Time-independent Generalization Bounds for SGLD in Non-convex Settings

Anonymous Author(s)
Affiliation
Address
email

Abstract

We establish generalization error bounds for stochastic gradient Langevin dynamics (SGLD) with constant learning rate under the assumptions of dissipativity and smoothness, a setting that has received increased attention in the sampling/optimization literature. Unlike existing bounds for SGLD in non-convex settings, ours are time-independent and decay to zero as the sample size increases. Using the framework of uniform stability, we establish time-independent bounds by exploiting the Wasserstein contraction property of the Langevin diffusion, which also allows us to circumvent the need to bound gradients using Lipschitz-like assumptions. Our analysis also supports variants of SGLD that use different discretization methods, incorporate Euclidean projections, or use non-isotropic noise.

1 Introduction

Investigating the generalization error of a learning algorithm is a fundamental problem in machine learning that has motivated the development of a rich theory connecting notions of model complexity and sensitivity to the ability of a learning algorithm to generalize well to unseen data. Recently, there has been increased interest in studying the generalization capabilities of stochastic gradient descent (SGD) and its variants, e.g. [13, 1, 6, 17, 21, 19]. Despite them being some of the most important methods of modern machine learning and statistics, their learning capabilities have not been fully explored, particularly in non-convex settings.

One variant that has attracted a lot of attention is stochastic gradient Langevin dynamics (SGLD) [28]. With the addition of independent Gaussian noise to each iteration, SGLD combines the SGD framework with the Langevin diffusion, a stochastic process that converges to the Gibbs distribution of a given objective function. By tuning the scale of the noise applied, SGLD has been shown to work well as both a sampling scheme and a statistical learning algorithm in a variety of settings, e.g. [28, 24, 4, 14]. Using its relationship with the Langevin equation, various tools from stochastic analysis have been adopted to develop a rich theory analyzing SGLD in settings where no comparable results seem to be available for SGD, e.g. [22, 31, 5, 30]. In particular, a recent work from Raginsky et al. [22] provides non-asymptotic excess risk bounds for SGLD for a broad class of non-convex learning problems and these bounds were derived using the exponential ergodicity of Langevin diffusions. Following this work, there has been ongoing progress in the development of bounds for both optimization error and mixing times, e.g. [31, 5, 30, 7, 4]. Many of these works are motivated by recent results showing that Langevin diffusions contract in particular Wasserstein distances [8, 9].

In parallel to this literature, there has been a growing body of work that considers generalization error bounds for SGLD. Using uniform stability, Mou et al. [17] showed that under the assumption of bounded gradient updates and bounded loss functions, SGLD attains generalization bounds that grow slower with time than those that are known for SGD in identical settings [13, 6]. By bounding

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

the generalization error using the mutual information between its input and output [29], Pensia et al. [21] established similar results while replacing the assumption of bounded loss functions with the weaker assumption that the loss function is subgaussian. This technique has also been used to develop data-dependent generalization bounds that depend on the gradient along the trajectory of the iterates [18, 12, 17, 19].

Although the assumptions made in the aforementioned bounds cover several settings of interest, we are not aware of any convergence analysis that considers Lipschitz, bounded, or subgaussian objective functions. Certainly the Gibbs distribution, an object that is fundamental to the design and known convergence analyses for SGLD [28, 24, 22, 5, 31], cannot possibly be defined in the case that the objective is bounded over the Euclidean space as the distribution would not integrate to one. Similarly, we are not aware of any generalization bounds for SGLD that use the assumptions of dissipativity and smoothness that are consistently applied in the non-convex sampling/optimization literature, e.g. [22, 5, 31].

Another commonality in existing generalization bounds for SGLD is that they grow indefinitely with time. Sampling from the Gibbs distribution, the limiting behavior of the Langevin diffusion, is known to be uniformly stable in multiple settings [29, 22] and it has also been shown that the continuous-time Langevin dynamics are uniformly stable when applied to bounded Lipschitz functions with weight decay [16]. However, convergence in generalization error for SGLD seems to have only been observed in preliminary empirical evidence that uses data-dependent bounds [18, 12]. Time-independent bounds have been obtained for SGD, for example, under the assumption of strong convexity in [13] and in non-convex settings, Lei and Ying [15] obtain bounds that decrease with time but don't decay to zero as the sample size increases.

1.1 Our contributions

In this paper, we obtain generalization error bounds for SGLD for learning problems under dissipativity and smoothness assumptions. We consider the setting where the learning rate is constant and proportional to $n^{-\gamma}$, where n is the sample size and γ is a positive constant, a type of scaling that has been used in the SGLD literature [4]. We analyze both the discrete-time algorithm as well as a continuous-time version and, in both cases, we obtain bounds that are time-independent, in the sense that they are uniform over all iterations, and decay to zero as the sample size increases.

At first, we focus on the special case of Lipschitz loss functions with weight decay, which is the primary example of a dissipative objective given in Raginsky et al. [22]. For the continuous-time algorithm, we obtain a bound with rate $O(1/n)$ and for the discrete-time algorithm, we obtain a bound with the slower rate $O(\sqrt{d/n})$, where d is the model dimension. Then we extend the result to the full dissipative case without Lipschitz requirements. In this setting, we obtain dimension-free bounds with rate $O(1/n^{1/3})$ for the continuous-time algorithm and with rate $O(1/n^{1/6})$ in the discrete-time case. Our method allows for the consideration of different discretization techniques and we provide an example of a discretization that attains $O(\sqrt{d}/n)$ bounds in the Lipschitz setting. We also explain how our analysis can be used to obtain generalization error bounds for modifications of SGLD that incorporate Euclidean projections or non-isotropic noise.

