Generalization Bounds for Label Noise Stochastic Gradient Descent

Jung Eun Huh
Department of Statistics
University of Oxford

Patrick Rebeschini
Department of Statistics
University of Oxford

Abstract

We develop generalization error bounds for stochastic gradient descent (SGD) with label noise in non-convex settings under uniform dissipativity and smoothness conditions. Under a suitable choice of semimetric, we establish a contraction in Wasserstein distance of the label noise stochastic gradient flow that depends polynomially on the parameter dimension d. Using the framework of algorithmic stability, we derive time-independent generalisation error bounds for the discretized algorithm with a constant learning rate. The error bound we achieve scales polynomially with d and with the rate of $n^{-2/3}$, where n is the sample size. This rate is better than the best-known rate of $n^{-1/2}$ established for stochastic gradient Langevin dynamics (SGLD)—which employs parameter-independent Gaussian noise—under similar conditions. Our analysis offers quantitative insights into the effect of label noise.

1 INTRODUCTION

One of the central objectives in statistical learning theory is to establish generalization error bounds for learning algorithms to assess the difference between the population risk of learned parameters and their empirical risk on training data. Ever since Bousquet and Elisseeff [2002] unveiled a fundamental connection between generalization error and algorithmic stability, which gauge a learning algorithm's sensitivity to perturbations in training data, numerous studies have used the framework of uniform stability to investigate generalization properties in gradient-based methods, encompassing both convex and non-convex settings, e.g. Hardt et al. [2016]; London [2017]; Mou et al. [2018]; Li et al. [2019a]; Feldman and Vondrak [2019]; Bassily et al. [2020]; Lei and Ying [2020]; Farnia and Ozdaglar [2021]; Farghly and Rebeschini [2021]; Lei et al. [2021]; Kozachkov et al. [2022]; Zhu et al. [2023].

A line of research has focused on understanding the generalization properties resulting from the incorporation of artificial noise into stochastic gradient descent (SGD) methods, as initiated by Keskar et al. [2016] [Pensia et al., 2018; Chaudhari and Soatto, 2018; Mou et al., 2018; Negrea et al., 2019; Amir et al., 2021; Wu et al., 2022]. Initial studies examined parameter-independent isotropic Gaussian noise, as used in stochastic gradient Langevin dynamics (SGLD) [Welling and Teh, 2011; Hardt et al., 2016; Xu and Raginsky, 2017; Raginsky et al., 2017; Mou et al., 2018; Negrea et al., 2019; Zhang et al., 2019; Li et al., 2019a; Chau et al., 2021; Farghly and Rebeschini, 2021; Zhu et al., 2023]. There is a growing interest in investigating the structural capabilities induced by parameter-*dependent* noise [Goyal et al., 2017; Shallue et al., 2018; Blanc et al., 2020; Damian et al., 2021; Li et al., 2022], where true labels at each iteration are replaced with noisy labels.

However, generalization error bounds for label noise SGD have not received the same attention as their noise-independent counterparts. Most results in label noise SGD have mainly offered local, asymptotic, and phenomenological insights, without focusing on generalization. Examples include unveiling local implicit bias phenomena by investigating the stability of global minimizers [Blanc et al., 2020], providing extensions to global implicit bias for sparsity in quadratically parametrised linear regression models [HaoChen et al., 2021; Damian et al., 2021], and establishing limiting processes with infinitesimal learning rate for analysing the dynamic of label noise SGD [Li et al., 2022; Pillaud-Vivien et al., 2022].

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1.1 Contributions

In this paper, we develop generalization error bounds for label noise in SGD within non-convex settings and offer a direct comparison with SGLD to emphasize the impact of label noise on generalization. Our analysis employs uniform dissipativity and smoothness assumptions, which are commonly considered in the literature on non-convex sampling and optimization [Eberle, 2016; Raginsky et al., 2017; Xu and Raginsky, 2017; Erdogdu et al., 2018; Zhang et al., 2019; Chau et al., 2021; Farghly and Rebeschini, 2021].

Under our assumptions, we establish an exponential Wasserstein contraction property for label noise SGD exhibiting a polynomial dependence on the parameter dimension d. This contraction property drives the convergence of our generalization error bounds, which also have polynomial dependence on the dimension. Specifically, leveraging a *uniform* dissipativity assumption, we employ the 2-Wasserstein contraction theorem presented in Wang [2016] to establish the exponential contraction of the Wasserstein distance. This analysis is tailored to a particular semimetric we use for the purpose of analyzing uniform stability. To carry out this analysis, we use the continuous counterpart of the algorithm, known as the stochastic gradient *flow* (SGF). This involves the utilization of Itô calculus and linear algebra techniques to handle the parameter-dependent rectangular matrix noise term.

By leveraging algorithmic stability, we employ our contraction result to establish time-independent generalization error bounds for label noise SGD. Our bounds approach zero as the sample size n increases at the rate of $\mathcal{O}(n^{-2/3})$, achieved by scaling the learning rate as $\mathcal{O}(n^{-2/3})$. This rate is faster than the best-known rate of $\mathcal{O}(n^{-1/2})$ established for SGLD (i.e. SGD with parameter-independent Gaussian noise) under similar assumptions [Farghly and Rebeschini, 2021], as detailed in the direct comparison in Section 5. The faster decay rate can be established due to the higher dependence of label noise SGD on the learning rate η, as shown in Table 1. This dependence is readily discernible through the presence of the multiplicative factor $\sqrt{\eta}$ in the diffusion part of SGF (4), in contrast to the parameter-independent noise flow dynamics of SGLD (12), where the term $\sqrt{\eta}$ is absent. This dependence has implications for the synchronous-type coupling technique we use to estimate the discretization error. It allows for a more favorable choice of the learning rate—$\mathcal{O}(n^{-2/3})$ instead of $\mathcal{O}(n^{-1/2})$ as seen in SGLD—resulting in a faster generalization error rate.

