbol{w}+\frac{1}{\beta'_k}\boldsymbol{\nabla}\log\tilde{\gamma},\boldsymbol{\nabla}\log\pi-\boldsymbol{\nabla}\log\tilde{\gamma}\rangle dw\\
&\leq-\left(\frac{1}{\beta'_k}-\frac{1}{2C}\right)\int_{\mathbb{R}^d}\pi\|\boldsymbol{\nabla}\log\pi-\boldsymbol{\nabla}\log\tilde{\gamma}\|^2dw+\frac{C}{2}\int_{\mathbb{R}^d}\pi\|\boldsymbol{h}_t\|^2dw
\end{aligned}\tag{113}$$

As in the ideal case, we choose $C=\beta'_k$ and use logarithmic Sobolev inequality for the first term. The variance parameter in the inequality can vary through time. Fortunately, since τ_k is typically small, we can use worst-case upper bounds for this parameter, which is easy to obtain as $\tilde{\sigma}_t^2$ is monotonic in both cases.

$$\tilde{\sigma}_t^2\leq\begin{cases}\tilde{\sigma}_0^2+\frac{\tau_k}{\beta'_k}, & \lambda=0\\ \max\left(\tilde{\sigma}_0^2,\frac{1}{\beta'_k\lambda}\right), & \lambda>0\end{cases}\tag{114}$$

Using the ODE approach in the analysis for ideal case, we can obtain an upper bound for KL divergence after gradient update.

$$D_{KL}\left(p_{k+1}|_{\boldsymbol{g}_k}\middle|\middle|\gamma_{k+1}\right)\leq e^{-\frac{\tau_k}{2b_k}}D_{KL}\left(p_k||\gamma_k\right)+\frac{\beta'_k\tau_k}{2}\int_0^{\tau_k}\int_{\mathbb{R}^d}\pi_t\|\boldsymbol{h}_t(\boldsymbol{w})\|^2dwdt\tag{115}$$

For the last integral, we have:

$$\begin{aligned}
\int_{\mathbb{R}^d}\pi_t\|\boldsymbol{h}_t(\boldsymbol{w})\|^2dw=&\int_{\mathbb{R}^d}p(\boldsymbol{\theta}_t=\boldsymbol{w})\left\|\int_{\mathbb{R}^d}\frac{p(\boldsymbol{\theta}_t=\boldsymbol{w},\boldsymbol{\theta}_0=\boldsymbol{y})}{p(\boldsymbol{\theta}_t=\boldsymbol{w})}\boldsymbol{g}_k(\boldsymbol{y})dy\right\|^2dw\\
\leq&\int_{\mathbb{R}^d}\frac{1}{p(\boldsymbol{\theta}_t=\boldsymbol{w})}\left(\int_{\mathbb{R}^d}p(\boldsymbol{\theta}_t=\boldsymbol{w},\boldsymbol{\theta}_0=\boldsymbol{y})dy\right)\left(\int_{\mathbb{R}^d}p(\boldsymbol{\theta}_t=\boldsymbol{w},\boldsymbol{\theta}_0=\boldsymbol{y})\|\boldsymbol{g}_k(\boldsymbol{y})\|^2dy\right)dw\\
=&\mathbb{E}\|\boldsymbol{g}_k(\boldsymbol{w}_k)\|^2
\end{aligned}\tag{116}$$

By convexity of KL divergence,

$$D_{KL}\left(p_{k+1}\middle|\middle|\gamma_{k+1}\right)\leq\mathbb{E}\left(D_{KL}\left(p_{k+1}|_{\boldsymbol{g}_k}\middle|\middle|\gamma_{k+1}\right)\right)\leq e^{-\frac{\tau_k}{2b_k}}D_{KL}\left(p_k||\gamma_k\right)+\frac{\beta'_k\tau_k}{2}\mathbb{E}\|\boldsymbol{g}_k(\boldsymbol{w}_k)\|^2\tag{117}$$

■

Proof of Theorem 16

We actually prove a more general version, which allows arbitrary choice of initialization variance and regularization parameter.