Using the framework of uniform stability, in the sense defined in [10], we derive generalization bounds by estimating how much SGLD diverges in Wasserstein distance when an element of the data set is changed. We exploit recent results that use reflection couplings to show that under dissipativity-type assumptions, Langevin diffusions contract in Wasserstein distance. Though this property has been adopted extensively in the sampling/optimization literature, we are not aware of any results for generalization bounds based on this property. Using the convexity of Wasserstein distances, we combine this with the stability induced by using stochastic gradients to obtain bounds that are time-independent.

A peculiarity of our results that arises from the methodology we use is that we require the learning rate to be sufficiently large relative to n to facilitate the contractive behavior of the Langevin equation. This contrasts with the usual approach of stability-based generalization bounds based on non-expansivity that often require the learning rate to decay sufficiently fast or the time-horizon to be sufficiently small to guarantee bounds that are non-vacuous [13, 17].

2 Background and notation

2.1 Stability and generalization

In this paper, we consider a loss function $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ where the Euclidean space $\mathbb{R}^d$ represents the set of possible model parameters and $\mathcal{Z}$ represents the data instance space. A common objective in learning theory is to minimize the *population risk* which, given a data distribution P, is defined by

$$F_P(x) = \mathbb{E}_{Z\sim P} f(x, Z).$$

In practice, we often cannot compute $F_P(x)$ so we instead collect independent samples from P to form a *data set* and we compute the average loss, or *empirical risk*, over the data set. We will use the notation $S = (z_1, ..., z_n)$ to denote the data set and define the empirical risk as

$$F_S(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, z_i).$$

When the parameter is chosen by a random algorithm that depends on the data set, denoted $A(S)$, we define the object central to this paper, the *generalization error*, as follows:

$$\mathrm{gen}(A, S) = F_P(A(S)) - F_S(A(S)).$$

To bound this quantity we employ the following notion of *uniform stability*.

Definition 2.1 ([13], Definition 2.1)**.** *An algorithm A is ε-uniformly stable if*

$$\varepsilon_{stab}(A) := \sup_{S\cong\hat{S}} \sup_{z\in\mathcal{Z}} \mathbb{E}\big[f(A(S), z) - f(A(\hat{S}), z)\big] \le \varepsilon.$$

where the first supremum is over data sets $S, \hat{S} \in \mathcal{Z}^n$ that differ by one element, denoted by $S \cong \hat{S}$.

The connection between generalization and stability under changes in the data set has received increased attention since the paper of Bousquet and Elisseeff [2], and these results have been extended to account for random algorithms by Elisseeff et al. [10]. The precise notion of stability that we consider in this paper is given in Hardt et al. [13].

Proposition 2.2 ([13], Theorem 2.2)**.** *Suppose A is an ε-uniformly stable algorithm, then the expected generalization error is bounded by*

$$\left|\mathbb{E}_{A,S}\,\mathrm{gen}(A, S)\right| \le \varepsilon.$$

Though we will present the results of this paper as bounds on the expected generalization error, uniform stability bounds can also give high probability bounds for the generalization error [11, 3]. Outside of generalization bounds, the concept of uniform stability has also been shown to be fundamentally related to differential privacy and learnability [27, 23].

2.2 Stochastic gradient Langevin dynamics

We will now define the algorithm of interest in this paper. Given a *mini-batch* $B \subset [n] := \{1, ..., n\}$, let the *mini-batch average* be defined by

$$F_S(x, B) = \frac{1}{|B|}\sum_{i\in B} f(x, z_i).$$

In our analysis we will consider uniformly sampled random mini-batches of fixed size $k \le n$. We now define SGLD, which given an initial distribution ρ_0 is given by

$$x_{m+1} = x_m - \eta\nabla F_S(x_m, B_{m+1}) + \sqrt{2\beta^{-1}\eta}\,\xi_{m+1}, \quad x_0 \sim \rho_0, \tag{1}$$

where $(B_m)_{m=1}^{\infty}$ is an i.i.d. sequence of random variables distributed uniformly on the set $\{B \subset [n] : |B| = k\}$ and $(\xi_m)_{m=1}^{\infty}$ is an i.i.d. sequence of $N(0, I_d)$ random variables. The parameters $\eta, \beta > 0$ are tunable and are referred to as the *learning rate* and *inverse temperature*.

If $k = n$ then (1) describes the Euler-Maruyama discretization of the (overdamped) Langevin equation with potential $F_S(\cdot)$. This is a stochastic differential equation of the form

$$dX_t = -\nabla F_S(x)dt + \sqrt{2\beta^{-1}}dW_t, \quad X_0 \sim \rho_0,$$

where W_t is a d-dimensional Wiener process [20]. In the more common case that $k < n$ we can still define a stochastic differential equation that approximates SGLD, given by

$$dX_t = -\nabla F_S(X_t, B_{\lceil t/\eta \rceil})dt + \sqrt{2\beta^{-1}}dW_t, \quad X_0 \sim \rho_0. \tag{2}$$

We will refer to the solution X_t as the *continuous-time* SGLD algorithm. Note that, under the smoothness assumptions imposed throughout this paper, all stochastic differential equations considered have strong solutions (see Theorem 3.1 of [20]).