The bounds we derive for label noise SGD exhibit a reduced dependency on the parameter dimension d, in contrast to previous bounds for SGLD [Farghly and Rebeschini, 2021]. This reduction stems from two factors, as elaborated in Section 5. Firstly, the noise term in SGF is dimension-independent, with the Wiener process in (4) being k-dimensional, where k denotes the minibatch size. In contrast, the Wiener process in the SGLD flow (12) is d-dimensional. Secondly, the contraction result we establish under uniform dissipativity has a polynomial dependence on d. In contrast, prior results used a weaker form of dissipativity and only established exponential-dependence on d.[1]

Section 2 introduces the framework of algorithmic stability and the assumptions we work with. Section 3 presents our contraction result and generalization error bounds for label noise SGD. Section 4 illustrates the proof schemes with supporting lemmas. Section 5 offers a comparison with prior work on generalization bounds for SGLD [Farghly and Rebeschini, 2021]. Section 6 is the conclusion. Proofs are in the Appendices.

2 SETUP AND PRELIMINARIES

To formalize the learning task, we consider an input-output space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^p$ represents the feature space and $\mathcal{Y} \subseteq \mathbb{R}$ represents the label space. The parameter space $\Omega \subseteq \mathbb{R}^d$ contains possible parameters of a data-generating distribution. We have a training dataset S that consists of n sample pairs $z_1, \dots, z_n \in \mathcal{Z}$, where each pair $z_i = (x_i, y_i)$ is drawn independently from a fixed probability distribution $\mathcal{P}$.

The goal is to learn a non-convex model function f belonging to the family $\mathcal{F}$, where $\theta \in \Omega$ serves as the parameter. Thus, $f(\theta, x_i)$ corresponds to the predicted output for a given input x_i and parameter θ.

Define our loss function $\ell : \Omega \times \mathcal{Z} \to \mathbb{R}$ of model $f : \Omega \times \mathcal{X} \to \mathcal{Y}$ as the squared loss:

$$\ell(\theta, z_i) := \frac{1}{2}(f(\theta, x_i) - y_i)^2.$$

We aim to find a parameter $\theta \in \Omega$ that minimizes the *population risk* $L_{\mathcal{P}}$ which is defined by:

$$L_P(\theta) := \mathbb{E}_{z \sim \mathcal{P}}[\ell(\theta, z)].$$

In settings where the data distribution $\mathcal{P}$ is unknown, calculating the population risk is often infeasible. Hence, we shift our focus to the *empirical risk* $L_S(\theta)$:

$$L_S(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, z_i). \tag{1}$$

[1] The stronger dissipativity assumption we use does not impact the influence of label noise on the relationship with the learning rate η—our choice of η does not depend on d—or the rates we establish as a function of n.

The squared loss is convex with respect to the model output, but the non-convexity of f makes the loss function non-convex with respect to the model parameters.

2.1 Generalization Error Bound via Uniform Stability

For an algorithm A trained by dataset S, we define the *generalization error* to be the difference between the empirical risk and the population risk:

$$\mathrm{gen}(A) := L_P(A(S)) - L_S(A(S)).$$

We bound the generalization error in expectation using the notion of *uniform stability*.

Definition 1 (Uniform stability [Hardt et al., 2016, Definition 2.1])**.** *A randomized algorithm A is ε-uniformly stable if*

$$\varepsilon_{stab}(A) := \sup_{S \simeq \widehat{S}} \sup_{z \in \mathcal{Z}} \mathbb{E}\left[\ell(A(S), z) - \ell(A(\widehat{S}), z)\right] \leq \varepsilon.$$

The first supremum is over pairs of datasets $S \simeq \widehat{S}$, where $S, \widehat{S} \in \mathcal{Z}^n$ differ by a single element independently drawn from the same data distribution.

The idea of bounding generalization error by uniform stability was proposed by Bousquet and Elisseeff [2002] and was further expanded by Elisseeff et al. [2005] to include random algorithms, with multiple further extensions in the literature. In this paper, we consider the notion of stability introduced in Hardt et al. [2016].

Theorem 1 (Generalization error in expectation [Hardt et al., 2016, Theorem 2.2])**.** *Let A be an ε-uniformly stable algorithm. Then,*

$$|\mathbb{E}_{A,S}[\mathrm{gen}(A)]| \leq \varepsilon.$$

2.2 Label Noise Stochastic Gradient Descent

Denote *mini-batch average* $L_S(\theta, B)$ as the average of the instance losses $\{\ell(\theta, z_i)\}$ over a uniformly sampled mini-batch $B \subset [n]$ of size $k \leq n$:

$$L_S(\theta, B) := \frac{1}{|B|} \sum_{i \in B} \ell(\theta, z_i) = \frac{1}{2k} \sum_{i \in B} (f(\theta, x_i) - y_i)^2. \quad (2)$$

We minimise the training loss in (2) with label noise SGD. Namely, during each gradient step $t > 0$, we explicitly introduce Gaussian random noise $\xi_t \sim \mathcal{N}(0, \delta I_n)$ to the label vector $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Define $\widetilde{S} = (\widetilde{z}_1, \ldots, \widetilde{z}_n)$, where $\widetilde{z}_i = (x_i, \widetilde{y}_i) = (x_i, y_i + (\xi_t)_i)$. The update rule of the algorithm started from θ_0 with initial distribution μ_0 corresponds to:

$$\begin{aligned} \theta_{t+1} &= \theta_t - \eta \nabla L_{\widetilde{S}}(\theta_t, B_{t+1}) \\ &= \theta_t - \eta \nabla L_S(\theta_t, B_{t+1}) \\ &\quad + \frac{\eta}{k} (\nabla \mathbf{f}(\theta_t, X_{B_{t+1}}))^\top (\xi_t)_{B_{t+1}}, \quad \theta_0 \sim \mu_0, \end{aligned} \quad (3)$$

where $\eta > 0$ is the learning rate, $(B_t)_{t=1}^\infty$ is an i.i.d. sequence of uniformly sampled batches of size k, $X_{B_{t+1}} \in \mathbb{R}^{k \times p}$ is a submatrix of $X := [x_1^\top, \ldots, x_n^\top]^\top$ with only rows of mini-batch B_{t+1} and $(\xi_t)_{B_{t+1}} \in \mathbb{R}^k$ is also a subvector of ξ_t corresponding to the mini-batch B_{t+1}. The matrix $\nabla \mathbf{f}(\theta, X_{B_{t+1}}) \in \mathbb{R}^{k \times d}$ consists of model gradients, where $(\nabla \mathbf{f}(\theta, X_{B_{t+1}}))_i := \nabla_\theta f(\theta, x_i) \in \mathbb{R}^d$.