Theorem 23 *(General version of Theorem 16) Assuming that for σ_k defined above, loss function $\ell(w;x)$ is s_k-subGaussian with respect to distribution $\mathcal{N}(0,\sigma_k^2 I_d)\times\mathcal{D}$. Assume that $f_i(w)$ is uniformly L-Lipschitz with respect to w. Assume $\eta_k\lambda\le\frac{1}{2},\forall k$. Given algorithmic parameters $N,\{\eta_k\},\beta,\sigma_0,\lambda$ fixed, the following inequalities uniformly holds for SGLD with probability $1-\delta$: (with respect to random draw of training data)*

$$\mathrm{err}_{gen}(\boldsymbol{w}_N)\le 2s_N\left(\frac{\beta}{n}\sum_{k=1}^N\eta_k e^{-R_{k,N}}\mathbb{E}\left[\|\boldsymbol{g}_k(\boldsymbol{w}_k)\|^2\right]+\frac{\log N/\delta+\log\log NL}{n}\right)^{\frac{1}{2}}\tag{118}$$

where the decaying factor $R_{k,N}$ is defined as follows:

- *If $\lambda=0$, $R_{k,N}=\sum_{j=k+1}^N\frac{\eta_j}{2\sigma_0^2\beta+6T_j}$.*
- *If $0<\lambda\le\frac{1}{\beta\sigma_0^2}$, $R_{k,N}=\frac{\lambda}{3}(T_N-T_k)$.*
- *If $\lambda>\frac{1}{\beta\sigma_0^2}$, $R_{k,N}=\begin{cases}\frac{\lambda}{4}(T_N-T_{k_1})+\frac{1}{2\beta\sigma_0^2}(T_{k_1}-T_k), & k<k_1\\ \frac{\lambda}{4}(T_N-T_k), & k\ge k_1,\end{cases}$*

where $k_1\triangleq\min\{k:T_k>\frac{1}{2\lambda}\ln(1+\frac{1}{2}\sigma_0^2\beta\lambda)\}$.

Proof Our analysis is divided into 3 cases based on choice of regularization parameter λ. Assuming that $\eta_k\lambda<0.5,\forall k$, the transformed parameters are at the same order with original ones, namely, $\frac{3}{4}\beta\le\beta_k'\le\beta$ and $\eta_k\le\tau_k\le 2\eta_k$.

Case I: $\lambda=0$.

In this case, the variance of each prior is $\sigma_k^2=\sigma_{k-1}^2+\frac{\tau_k}{\beta_k'}\le\sigma_0^2+\frac{4}{3\beta}\sum_{j=1}^k\tau_j$. So we have $b_k=\sigma_0^2\beta+\frac{4}{3}\sum_{j=1}^k\tau_j\le\sigma_0^2\beta+3\sum_{j=1}^k\eta_j$. By iteratively using Lemma 15, we get

$$D_{KL}(p_N||\gamma_N)\le\beta\sum_{k=1}^N\eta_k\exp\left(-\sum_{j=k+1}^N\frac{\eta_j}{2\sigma_0^2\beta+6\sum_{l=1}^j\eta_l}\right)\mathbb{E}\left[\|\boldsymbol{g}_k(\boldsymbol{w}_k)\|^2\right]\tag{119}$$

Case II: $0<\lambda\le\frac{3}{2\beta\sigma_0^2}$.

In this case, note that by construction,

$$\sigma_{k+1}^2=e^{-2\lambda\tau_k}\sigma_k^2+\frac{1-e^{-2\lambda\tau_k}}{\beta_k'\lambda}=(1-\lambda\eta_k)^2\sigma_k^2+\frac{2\eta_k}{\beta}$$

We can easily prove by a simple induction argument that $\forall k, \sigma_k^2 \leq \frac{3}{2\lambda\beta}$. So we have $b_k \leq \frac{3}{2\lambda}$ Using Lemma 15 iteratively, we have the following upper bound for KL divergence:

$$D_{KL}(p_N||\gamma_N) \leq \beta \sum_{k=1}^{N} \eta_k e^{-\frac{\lambda}{3}(T_N - T_k)} \mathbb{E}\left[\|\boldsymbol{g}_k(\boldsymbol{w}_k)\|^2\right] \tag{120}$$

where $T_k = \sum_{j=1}^{k} \eta_j$.