Since x_m is a Markov process, we can define its Markov kernel R_x. Though, we will not frequently reference this fact, we will use the notation μR_x^m to denote the law of x_{m+t} under the condition $x_t \sim \mu$. Because of its dependence on B_k, X_t is not necessarily a Markov process in continuous-time, but the discrete-time process $(X_{k\eta})_{k=0}^{\infty}$ does in fact satisfy the Markov property and so we denote its kernel by R_X.

Given a deterministic set $B \subset [n]$ we use the notation μP_t^B to denote the law of θ_t, the solution to the SDE,

$$d\theta_t = -\nabla F_S(\theta_t, B)dt + \sqrt{2\beta^{-1}}dW_t, \quad \theta_0 \sim \mu,$$

which is a Markov (diffusion) process. Thus μR_X can be computed by integrating μP_η^B over B with respect to the mini-batch distribution.

Since this work is concerned with stability under changes in the data set, we will also be interested in SGLD when S is replaced with some data set $\hat{S}$ that differs by a single element. We will use the notation $\hat{x}_m$ and $\hat{X}_t$ to denote the respective counterparts of x_m of X_t when trained using $\hat{S}$ instead of S and similarly we will use $\hat{R}_x$, $\hat{R}_X$ and $\hat{P}_t^B$.

2.3 Wasserstein distance

For probability measures μ, ν on $\mathbb{R}^d$ with finite p^{th} moment, we define the *Wasserstein distance*

$$\mathrm{W}_p(\mu, \nu) = \Big(\inf_{\pi \in \mathcal{C}(\mu,\nu)} \int \|x - y\|^p \, \pi(dx, dy) \Big)^{1/p},$$

where $\|\cdot\|$ is the Euclidean norm. Here, the infimum is over all couplings of μ and ν, that is the set of all probability measures with marginals μ and ν. In our analysis, we will also consider Wasserstein distances based on any semimetric d, which under certain integrability assumptions is defined by

$$\mathrm{W}_d(\mu, \nu) = \inf_{\pi \in \mathcal{C}(\mu,\nu)} \int d(x, y) \, \pi(dx, dy).$$

Note that throughout this paper all probability measures considered, outside of the data distribution, are Borel probability measures on $\mathbb{R}^d$.

A property of the Wasserstein distance that will prove to be of central importance in our results is its convexity.

Lemma 2.3 (Convexity of the Wasserstein distance)**.** *Suppose that d is a semimetric and $\mu_1, \mu_2, \nu_1, \nu_2$ are probability measures. Then, for any $r \in [0, 1]$,*

$$\mathrm{W}_d(\mu, \nu) \leq r \, \mathrm{W}_d(\mu_1, \nu_1) + (1 - r) \, \mathrm{W}_d(\mu_2, \nu_2),$$

where we define $\mu(dx) = r\mu_1(dx) + (1 - r)\mu_2(dx)$ and $\nu(dx) = r\nu_1(dx) + (1 - r)\nu_2(dx)$.

The proof of this result, as well as a more general statement, can be found in [25], Theorem 4.8.

3 The Lipschitz setting with weight decay regularization

To begin our analysis, we focus on the special case of dissipativity. We will assume that the loss function $f(x, z)$ is Lipschitz in x for any fixed z, and when performing gradient updates we will use the empirical risk with an added weight decay term:

$$\tilde{F}_S(x, B) = F_S(x, B) + \frac{\lambda}{2}\|x\|.$$

This setting has been considered in the data-dependent literature for SGLD, for example by Mou et al. [17], and under the additional assumption of boundedness, this setting has also been considered in the stability-based analysis of the continuous Langevin dynamics by Li et al. [16].

More formally, the following assumptions are imposed:

Assumption 3.1. *For each $z \in \mathcal{Z}$, $f(\cdot, z)$ is L-Lipschitz: for all $x_1, x_2 \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

$$|f(x_1, z) - f(x_2, z)| \leq L\|x_1 - x_2\|.$$

Assumption 3.2. *For each $z \in \mathcal{Z}$, $f(\cdot, z)$ is M-smooth: for all $x_1, x_2 \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

$$\|\nabla f(x_1, z) - \nabla f(x_2, z)\| \leq M\|x_1 - x_2\|.$$

Assumption 3.3. *The initial condition ρ_0 has finite first moment $\sigma_1 := \rho_0(\|\cdot\|) < \infty$.*

Assumption 3.3 is used to guarantee that SGLD has finite first moment so that we can perform our analysis using the 1-Wasserstein distance. Note that the weight decay term only affects the gradient updates of SGLD and is not applied to the generalization error, which is still defined in terms of F_S and F_P.

Theorem 3.1. *Suppose Assumptions 3.1–3.3 hold. If $\eta \in (0, 1)$ and $n \geq C_1 k/\eta$, then the continuous-time algorithm attains the generalization bound*

$$\sup_{m \in \mathbb{N}} |\mathbb{E}\,\mathrm{gen}(X_{m\eta}, S)| < \frac{C_2}{n}.$$

If for any $\alpha \geq C_1$ the learning rate is set to $\eta = \alpha k/n$ and $n \geq \alpha k/(1 \vee \lambda)$, then the discrete-time algorithm attains the generalization bound

$$\sup_{m \in \mathbb{N}} |\mathbb{E}\,\mathrm{gen}(x_m, S)| < C_3 \frac{1 + \sqrt{\alpha d}}{\sqrt{n}}.$$

The positive constants $C_1 \equiv C_1(M, L, \lambda, \beta), C_2 \equiv C_2(M, L, \lambda, \beta), C_3 \equiv C_3(M, L, \lambda, \beta, \sigma_1)$ are given in the proof in equations (6), (7) and (8) respectively.