In this paper, when the context is clear, we may use $\nabla \mathbf{f}$ or $\nabla \mathbf{f}(\theta)$ instead of $\nabla \mathbf{f}(\theta, X)$. Unless specified, ∇ denotes the gradient with respect to the parameter θ.

2.2.1 Label Noise Stochastic Gradient Flow

To understand the label noise SGD dynamics on the non-convex objective in (2), we use a continuous-time model known as the stochastic gradient flow (SGF). Recent studies have also explored (stochastic) diffusion processes to represent and analyze the dynamics in discrete sequential processes [Li et al., 2019b; Farghly and Rebeschini, 2021; Pillaud-Vivien et al., 2022].

The update rule (3) corresponds to the Euler-Maruyama discretization of the stochastic differential equation (SDE):

$$\begin{aligned} d\Theta_t = &- \nabla L_S(\Theta_t, B_{\lceil \frac{t}{\eta} \rceil}) dt \\ &+ \frac{\sqrt{\delta \eta}}{|B_{\lceil \frac{t}{\eta} \rceil}|} \left(\nabla \mathbf{f}(\Theta_t, X_{B_{\lceil \frac{t}{\eta} \rceil}}) \right)^\top dW_t, \end{aligned} \quad (4)$$

where W_t is a k-dimensional standard Wiener process, $\Theta_t \in \mathbb{R}^d$, $\nabla L_S(\Theta_t) \in \mathbb{R}^d$, and $\delta, \eta, n \in \mathbb{R}$.

The same SDE can be derived for any label noise with zero mean and δI covariance through the construction detailed in Appendix 7.1. Under smoothness assumptions on the loss function, these SDEs are considered to have strong solutions [Øksendal, 2003, Theorem 3.1].

As θ_t in the update rule (3) is a Markov process, we define its Markov kernel as R_θ. We will also use the notation μR_θ^s to represent the law of θ_{t+s} given that θ_t follows the distribution μ. The process Θ_t in the SDE (4) may not necessarily be a continuous-time Markov process due to its dependence on B_k. However, the discrete-time process $(\Theta_{t\eta})_{t=0}^\infty$ satisfies the Markov property, and we denote its kernel as R_Θ. We use the notation μP_t^B to denote the law of λ_t, the solution to the SDE with a deterministic batch $B \subset [n]$. Hence, μR_Θ is obtained by integrating μP_η^B over B with respect to the mini-batch distribution. We use $\widehat{\theta}_t, \widehat{\Theta}_t, \widehat{R}_\theta, \widehat{R}_\Theta$ and $\widehat{P}_t^B$ to denote the corresponding counterparts of $\theta_t, \Theta_t, R_\theta, R_\Theta$ and P_t^B when trained with a perturbed dataset $\widehat{S}$ instead of S, where S and $\widehat{S}$ differ in a single element as specified in Definition 1.

2.3 Wasserstein Distance

Algorithmic stability is often measured in terms of *p-Wasserstein distance* [Raginsky et al., 2017; Farghly and Rebeschini, 2021], defined as

$$W_p(\mu,\nu) := \left(\inf_{\pi\in\mathcal{C}(\mu,\nu)}\int \|x-y\|^p \,\pi(dx,dy)\right)^{1/p},$$

where $\|\cdot\|$ is the Euclidean norm and $\mathcal{C}(\mu,\nu)$ is the set of all couplings of μ and ν, that is, the set of all probability measure with marginals μ and ν.

The conventional approach for bounding uniform stability relies on the Lipschitz continuity of the loss function [Raginsky et al., 2017; Farghly and Rebeschini, 2021]. Without this continuity, the 2-Wasserstein distance metric becomes insufficient for bounding uniform stability. Following the approach in Farghly and Rebeschini [2021], we introduce the *semimetric*[2]

$$\rho_g(x,y) := g(\|x-y\|_2)(1+2\varepsilon+\varepsilon\|x\|_2^2+\varepsilon\|y\|_2^2), \quad (5)$$

where $\varepsilon < 1$, $g : \mathbb{R}^+\cup\{0\} \to \mathbb{R}^+\cup\{0\}$ is concave, bounded, and non-decreasing. We consider the ρ_g-*Wasserstein distance* based on the semimetric ρ_g:

$$W_{\rho_g}(\mu,\nu) := \inf_{\pi\in\mathcal{C}(\mu,\nu)}\int \rho_g(x,y)\pi(dx,dy). \quad (6)$$

2.4 Assumptions

Our analysis relies on four assumptions we now introduce. The first assumption concerns dissipativity, which is commonly (c.f. introduction) imposed to ensure that the diffusion process converges towards the origin rather than diverging when it is far from it, as noted by Erdogdu et al. [2018].

Definition 2. *A stochastic process $d\theta_t = b(\theta_t)dt + G(\theta_t)dW_t$ is α-uniformly dissipative for $p \in [1,\infty)$ and $\alpha > 0$, if $\forall\theta,\theta' \in \mathbb{R}^d$*

$$2\langle b(\theta)-b(\theta'),\theta-\theta'\rangle + \|G(\theta)-G(\theta')\|_F^2 + (p-2)\|G(\theta)-G(\theta')\|_{op}^2 \le -\alpha\|\theta-\theta'\|_2^2. \quad (7)$$

Assumption 1 (A1). *The diffusion process (4) is α-uniformly dissipative for $p = 2$.*

The remaining three assumptions specify conditions for the loss function ℓ, the model function f, and the initial parameter condition μ_0. In particular, we impose **A3** to ensure the boundedness of our noise term.

[2] A semimetric is a function defined on $\mathbb{R}^d\times\mathbb{R}^d$ that is symmetric and non-negative with $\rho_g(x,y) > 0$ for $x \neq y$ but is not necessarily satisfying the triangle inequality.