Case III: $\lambda > \frac{3}{2\beta\sigma_0^2}$**.**

In this case, using the fact that $(1 - \lambda\eta_k)^2 \leq e^{-2\lambda\eta_k}$, we can easily expand the iteration formula and get the upper bound:

$$\sigma_k^2 \leq e^{-2\lambda T_k}\sigma_0^2 + \frac{(1 - e^{-2\lambda T_k})}{(1 - e^{-1})\beta\lambda} \leq e^{-2\lambda T_k}\sigma_0^2 + \frac{2(1 - e^{-2\lambda T_k})}{\beta\lambda}.$$

And it is easy to see that $b_k \leq \sigma_{k-1}^2\beta$. For simplicity, we divide the procedure into two parts:

- For $T_k \leq \frac{1}{2\lambda}\ln(1 + \frac{1}{2}\sigma_0^2\beta\lambda)$, we have $\sigma_k^2 \leq \sigma_0^2$, and $b_{k+1} \leq \sigma_0^2\beta$.
- For $T_k > \frac{1}{2\lambda}\ln(1 + \frac{1}{2}\sigma_0^2\beta\lambda)$, we have $\sigma_k^2 \leq \frac{2}{\beta\lambda}$, and $b_{k+1} \leq \frac{2}{\lambda}$.

Let $k_1 \triangleq \min\{k : T_k > \frac{1}{2\lambda}\ln(1 + \frac{1}{2}\sigma_0^2\beta\lambda)\}$. We can obtain the KL divergence bound by treating two parts differently.

$$D_{KL}(p_N||\gamma_N) \leq \beta \sum_{k=1}^{k_1} \eta_k e^{-\frac{\lambda}{4}(T_N - T_{k_1}) - \frac{1}{2\beta\sigma_0^2}(T_{k_1} - T_k)} \mathbb{E}\left[\|\boldsymbol{g}_k(\boldsymbol{w}_k)\|^2\right] + \beta \sum_{k=k_1+1}^{N} \eta_k e^{-\frac{\lambda}{4}(T_N - T_k)} \mathbb{E}\left[\|\boldsymbol{g}_k(\boldsymbol{w}_k)\|^2\right] \tag{121}$$

In this case, the contribution of each round will first decay with a slower rate ($\frac{1}{2\beta\sigma_0^2}$ on the exponent). As variance for each prior becomes smaller along iterations, faster rate of decay with $\frac{\lambda}{4}$ on the exponent will be achieved.

Putting them together, we get the final PAC-Bayesian results. ■

Remarks:

- For the case of $\lambda = 0$, we can still get rid of the parameter norm dependence using a varying prior. In this case, the $\boldsymbol{\nabla}\log\gamma$ term cancels out with a term from time derivative of γ. However, we actually pay two prices for this norm-free properties: On the one hand, unless the loss class is uniformly bounded, the Orlicz norm for ℓ can grow with N since σ_N^2 grows linearly with N; On the other hand, the exponential decaying factor becomes significantly weakened. For example, if we take η_k to be a fixed constant, we will have $R_{k,N} \sim \frac{1}{6}\ln\frac{N}{k}$ and the rate of decaying factor is $(N-k)^{-\frac{1}{6}}$, which is much slower.
- Although the exponential decaying factor comes from the Gaussian initialization at the first glance. If we are trying to get bounds that are independent of parameter norm, the choice of initialization variance does not affect the time-varying prior very much. And the decaying factor will be eventually depending only on the ℓ_2 regularization parameter λ.