Note that the bound for the continuous-time algorithm has no explicit dependence on dimension and in the discrete-time case the bound grows with the rate $\sqrt{d}$. However, in the case $\lambda < M$, the constants C_1, C_2, C_3 depend exponentially on the parameters M, L, λ, β. A similar dependence was found in the time-independent bounds for the Langevin equation [16] and it is also found in convergence analyses in non-convex settings leading to the time horizon required to attain a certain guaranteed accuracy having to scale exponentially with parameters including d and β, e.g. [22, 5, 31].

3.1 Methodology

In this section we will summarize the method used to obtain these bounds. The main technical contribution that leads to Theorem 3.1 is summarized in the following proposition:

Proposition 3.2. *Suppose Assumptions 3.1-3.3 hold, $\eta \in (0, 1)$ and $n \geq C_1 k/\eta$. Then, for any $m \in \mathbb{N}$,*

$$\mathrm{W}_1(\rho_0 R_X^m, \rho_0 \hat{R}_X^m) \leq \frac{C_2}{Ln}.$$

If we also assume that $\eta \leq \lambda^{-1}$, then

$$\mathrm{W}_1(\rho_0 R_x^m, \rho_0 \hat{R}_x^m) \leq \frac{C_3}{L}\left(\frac{1}{n} + \sqrt{d\eta}\right). \tag{3}$$

Under assumption 3.1, such bounds are sufficient to obtain uniform stability bounds. Indeed, it follows directly from the definition of ε_{stab} that

$$\varepsilon_{stab}(A) \leq L \sup_{S \cong \hat{S}} \mathrm{W}_1\left(law(A(S)),\, law(A(\hat{S})\right). \tag{4}$$

A property similar to (4) has been exploited in [22], where it was used to prove the stability of the Gibbs algorithm.

To prove Proposition 3.2, we begin with a weak estimate for the divergence in 1-Wasserstein distance using synchronous couplings between X_t and $\hat{X}_t$, that is, we couple the processes by having them share the same Brownian motion. This is useful for analyzing how much diffusion processes diverge since we can derive equations of the following form:

$$X_t - \hat{X}_t = X_0 - \hat{X}_0 + \int_0^t \Big(\nabla F_S(\hat{X}_s, B_{\lceil s/\eta \rceil}) - \nabla F_S(X_s, B_{\lceil s/\eta \rceil})\Big) ds.$$

By exploiting the smoothness and Lipschitz assumptions and applying Grönwall's inequality, we obtain the following result.

Lemma 3.3. *Suppose Assumptions 3.1 and 3.2 hold, then for any two probability measures μ, ν on $\mathbb{R}^d$ we have*

$$\mathrm{W}_1(\mu P_t^B, \nu \hat{P}_t^B) \leq \exp((\lambda + M)t)\, \mathrm{W}_1(\mu, \nu) + \mathbb{1}(n \in B) \exp((\lambda + M)t) \frac{2Lt}{k}.$$

Proof. The proof is given in Appendix B.1. □

In the case $\hat{P}^B = P^B$, we can obtain sharper bounds using reflection couplings. In a recent result from Eberle [8], reflection couplings were used to show that, under conditions not dissimilar from dissipativity, diffusion processes contract in Wasserstein distance. The Wasserstein distance considered in this result does not use a standard metric however, it uses the metric $d_g(x, y) := g(\|x - y\|)$ where g is a strictly-increasing concave function that is constructed depending on the objective function (or drift term) used. In the lemma that follows we give a special case of such a result in the setting that we consider. Throughout this section we use the notation $\mathrm{W}_g(\mu, \nu) := \mathrm{W}_{d_g}(\mu, \nu)$.

Lemma 3.4. *Suppose Assumptions 3.1 and 3.2 hold. Then there exists a strictly-increasing concave function $g : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ such that for any two probability measures μ, ν on $\mathbb{R}^d$, any $B \subset [n]$ and $t \geq 0$, we have*

$$\mathrm{W}_g(\mu P_t^B, \nu P_t^B) \leq e^{-ct}\, \mathrm{W}_g(\mu, \nu), \quad c = \begin{cases} (c_1 c_2)^{-1}, & \text{if } \lambda < M, \\ c_3^{-1}, & \text{if } \lambda \geq M, \end{cases}$$

with constants $c_1 = \exp[2\beta L^2(M - \lambda)/\lambda^2]$, $c_2 = 8(L^2\beta/\lambda + 1)/\lambda$, $c_3 = \max(16L^2\beta/\lambda^2, 2/\lambda)$. Furthermore, W_g is equivalent to the 1-Wasserstein distance in the sense that

$$\frac{1}{2\max(c_1, 1)}\, \mathrm{W}_1(\mu, \nu) \leq \mathrm{W}_g(\mu, \nu) \leq \mathrm{W}_1(\mu, \nu).$$

Proof. We refer to Appendix C for the proof and a broader discussion on this result. □

In the case that n is not in the random mini-batch, i.e. $n \notin B$, which occurs with probability $1 - k/n$, it follows that $\hat{P}^B = P^B$ and so the processes $X_{k\eta}$ and $\hat{X}_{k\eta}$ contract with respect to the Wasserstein distance W_g. In the case that n is in the random mini-batch, $n \in B$, Lemma 3.3 gives uniform bounds for how much $X_{k\eta}$ and $\hat{X}_{k\eta}$ can diverge. The remainder of the proof works to combine these facts and shows that if n is sufficiently large, $X_{k\eta}$ and $\hat{X}_{k\eta}$ can only diverge by some fixed amount. We combine the two cases using the convexity of the Wasserstein distance and we find that given a sufficiently large n, the resulting bound does indeed converge over time.