Assumption 2 (A2). *For each $z \in \mathcal{Z}$, $\ell(\cdot,z)$ is differentiable and M-smooth, where $M < \alpha/2$: $\forall\theta_1,\theta_2 \in \mathbb{R}^d$ and $\forall z \in \mathcal{Z}$,*

$$\|\nabla\ell(\theta_1,z) - \nabla\ell(\theta_2,z)\| \le M\|\theta_1-\theta_2\|.$$

Assumption 3 (A3). *For each $z \in \mathcal{Z}$, $\ell(\cdot,z)$ is ℓ_f-Lipschitz: $\forall\theta_1,\theta_2 \in \mathbb{R}^d$ and $\forall x \in \mathcal{X}$,*

$$|f(\theta_1,x) - f(\theta_2,x)| \le \ell_f\|\theta_1-\theta_2\|.$$

Assumption 4 (A4). *The initial condition μ_0 of θ_0 has finite fourth moment σ_4.*

To simplify the direct application of existing lemmas on dissipativity properties, we employ the notation: $m := \alpha/4$ and $b := \left(1+4/(\alpha^2-4M^2)\right)\eta_{\max}\delta\ell_f^2/(2k)$. Here, $\eta_{\max}$ denotes the maximum allowable learning rate as defined in Theorem 3.

3 MAIN RESULTS

3.1 The Wasserstein Contraction Property

Following the approach pursued by Farghly and Rebeschini [2021] (c.f. Lemma 4.3 in there), we aim to ensure uniform stability by establishing the contraction in the ρ_g-Wasserstein distance. Prior studies, e.g. [Eberle, 2016; Farghly and Rebeschini, 2021], have used reflection couplings to establish this contraction under conditions close to uniform dissipativity (**A1**). However, these studies considered diffusions characterized by constant (i.e. parameter-independent) noise terms. In our work, we adapt a more general 2-Wasserstein contraction result from Wang [2016] to the ρ_g-semimetric, thereby establishing exponential ρ_g-Wasserstein contraction property for label noise SGD, even with parameter-dependent noise terms.

Theorem 2 (Wasserstein contraction)**.** *Suppose **A1** and **A2** hold. Then there exists a function g such that for any $t \ge 0$ and $1 \le r < a$ we have*

$$W_{\rho_g}(\mu P_t^B, \nu P_t^B) \le C_1 e^{-\alpha t} W_{\rho_g}(\mu,\nu),$$

where

$$C_1 := \frac{1}{\varphi a\zeta_r(a)}\left(1+\varepsilon\left\{2+2\sigma_4^{1/2}+2\tilde{c}(2)^{1/2}+\frac{4b}{m} + \frac{2\delta\eta_{\max}}{km}(d+2)\ell_f^2\right\}\right),$$

with $a > 1$ and $(a-1)/a^2 < \zeta_r(a) \le 1/a$ defined in the proof in Appendix 8.2.3. Also, g is constant on $[R,\infty)$ and $\varphi r \le g(r) \le r$ for some $R,\varphi \in \mathbb{R}^+$.

Remark 1 (Bound on the contraction coefficient)**.** *We can bound C_1 as $C_1 \le e^{\alpha\eta}$ by appropriately selecting $a \equiv a(\alpha,\eta,\varphi,s)$ and $\varepsilon \equiv \varepsilon(\delta,d,m,b,\eta,\ell_f,\sigma_4,k,s)$.*

The exponential dependence of C_1 on η can be restricted by the constraint $\eta \leq \eta_{\max}$. All other constants and parameters involved exhibit only polynomial dependencies. Detailed expressions for the parameters φ, a, ε, and s are available in Appendix 8.3.4.

The parameters $\varphi, a, \zeta_r(a), \varepsilon$, and $\tilde{c}(2)$ in Theorem 2 do not depend on sample size. The parameters $\varphi, a, \zeta_r(a)$, and $\tilde{c}(2)$ are independent of dimensionality, with the exception of ε, which we have defined to exhibit a polynomial dependence on the dimension d. As a result, our final generalization error bound in Section 3.2 preserves its polynomial dependence on dimension.

Remark 2 (Dimension dependence)**.** *The primary factor enabling to achieve polynomial dependence on the dimension d is the use of the uniform dissipativity condition **A1** and the contraction result from Wang [2016], rather than the presence of label noise itself.*

3.2 Generalization Error Bounds

We now derive an upper bound on the expected generalization error $|\mathbb{E}_{A,S}[\mathrm{gen}(A)]|$ for a randomly selected dataset S and a randomized algorithm A which belongs to the class of iterative algorithms described in Section 2.2. The explicit expressions for all parameters can be found in the proof provided in Appendix 8.3.3.

Theorem 3 (Generalisation error bounds)**.** *Suppose **A1**, **A2**, **A3**, and **A4** hold and $\eta \leq \eta_{\max} := \min\{\frac{1}{m}, \frac{m}{2M^2}\}$. Then, for any $t \in \mathbb{N}$, the continuous-time algorithm attains the generalization error bound*

$$|\mathbb{E}\mathrm{gen}(\Theta_{\eta t})| \leq C_2 \min\left\{\eta t, \frac{n(\eta + 2/\alpha)}{(n-k)}\right\} \cdot \frac{1}{n}\left(\frac{\eta}{k^{1/2}} + \eta^{1/2} + k^{1/2} + \frac{k}{\eta^{1/2}}\right). \quad (8)$$

The discrete-time algorithm attains the bound

$$|\mathbb{E}\mathrm{gen}(\theta_t)| \leq C_3 \min\left\{\eta t, \frac{n(\eta + 2/\alpha)}{(n-k)}\right\} \cdot \left[\frac{1}{n}\left\{\frac{\eta}{k^{1/2}} + \eta^{1/2} + k^{1/2} + \frac{k}{\eta^{1/2}}\right\} + \eta + \frac{\eta}{k^{1/2}}\right]. \quad (9)$$

The positive parameters $C_2 \equiv C_2(\delta, d, m, b, M, \ell_f, \sigma_4, \varphi, R, \varepsilon)$ and $C_3 \equiv C_3(\delta, d, m, b, M, \ell_f, \sigma_4, \varphi, R, \varepsilon)$ are given in (18) *and* (21) *in Appendix 8.3.3.*

Remark 3 (Sample size dependence)**.** *Choosing $\eta = \mathcal{O}(n^{-2/3})$ achieves the fastest decaying generalization error bound of $\mathcal{O}(n^{-2/3})$ as derived in Appendix 7.6.*

Remark 4 (Dimension dependence)**.** *Our parameters C_2 and C_3 exhibit polynomial dependencies on parameter dimension d, as well as on $\delta, m, b, M, \ell_f, \sigma_4, \varphi, R, \varepsilon$. The generalization error bounds for both the continuous and the discrete-time algorithm increase at a rate of $d^{5/2}$ as detailed in Table 1. Our bound remains independent of the feature dimension p since the feature vectors only affect our algorithm through the model function f, which has an output dimension of 1.*

Remark 5 (Time independence)**.** *Following prior works, we adopt the term "time-independent" bounds to denote bounds that remain constant as a function of time t for t large enough. See Lemma 7 for a reference.*

4 Proof Schemes

4.1 Proof for Wasserstein Contraction

The ρ_g-Wasserstein contraction analysis in Theorem 2 builds upon the 2-Wasserstein contraction result in Wang [2016] under uniform dissipativity. We can draw from Theorem 2.5 in Wang [2016] that if **A1** and **A3** are satisfied, then for any $t \geq 0$, the following contraction property holds:

$$W_2(\mu P_t^B, \nu P_t^B) \leq e^{-\alpha t/2} W_2(\mu, \nu). \quad (10)$$

Transition from 2-Wasserstein contraction to ρ_g-Wasserstein contraction is achieved by leveraging the properties of the semimetric ρ_g as elaborated in Farghly and Rebeschini [2021] and the Reverse Jensen's Inequality in Wunder et al. [2021].

Refer to Lemma D.3 in Farghly and Rebeschini [2021] for the following inequality, which holds for any two probability measures μ and ν on $\mathbb{R}^d$:

$$W_{\rho_g}(\mu, \nu) \leq W_2(\mu, \nu)(1 + 2\varepsilon + \varepsilon\mu(\|\cdot\|^4)^{\frac{1}{2}} + \varepsilon\nu(\|\cdot\|^4)^{\frac{1}{2}}). \quad (11)$$

To utilize inequality (11), we derive the moment bound (Lemma 1) and moment estimate bound (Lemma 2) tailored for label noise SGD. These derivations require adjustments using Itô calculus tools to accommodate the parameter-dependent nature of label noise SGD and the non-square matrix form of the noise term.

Lemma 1 (Moment bound)**.** *Suppose **A1** and **A3** hold and μ is a probability measure on $\mathbb{R}^d$. Then, for any $B \subset [n]$, we have*

$$\mu P_t^B(\|\cdot\|^p) \leq \mu(\|\cdot\|^p) + \left[\frac{2b}{m} + \frac{\delta\eta}{km}(p + d - 2)\ell_f^2\right]^{p/2}.$$

Combining the results in (10) and (11) with the moment bound in Lemma 1, we obtain the following inequality, which holds under the assumptions **A1** and **A3**:

$$W_{\rho_g}(\mu P_\eta^B, \nu \widehat{P}_\eta^B) \leq e^{-\alpha t/2} W_2(\mu, \nu)\left(1 + \varepsilon\left\{2 + \mu(\|\cdot\|^2)^{\frac{1}{2}} + \nu(\|\cdot\|^2)^{\frac{1}{2}} + \frac{4b}{m} + \frac{2\delta\eta}{km}(d+2)\ell_f^2\right\}\right).$$

Our analysis focuses on the contraction of ρ_g-Wasserstein distance between the distributions $\mu = \mu_0 R_\Theta^t$ and $\nu = \mu_0 R_\Theta^t$, which represent the laws of our processes when initiated from the same distribution μ_0 and trained with datasets that differ in a single element. We establish the following moment estimate bound by further utilizing the smoothness (**A2**) and finite fourth moment (**A4**) assumptions.

Lemma 2 (Moment estimate bound)**.** *Suppose **A1**, **A2**, **A3**, and **A4** hold. Then*

$$\mu R_\theta^t(\|\cdot\|^{2p}) \le \mu(\|\cdot\|^{2p}) + \tilde{c}(p),$$

where $\eta < \eta_{\max} := \min\{\frac{1}{m}, \frac{m}{2M^2}\}$ and

$$\begin{aligned}\tilde{c}(p) = \eta_{\max}\Big\{ &(3b)^p\,(\eta_{\max}+2/m)^{p-1} \\ &+ p(2p-1)\delta\ell_f^2(\eta_{\max}+2/m)^{p-2}\,(3b)^{p-1}\,\eta_{\max}^2 \\ &+ \{p(2p-1)\}^{p+1}\delta^p\ell_f^{2p}\eta_{\max}^{2p-1}\Big\}.\end{aligned}$$

As a consequence of Lemma 2, we have:

$$\begin{aligned}W_{\rho_g}(\mu P_\eta^B, \nu\widehat{P}_\eta^B) \le e^{-\alpha t/2} W_2(\mu,\nu)\Bigg(1+\varepsilon\Bigg\{2+2\sigma_4^{1/2} \\ + 2\tilde{c}(2)^{1/2} + \frac{4b}{m} + \frac{2\delta\eta}{km}(d+2)\ell_f^2\Bigg\}\Bigg).\end{aligned}$$

Lastly, we establish the following lemma, based on the Reverse Jensen's Inequality in Wunder et al. [2021].

Lemma 3. *There exists a function g constant on $[R,\infty)$ with $\varphi r \le g(r) \le r$ for some $R, \varphi \in \mathbb{R}^+$ such that, for any two probability measures μ and ν on $\mathbb{R}^d$, we have*

$$W_2(\mu,\nu) \le \frac{1}{\varphi a \zeta_b(a)} W_{\rho_g}(\mu,\nu).$$

Lemma 3 guarantees the existence of a function g such that the semimetric ρ_g exhibits the exponential contraction property outlined in Theorem 2.

4.2 Proof for Generalization Error Bounds

The proof of Theorem 3 follows the stability framework outlined in Section 2.1, adhering to the dissipativity and smoothness conditions we consider. A result from Farghly and Rebeschini [2021] shows that a uniform stability bound can be obtained by controlling the ρ_g-Wasserstein distance between the laws of algorithms $A(S)$ and $A(\widehat{S})$, where $S \simeq \widehat{S}$.