Before presenting the proof, we state the lemma that allows this result to be extended to the discrete-time case by bounding the amount to which the continuous-time and discrete-time algorithms diverge in a single step.

Lemma 3.5. *Suppose Assumptions 3.1 and 3.2 hold and $\lambda\eta < 1$. Then, for any probability measure μ on $\mathbb{R}^d$, we have*

$$\mathrm{W}_1(\mu R_x, \mu R_X) \leq \eta(\lambda + M)\Big[\eta(\lambda\mu(\|\cdot\|) + 2L) + \sqrt{2d\beta^{-1}\eta}\Big] \exp(M + 1).$$

Furthermore, for each $m \in \mathbb{N}$,

$$\mu R_x^m(\|\cdot\|) \leq \mu(\|\cdot\|) + \frac{L + \sqrt{2\beta^{-1}d\eta^{-1}}}{\lambda}.$$

Proof. The proof is given in Appendix B.2. □

The proof of Proposition 3.2 is given in sketch form below; the precise argument can be found in Appendix D.

Proof of Proposition 3.2 (sketch). Since the probability of a random mini-batch containing the element n is k/n, Lemma 2.3 is used to bound the Wasserstein distance with the decomposition

$$\mathrm{W}_g(\mu R_X, \nu \hat{R}_X) \leq \frac{k}{n} \sup_{B:n\in B} \mathrm{W}_g(\mu P_\eta^B, \nu \hat{P}_\eta^B) + \left(1 - \frac{k}{n}\right) \sup_{B:n\notin B} \mathrm{W}_g(\mu P_\eta^B, \nu \hat{P}_\eta^B).$$

Bounding the first term using Lemma 3.3 and the second term using Lemma 3.4, it follows that

$$\mathrm{W}_g(\mu R_X, \nu \hat{R}_X) \leq \tilde{c}_1 \, \mathrm{W}_g(\mu, \nu) + \tilde{c}_2, \tag{5}$$

where $\tilde{c}_1 = \frac{k}{n} \exp((\lambda + M)\eta) + e^{-c\eta}$ and $\tilde{c}_2 = \exp((\lambda + M)\eta)\frac{2L\eta}{n}$. It follows by induction that

$$\mathrm{W}_g(\rho_0 R_X^m, \rho_0 \hat{R}_X^m) \leq \sum_{t=1}^{m} \tilde{c}_1^{m-t} \tilde{c}_2 + \tilde{c}_1^m \, \mathrm{W}_g(\rho_0, \rho_0) < \frac{1 - \tilde{c}_1^m}{1 - \tilde{c}_1} \tilde{c}_2,$$

and thus it is sufficient to guarantee $\tilde{c}_1 < 1$ to obtain time-independent bounds. This follows from the assumption $n \geq C_1(M, L, \lambda, \beta) k \eta^{-1}$ with

$$C_1(M, L, \lambda, \beta) := 2 \exp(\lambda + M)(1 - e^{-c})^{-1}. \tag{6}$$

The first bound in the statement of the proposition is obtained once it is shown that $\tilde{c}_2/(1 - \tilde{c}_1)$ is bounded above by C_2/Ln where

$$C_2(M, L, \lambda, \beta) := \frac{8L^2 \max\{c_1, 1\} \exp(\lambda + M)}{(1 - e^{-c})}. \tag{7}$$

This result is extended to the discrete-time case by applying the triangle inequality,

$$\mathrm{W}_g(\mu R_x, \nu \hat{R}_x) = \mathrm{W}_g(\mu R_x, \mu R_X) + \mathrm{W}_g(\mu R_X, \nu \hat{R}_X) + \mathrm{W}_g(\nu \hat{R}_X, \nu \hat{R}_x).$$

The first and third terms are bounded using the discretization error bound in Lemma 3.5 and an identical argument is applied to obtain the bound,

$$\mathrm{W}_g(\rho_0 R_x^m, \rho_0 \hat{R}_x^m) \leq \tilde{c}_1 \, \mathrm{W}_g(\rho_0 R_x^{m-1}, \rho_0 \hat{R}_x^{m-1}) + \tilde{c}_2 + \tilde{c}_3 \leq \frac{\tilde{c}_2 + \tilde{c}_3}{1 - \tilde{c}_1}.$$

where $\tilde{c}_3 = \eta^{3/2}(\lambda + M)(\lambda \rho_0(\|\cdot\|) + 3L + 2\sqrt{2d\beta^{-1}}) \exp(M + 1)$. The right-hand side is bounded by $C_3(n^{-1} + \sqrt{\eta d})/L$ where

$$C_3(M, L, \lambda, \beta, \sigma_1) := \frac{8 \max\{c_1, 1\} \exp(M + 1)}{1 - e^{-c}} \Big(L + (\lambda + M) \Big(\lambda \sigma_1 + 3L + 2\sqrt{2\beta^{-1}} \Big) \Big). \tag{8}$$