Lemma 4 ([Farghly and Rebeschini, 2021, Lemma 4.3])**.** *Suppose **A1** and **A2** hold and let A be a random algorithm. Then*

$$\varepsilon_{stab}(A) \le \frac{M(b/m+1)}{\varphi\varepsilon(R\vee 1)} \sup_{S\simeq\widehat{S}} W_{\rho_g}\Big(law(A(S)), law(A(\widehat{S}))\Big).$$

Remark 6. *Lemma 4.3 in Farghly and Rebeschini [2021] applies under a weaker assumption of dissipativity and remains valid under our stronger assumption (**A1**). This connection is elaborated in Section 5.1.*

Hence, it is sufficient to control the quantities $W_{\rho_g}(\mu_0 R_\Theta^t, \mu_0 \widehat{R}_\Theta^t)$ and $W_{\rho_g}(\mu_0 R_\theta^t, \mu_0 \widehat{R}_\theta^t)$ to prove Theorem 3. Recall from Section 2.2 that μR_Θ is obtained by integrating μP_η^B over B with respect to the mini-batch distribution. To begin, we estimate the divergence $W_{\rho_g}(\mu P_\eta^B, \nu\widehat{P}_\eta^B)$ using the divergence bound (Lemma 5), the moment bound (Lemma 1) and the moment estimate bound (Lemma 2) with $p=4$.

Lemma 5 (Divergence bound)**.** *Suppose **A1**, **A2**, and **A3** hold. Then*

$$\mathbb{E}\,\|\theta_t - \theta_0\|^2 \le 4M^2\left(\mathbb{E}\,\|\theta_0\|^2 + \frac{3b}{m} + \frac{\delta\eta d}{km}\ell_f^2\right)t + \frac{2\delta\eta}{k}\ell_f^2 t.$$

This lemma computes the extent to which the process θ_t deviates from the initial condition θ_0.

Without loss of generality, assume that the datasets S and $\widehat{S}$ differ only at i^{th} element. Considering that $\mathbb{P}(i \in B) = k/n$, the convexity of the ρ_g-Wasserstein distance (Lemma 12 in Appendix 7.4) gives the following inequality:

$$\begin{aligned}W_{\rho_g}(\mu R_\Theta, \nu\widehat{R}_\Theta) \le \frac{k}{n}\sup_{B:n\in B} W_{\rho_g}(\mu P_\eta^B, \nu\widehat{P}_\eta^B) \\ + \left(1-\frac{k}{n}\right)\sup_{B:n\notin B} W_{\rho_g}(\mu P_\eta^B, \nu\widehat{P}_\eta^B).\end{aligned}$$

If $i \notin B$, then $\widehat{P}^B = P^B$ so the processes $\Theta_{t\eta}$ and $\widehat{\Theta}_{t\eta}$ contract in ρ_g-Wasserstein distance by Theorem 2. If $i \in B$, the divergence $W_{\rho_g}(\mu P_\eta^B, \nu\widehat{P}_\eta^B)$ obtained above provides uniform bounds on the extent to which $\Theta_{t\eta}$ and $\widehat{\Theta}_{t\eta}$ can deviate from each other.

Using induction and auxiliary inequalities, we derive a bound for $\varepsilon_{stab}(\Theta_{\eta t})$ in terms of Lemma 4. By Theorem 1, this bound serves as the generalization error bound for our continuous-time algorithm, as in (8).

So far, we analysed the continuous-time dynamics of our algorithm. The discrete-time process (3) corresponds to the Euler-Maruyama discretization of (4). We derive discretization error bounds using synchronous-type couplings between θ_η and $\Theta_{t\eta}$, with both processes sharing the same Brownian motion.

Lemma 6 (Discretization error bound)**.** *Suppose **A1**, **A2**, and **A3** hold. Then, for any probability measure μ on $\mathbb{R}^d$, we have*

$$\begin{aligned}W_2(\mu R_\theta, \mu R_\Theta)^2 \le 8\eta^4 \exp\left(4\eta^2 M^2\right) \\ \cdot\left[\frac{2}{3}M^4\left(\mu\,\|\cdot\|^2 + \frac{b}{m}\right) + \left(M^2+2\right)\frac{\delta}{2k}\ell_f^2\right].\end{aligned}$$

To extend the generalization error bound established for the continuous-time algorithm (8) to its discrete-time counterpart (9), we add the one-step discretization error to the continuous-time error bound using the weak triangle inequality (Lemma 15 in Appendix 7.4).

5 COMPARISON WITH SGLD

Label noise SGD is a parameter-dependent noisy algorithm often compared with parameter-independent noisy algorithms like SGLD [HaoChen et al., 2021]. Farghly and Rebeschini [2021] present a discrete-time generalization error bound for SGLD in a dissipative and smooth setting, which decays to zero at a rate of $\mathcal{O}(n^{-1/2})$ with an appropriate learning rate scaling as $\mathcal{O}(n^{-1/2})$. In comparison, our result exhibits a faster rate of decay, as discussed in the introduction.

Lemma 7 ([Farghly and Rebeschini, 2021, Theorem 4.1])**.** *If $\eta \in (0,1)$ then for any $t \in \mathbb{N}$, the continuous-time algorithm attains the generalization bound*

$$|\mathbb{E}\mathrm{gen}(\Theta_{\eta t})| < C_5 \min\left\{\eta t, \frac{(C_4+1)n}{n-k}\right\} \frac{k}{n\eta^{1/2}}$$

Furthermore, if $\eta \le 1/2m$, then the discrete-time algorithm attains the generalization bound

$$|\mathbb{E}\mathrm{gen}(\theta_t)| < C_6 \min\left\{\eta t, \frac{(C_4+1)n}{n-k}\right\} \left(\frac{k}{n\eta^{1/2}} + \eta^{1/2}\right).$$

The parameters C_4, C_5, C_6 depend on M, m, b, d, β. Here, $\beta^{-1} > 0$ represents the noise level.