□

Remark 3.6. *In both the analysis of the continuous-time algorithm and the discrete-time algorithm we bound the term $1 - \tilde{c}_1^m < 1$. Since $\tilde{c}_1 \geq e^{-c\eta}$, we can instead use the bound $1 - \tilde{c}_1^m \leq 1 - e^{-c\eta m} \leq c\eta m$. With this, we can sharpen our generalization bounds for the case of $m \ll 1/\eta$ for which both bounds have the rate ηm.*

Remark 3.7. *For the discrete-time algorithm, the bound is obtained by simply appending the one-step discretization error to the bound in (5) for the continuous-time algorithm. Thus the result can easily be extended to a broader range of discretizations of $(X_t)_{t\geq 0}$. Suppose K_x is a Markov kernel on $\mathbb{R}^d$ such that for any probability measure μ on $\mathbb{R}^d$, $\mathrm{W}_1(\mu K_x, \mu R_X) \leq \delta(\eta)$. Then, we immediately obtain the bound*

$$\mathrm{W}_1(\rho_0 K_x^m, \rho_0 \hat{K}_x) \leq \frac{C_2}{Ln} + \frac{2\delta(\eta)}{\eta(1 - e^{-c})}.$$

In the case of the standard discrete-time algorithm we obtain $\delta(\eta) = O(\eta^{3/2})$ but if we consider a discretization that attains $O(\eta^2)$ it would be possible to derive bounds with $1/n$ rates. For example, if instead of approximating X_η with a single Euler-Maruyama step we use $T = \lfloor 1/\eta \rfloor$ steps with step-size $\gamma = \eta/T$, then we can show that $\delta(\eta) \leq O(\sqrt{\gamma}) \leq O(\eta^2)$. This is presented more formally and proven in Appendix D.

Remark 3.8. *With our approach, it is possible to derive bounds for projected versions of SGLD. If $\Omega \subset \mathbb{R}^d$ is a compact set, we can define the algorithm,*

$$x_{m+1}^{\Pi} = \Pi_\Omega\Big[x_m^{\Pi} - \eta \nabla F_S(x_m, B_{m+1}) + \sqrt{2\beta^{-1}\eta}\zeta_{m+1}\Big], \tag{9}$$

where Π_Ω is the Euclidean projection onto the set Ω. A well-known property that is exploited in the optimization literature is that for any $x_1, x_2 \in \mathbb{R}^d$, $\|\Pi_\Omega(x_1) - \Pi_\Omega(x_2)\| \leq \|x_1 - x_2\|$ and, given two measures μ, ν on $\mathbb{R}^d$, this naturally extends to

$$\mathrm{W}_1(\mu^*\Pi_\Omega, \nu^*\Pi_\Omega) \leq \mathrm{W}_1(\mu, \nu),$$

where $\mu^*\Pi_\Omega$ denotes the push-forward of μ with respect to the mapping Π_Ω. If we let R_x^{Π} be the kernel of the process given in (9)*, it follows immediately that,*

$$\mathrm{W}_1(\mu R_x^{\Pi}, \nu \hat{R}_x^{\Pi}) \leq \mathrm{W}_1(\mu R_x, \nu \hat{R}_x),$$

and hence the Wasserstein distance bound in (3) also applies to the projected algorithm and therefore, the generalization bound in Theorem 4.1 also holds for the projected algorithm.

Remark 3.9. *Our technique supports the case where, instead of assuming ξ_m is a standard Gaussian random vector, we set $\xi_m \sim N(0, \Sigma)$ for some symmetric positive semidefinite matrix $\Sigma \in \mathbb{R}^{d\times d}$. If the operator norm of Σ is 1, the conclusions of Lemma 3.3 and Lemma 3.4 still hold true and, as a result, the conclusion of Theorem 3.1 would also hold.*

4 The dissipative setting

Now we extend the results of the previous section to the full dissipative smooth case.

Assumption 4.1. *For each $z \in Z$, $f(\cdot, z)$ is (m, b)-dissipative: for all $x \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

$$\langle \nabla f(x, z), x\rangle \geq m\|x\|^2 - b.$$

Assumption 4.2. *Same as assumption 3.2.*

Assumption 4.3. *The initial condition ρ_0 has finite fourth moment $\sigma_4 := \rho_0(\|\cdot\|) < \infty$.*

Without the Lipschitz assumption, the analysis is fundamentally more difficult as its no longer sufficient to control the uniform argument stability or 1-Wasserstein distance. For this reason, we require n to grow faster with η than in the previous theorem.

Theorem 4.1. *Suppose Assumptions 4.1–4.3 hold and $\beta \geq 1$. If $\eta \in (0, 1)$ and $n \geq C_4 k^2/\eta^{3/2}$, then the continuous-time algorithm attains the generalization bound*

$$\sup_{m\in\mathbb{N}} |\mathbb{E}\,\mathrm{gen}(X_{m\eta}, S)| < \frac{C_5}{(nk)^{1/3}}.$$

If for any $\alpha \geq C_4$ the learning rate is set to $\eta = (\alpha k^2/n)^{2/3}$ and $n \geq \alpha k^2/(1 \vee m^{3/2}/8M^3)$, then the discrete-time algorithm attains the generalization bound

$$\sup_{m\in\mathbb{N}} |\mathbb{E}\,\mathrm{gen}(x_m, S)| < C_6 \frac{1 + k^{1/3}\alpha^{1/6}}{n^{1/6}}.$$

The constants $C_4 \equiv C_4(M, m, b, d/\beta), C_5 \equiv C_5(M, m, b, d/\beta, \sigma_4), C_6 \equiv C_6(M, m, b, d/\beta, \sigma_4)$ are given in the proof of Proposition 4.2.

Similar to Theorem 3.1, the constants C_4, C_5, C_6 depend exponentially on parameters $M, m, b, d/\beta$ but unlike Theorem 3.1, if we take $\beta \propto d$, both bounds have no explicit dependence on dimension or β.

4.1 Methodology

Without the Lipschitz assumption it is no longer sufficient to control the 1-Wasserstein distance, however, under the assumption of dissipativity and smoothness we can derive uniform stability bounds by estimating the 2-Wasserstein distance. This guides our main technical contribution for this setting which is stated in Proposition 4.2 below. Since this proposition relies on similar techniques to Proposition 3.2, we will postpone the proof to the appendix and in this section, we will focus on the challenges faced when extending to the dissipative setting.

Proposition 4.2. *Suppose Assumptions 4.1-4.3 hold, $\eta \in (0,1)$ and $n \geq C_4 k^2/\eta^{3/2}$, then for any $m \in \mathbb{N}$,*

$$\mathrm{W}_2(\rho_0 R_X^m, \rho_0 \hat{R}_X^m) \leq \frac{C_5}{M(\sqrt{\sigma_4} + \sqrt{b/m})n^{1/3}}.$$

If we also assume that $\eta \leq 1 \wedge (m/4M^2)$, then

$$\mathrm{W}_2(\rho_0 R_x^m, \rho_0 \hat{R}_x^m) \leq \frac{C_6}{M(\sqrt{\sigma_4} + \sqrt{b/m})}\left(\frac{1}{(nk)^{1/3}} + \eta^{1/4}\right).$$

Proof. The proof is given in appendix E. □

We show that this result can be used to derive stability bounds in Lemma E.1 where, under Assumptions 4.1, 4.2 and $\mathbb{E}\|A(S)\|^2 \leq s^2$ for all $S \in \mathcal{Z}^n$, we derive the following bound:

$$\varepsilon_{stab}(A) \leq M\big(s + \sqrt{b/m}\big) \sup_{S \cong \hat{S}} \mathrm{W}_2\left(law(A(S)),\, law(A(\hat{S}))\right).$$

The proof of Proposition 4.2 proceeds by establishing similar results to Lemma 3.3 and Lemma 3.4 using synchronous and reflection couplings. As has been noted in [8] and [26], contractions in 2-Wasserstein distance are more difficult to obtain than in 1-Wasserstein distance. The best known contraction result under the assumptions we consider is from Eberle et al. [9] and uses a Wasserstein distance W_d based on a semimetric d. Since W_d is not a metric, we don't have access to the triangle inequality and it is not possible to show W_d is equivalent to any standard Wasserstein distances, as was done in Lemma 3.4. Instead, to convert between bounds in W_d and the 2-Wasserstein distance we use the inequality,

$$\frac{C_{10}}{2}\,\mathrm{W}_2(\mu,\nu)^2 \leq \mathrm{W}_d(\mu,\nu) \leq C_{11}\,\mathrm{W}_2(\mu,\nu)\Big[1 + \mu(V_4) + \nu(V_4)\Big].$$

It is these obstacles that result in the slower $1/n^{1/3}$ rates in the continuous-time case, as opposed to the $1/n$ rates obtained in the Lipschitz setting with weight decay.

Remark 4.3. *As stated in Remark 3.6, our method does give rates with respect to m for $m \ll 1/\eta$. Here, both bounds have rate $O(\sqrt{\eta m})$ which is the same rate that Mou et al. [17] and Pensia et al. [21] obtain while under Lipschitz assumptions.*

5 Conclusion

In this paper, we derive time-independent generalization error bounds for SGLD under the assumptions of dissipativity and smoothness. We obtain bounds scaling as $O(1/n^{1/3})$ for the continuous-time algorithm and scaling as $O(1/n^{1/6})$ in the discrete-time case. In the special case of Lipschitz loss functions with weight decay regularization, we obtain faster rates of $O(1/n)$ for the continuous-time algorithm and $O(\sqrt{d/n})$ for the discrete-time case.

Within the framework of uniform stability, we use a combination of synchronous and reflection couplings to control the Wasserstein distance between versions of the algorithm with perturbations on the data set. Using the convexity of the Wasserstein distance, we show that once n is taken to be sufficiently large relative to η, the two versions of the algorithm can only diverge by a fixed amount that decays to zero as n increases.

The methodology used in our analysis allows for the extension to a broader class of discretizations and in the special case of Lipschitz smooth functions with weight decay we give an example of a discrete-time algorithm that attains bounds scaling as $O(\sqrt{d}/n)$. We are also able to explore modifications of the algorithm that incorporate Euclidean projections or non-isotropic Gaussian noise.