5.1 Comparison of Settings

The proof of the generalization error bound for SGLD in Farghly and Rebeschini [2021] also relies on uniform stability and is built upon largely the same assumptions we use, except for one significant difference that arises in the analytical framework regarding the concept of dissipativity. Farghly and Rebeschini [2021] consider the following assumption in place of the uniform dissipativity assumption (**A1**) we use:

Assumption 1′ (A1′). *The loss function $\ell(\cdot, z)$ is (m,b)-dissipative: there exists $m > 0$ and $b \ge 0$ such that, for all $\theta \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

$$\langle \theta, \nabla \ell(\theta, z)\rangle \ge m\|\theta\|^2 - b \quad \forall \theta \in \mathbb{R}^d.$$

Uniform dissipativity is the key factor that allows our results to circumvent the exponential dependence on the parameter dimension d as established in SGLD's bounds in Farghly and Rebeschini [2021], leading to polynomial dependence. However, it is important to stress that the contraction result is unrelated to the dependence of our final generalization error bound the learning rate η and sample size n. Consequently, the faster decay rate as a function of n highlighted in our generalization error bounds is attributable to the advantages provided by label noise rather than the imposition of uniform dissipativity. This observation is further supported by the following lemma, where we establish a relationship between the assumptions of uniform dissipativity (**A1**) and the dissipativity (**A1′**).

Lemma 8. *Under **A2** and **A3**, the uniform dissipativity assumption **A1** implies the dissipativity assumption **A1′** with $m = \alpha/4$ and $b = \left(\frac{4}{\alpha^2 - 4M^2} + 1\right)\frac{\eta_{\max}}{2k}\delta\ell_f^2$. The converse holds if $m^3 < M^2 b$ and $\|\theta\| < B$ for all θ, where B is within the interval*

$$\frac{1}{2M}\left(m - M\sqrt{\frac{b}{m}} \pm \sqrt{\left(M\sqrt{\frac{b}{m}} - m\right)^2 - 4M\left(b + \frac{\eta\delta\ell_f^2}{k}\right)}\right).$$

This lemma illustrates that by bounding the parameter space and imposing constraints on dissipativity and smoothness constants, we can treat the analytical framework of Farghly and Rebeschini [2021] and our own as equivalent. This enables a direct comparison between the two algorithms, label noise SGD and SGLD. The proof of Lemma 8 is in Appendix 8.1.

Regarding the absence of the Lipschitzness assumption on the model (**A3**) in Farghly and Rebeschini [2021], it is worth noting the strong connection between **A2** and **A3** in label noise SGD with squared loss L_S. Near the global minimizer θ^*, it is observed that $\left\|\frac{1}{k}\nabla\mathbf{f}(\theta^*)^\top \nabla\mathbf{f}(\theta^*)\right\|_2 \approx \left\|\nabla^2 L_S(\theta^*)\right\|_2$, as discussed in Damian et al. [2021] and Li et al. [2022]. Thus, under **A3**, we have the following inequalities for our loss L_S:

$$k\left\|\nabla^2 L_S(\theta^*)\right\|_2 \approx \left\|\nabla\mathbf{f}^\top\nabla\mathbf{f}\right\|_2 = \sigma_{\max}(\nabla\mathbf{f}^\top\nabla\mathbf{f})$$
$$\le \sum_i^k \lambda_i(\nabla\mathbf{f}^\top\nabla\mathbf{f}) = \mathrm{Tr}(\nabla\mathbf{f}^\top\nabla\mathbf{f}) = \sum_{i=1}^k \|\nabla f_i\|^2 < k\ell_f^2.$$

This confirms the ℓ_f^2-smoothness of L_S (**A2** with $M = \ell_f^2$) near the global minimum.

5.2 Label Noise and Faster Decay Rate

We pinpoint the reasons for the faster rate of decay (as a function of the sample size n) in the generalization error bound of label noise SGD compared to SGLD by closely examining the differences in the proof components, as outlined in Table 1.

The noise terms in SGLD and label noise SGD exhibit different dependencies on the learning rate η, dimension d, and batch size k. In the update rule of SGLD, the noise term exhibits a square root dependence on the learning rate η:

$$\theta_{t+1} = \theta_t + \eta\nabla L_S(\theta_t, B_{t+1}) + \sqrt{2\beta^{-1}\eta}\xi_{t+1}, \quad \theta_0 \sim \mu_0.$$

Table 1: Bounds with respect to η, d, and n.

Bound Term	SGLD	Label Noise SGD
Divergence	$\mathcal{O}(d+1)$	$\mathcal{O}(\eta d+\eta)$ (5)
Moment	$\mathcal{O}(d^2+1)$	$\mathcal{O}(\eta^2 d^2+1)$(1)
Moment estimate	$\mathcal{O}(d^2+d+1)$	$\mathcal{O}(1)$ (2)
Discretization error	$\mathcal{O}(d\eta^3 e^{\eta^2})$	$\mathcal{O}(\eta^4 e^{\eta^2})$ (6)

$\lvert\mathbb{E}\text{gen}(\theta_t)\rvert$ for SGLD
$\mathcal{O}\left(e^{(d+\sqrt{d})}\left(\frac{d^{7/2}}{n\eta^{1/2}}+d^{3/2}\eta^{1/2}\right)\right)$
$\lvert\mathbb{E}\text{gen}(\theta_t)\rvert$ for Label Noise SGD
$\mathcal{O}\left(\frac{d^{3/2}}{n}\left[d\eta+(d\eta)^{1/2}+1+(d\eta)^{-1/2}\right]+d\eta\right)$ (3)

Consequently, the resulting noise term in the associated stochastic process becomes independent of η by the derivation detailed in Appendix 7.1. The stochastic process associated to SGLD is expressed as:

$$d\Theta_t = -\nabla L_S(\Theta_t, B_{\lceil t/\eta \rceil})dt + \sqrt{2\beta^{-1}}d\widetilde{W}_t, \Theta_0 \sim \mu_0, (12)$$

where $\widetilde{W}_t$ is a d-dimensional standard Wiener process. The noise term in this stochastic process is independent of both η and k.

In contrast, the update rule of label noise SGD (3) has a linear dependence of the noise term on η. This linear relationship arises because label noise impacts the loss function, and its gradient is directly scaled by the learning rate η in the update rule. Thus, as shown in Appendix 7.1, the noise term in the stochastic process of label noise SGD (4) is linearly dependent on η/k.