The fact that our proof relies heavily on properties of the Langevin diffusion does introduce some shortcomings. For example, the discretization error introduced when considering the discrete-time algorithm leads to an explicit dependence on the dimension and, furthermore, the dependence on model parameters aside from d is often exponential. In addition, Gaussian noise is fundamental in the construction and analysis of the Langevin diffusion, so it would be difficult to extend this analysis to algorithms where different types of noise are used.

References

[1] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems*, volume 33, pages 4381–4391, 2020.

[2] Olivier Bousquet and André Elisseeff. Stability and Generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[3] Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 610–626. PMLR, 2020.

[4] Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient langevin dynamics. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[5] Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient langevin dynamics with dependent data streams: the fully non-convex case. *arXiv preprint arXiv:1905.13142*, 2019.

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

[7] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

[8] Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166(3-4):851–886, 2013.

[9] Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Quantitative Harris-type theorems for diffusions and McKean–Vlasov processes. *Transactions of the American Mathematical Society*, 371(10):7135–7173, 2018.

[10] Andre Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of Randomized Learning Algorithms. *Journal of Machine Learning Research*, 6(3):55–79, 2005.

[11] Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1270–1279. PMLR, 2019.

[12] Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened Generalization Bounds based on Conditional Mutual Information and an Application to Noisy, Iterative Algorithms. In *Advances in Neural Information Processing Systems*, volume 33, pages 9925–9935, 2020.

[13] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *33rd International Conference on Machine Learning, ICML 2016*, 3:1868–1877, 2015.

[14] Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. Distributed Bayesian Learning with Stochastic Natural Gradient Expectation Propagation and the Posterior Server. *J. Mach. Learn. Res.*, 18(1): 3744–3780, 2017.

[15] Yunwen Lei and Yiming Ying. Fine-Grained Analysis of Stability and Generalization for Stochastic Gradient Descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5809–5819. PMLR, 2020.

[16] Jian Li, Xuanyuan Luo, and Mingda Qiao. On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning. In *International Conference on Learning Representations*, 2020.

[17] Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 605–638. PMLR, 2018.

[18] Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[19] Gergely Neu. Information-theoretic generalization bounds for stochastic gradient descent. *arXiv preprint arXiv:2102.00931*, 2021.

[20] Grigorios A. Pavliotis. *Stochastic Processes and Applications*, volume 60 of *Texts in Applied Mathematics*. Springer New York, 2014.

[21] Ankit Pensia, Varun Jog, and Po Ling Loh. Generalization Error Bounds for Noisy, Iterative Algorithms. In *IEEE International Symposium on Information Theory - Proceedings*, volume 2018-June, pages 546–550. Institute of Electrical and Electronics Engineers Inc., 2018.

[22] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis. In *Proceedings of Machine Learning Research*, volume 65, pages 1–30. PMLR, 2017.

[23] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[24] Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and Fluctuations For Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research*, 17(7):1–33, 2016.

[25] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2009.

[26] Feng Yu Wang. Exponential Contraction in Wasserstein Distances for Diffusion Semigroups with Negative Curvature. *Potential Analysis*, 53(3):1123–1144, 2020.

[27] Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. *The Journal of Machine Learning Research*, 17(1):6353–6392, 2016.

[28] M Welling and YW Teh. Bayesian learning via stochastic gradient langevin dynamics. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, 2011.

[29] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[30] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[31] Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for stochastic gradient langevin dynamics under local conditions in nonconvex optimization. *arXiv preprint arXiv:1910.02008*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The claims made in the abstract are focused around Theorems 3.1 and 4.1 and they are portrayed accurately. In the abstract, we explain the dependence on key quantities: the sample size and the number of iterations, and later in the contributions section we also include the dependence on dimension. In both the abstract and contributions section we state the assumptions used but we omit the detail that the initial distribution has to have finite fourth moment, but this is mostly a technical detail and is satisfied for most common initial distributions (e.g. Gaussian distribution).
   (b) Did you describe the limitations of your work? [Yes] The shortcomings are succinctly put forward in the conclusion, section 5. Also after the statement of both theorems 3.1 and 4.1, we emphasize the fact that there are cases where the bound grows exponentially with model parameters aside from d.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is a theoretic analysis of an existing algorithm for stochastic optimization problems and it does not propose anything beyond minor modifications to this algorithm. The quantity of interest, the generalization error, is not controversial and without extrapolating to some unreasonable prediction we see no setting where knowledge of the generalization error will be misused to, say, justify the use of an algorithm in an unethical manner.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See assumptions 3.1–3.3 and 4.1–4.3. In every theorem, proposition and lemma that depends on these assumptions, we make the dependence explicitly clear in its statement.
   (b) Did you include complete proofs of all theoretical results? [Yes] The complete proofs are contained in the appendices. For Theorem 3.1, one of the two main theorems, a full sketch proof is provided in section 3.1. For Theorem 4.1, the techniques are broadly similar so we discuss the differences and the challenges faced in extending the result in section 4.1.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] No experiments were performed in this paper.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] No experiments were performed in this paper.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] No experiments were performed in this paper.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] No experiments were performed in this paper.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] Our work did not involve the use of any assets.
   (b) Did you mention the license of the assets? [N/A] Our work did not involve the use of any assets.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] Our work did not involve the use of any assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Our work did not involve the use of any assets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Our work did not involve the use of any assets.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Our work did not involve crowdsourcing or conducted research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Our work did not involve crowdsourcing or conducted research with human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Our work did not involve crowdsourcing or conducted research with human subjects.