The faster decay rate of the label noise SGD bound compared to the SGLD bound is primarily attributed to its discretization error bound. Table 1 shows that the SGLD discretization error bound scales as $\mathcal{O}(\eta^3)$, whereas that of label noise SGD scales as $\mathcal{O}(\eta^4)$, a consequence of the synchronous-type coupling method detailed in Appendix 8.3.2. The noise term's dependency on model parameters in label noise SGD unavoidably introduces η-dependent noise in our coupling method, unlike the synchronous-type coupling method used for parameter-independent noise. This, coupled with the divergence bound's dependence on η, strengthens the dependence of the discretization error bound on η and leads to a faster decay rate of our discrete-time generalization bound through an appropriate choice of η.

5.3 Dimensionality Dependencies

A distinguishing trait of the generalization error bound presented in Farghly and Rebeschini [2021] is its exponential dependence on the parameter dimension d. This dependence is a consequence of the contraction result employed by the authors under the dissipativity assumption they considered. In contrast, our approach, inspired by the 2-Wasserstein contraction result from Wang [2016] under uniform dissipativity, allows us to circumvent this dependency, leading our generalization error bound displaying polynomial scaling with respect to the dimension d.

Furthermore, the difference in the dimension of the Wiener process, which is k-dimensional in label noise SGD (4) and d-dimensional in SGLD (12), leads to reduced dependence on the parameter dimension within our proof components and, consequently, our generalization bounds. This indicates the advantages offered by label noise. Upon examining the parameters in the proof of Theorem 3, it is noteworthy that the divergence bound and moment bound of label noise SGD depends on η, which is different from that of SGLD. This leads to similar scaling of η and d in each term of the generalization error bound for label noise SGD, indicating that controlling η can alleviate the increase in bounds due to high dimensionality.

6 CONCLUSION

The proof technique we employ to establish the contraction property for label noise SGD with polynomial dependence on the dimension d hinges on the assumption of uniform dissipativity. This assumption enables us to avoid the need for reflection coupling, which was utilized in prior research involving parameter-independent noise SGLD [Farghly and Rebeschini, 2021]. This, in turn, allows us to direct our attention toward understanding the impact of label noise on the selection of learning rate scaling, thereby achieving improved generalization error bounds as a function of the sample size n.

We defer the task of establishing results for label noise SGD under a less restrictive form of dissipativity to future research. This pursuit may involve employing Kendall-Cranston couplings [Kendall, 1986; Cranston, 1991] for parameter-dependent noise terms, i.e. non-constant diffusion coefficients.

References

I. Amir, T. Koren, and R. Livni. SGD generalizes better than GD (and regularization doesn't help). In *Conference on Learning Theory*, pages 63–92, 2021.

R. Bassily, V. Feldman, C. Guzmán, and K. Tal-

war. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems*, volume 33, pages 4381–4391, 2020.

G. Blanc, N. Gupta, G. Valiant, and P. Valiant. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *Conference on learning theory*, pages 483–513, 2020.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2: 499–526, 06 2002.

N. H. Chau, E. Moulines, M. Rasonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams: The fully nonconvex case. *SIAM Journal on Mathematics of Data Science*, 3(3):959–986, 2021.

P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop*, pages 1–10, 2018.

M. Cranston. Gradient estimates on manifolds using coupling. *Journal of functional analysis*, 99(1):110–124, 1991.

A. Damian, T. Ma, and J. D. Lee. Label noise SGD provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

A. Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166:851–886, 2016.

A. Elisseeff, T. Evgeniou, M. Pontil, and L. P. Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

M. A. Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. *Advances in Neural Information Processing Systems*, 31, 2018.

T. Farghly and P. Rebeschini. Time-independent generalization bounds for SGLD in non-convex settings. *Advances in Neural Information Processing Systems*, 34:19836–19846, 2021.

F. Farnia and A. Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pages 3174–3185, 2021.

V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279, 2019.

P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch SGD: Training Imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

T. H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4): 292–296, 1919.

J. Z. HaoChen, C. Wei, J. D. Lee, and T. Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357, 2021.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234, 2016.

K. Itô. Stochastic integral. *Proceedings of the Imperial Academy*, 20(8):519–524, 1944.

W. S. Kendall. Nonnegative Ricci curvature and the Brownian coupling property. *Stochastics: An International Journal of Probability and Stochastic Processes*, 19(1-2):111–129, 1986.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

L. Kozachkov, P. M. Wensing, and J. Slotine. Generalization in supervised learning through Riemannian contraction. *arXiv preprint arXiv:2201.06656*, 2022.

A. Krogh and J. Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

Y. Lei and Y. Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 5809–5819, 2020.

Y. Lei, M. Liu, and Y. Ying. Generalization guarantee of SGD for pairwise learning. *Advances in Neural Information Processing Systems*, 34:21216–21228, 2021.

J. Li, X. Luo, and M. Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. *arXiv preprint arXiv:1902.00621*, 2019a.

Q. Li, C. Tai, and W. E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, (1):1474–1520, 2019b.

Z. Li, T. Wang, and S. Arora. What happens after SGD reaches zero loss? –a mathematical framework, 2022.

B. London. Generalization bounds for randomized learning with application to stochastic gradient descent. In *Knowledge Discovery and Data Mining 2017*, 2017.

X. Mao. Brownian motions and stochastic integrals. pages 1–46. Woodhead Publishing, second edition edition, 2011.

L. Mirsky. A trace inequality of John von Neumann. *Monatshefte für Mathematik*, 79:303–306, 1975.

W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638, 2018.

J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. *Advances in Neural Information Processing Systems*, 32, 2019.

B. Øksendal. *Stochastic differential equations*. Springer, 2003.

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

L. Pillaud-Vivien, J. Reygner, and N. Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In *Conference on Learning Theory*, pages 2127–2159, 2022.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Nonconvex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703, 2017.

C. J. Shallue, J. Lee, J. Antognini, J. Sohl-Dickstein, R. Frostig, and G. E. Dahl. Measuring the effects of data parallelism on neural network training. *arXiv preprint arXiv:1811.03600*, 2018.

F. Y. Wang. Exponential contraction in Wasserstein distances for diffusion semigroups with negative curvature. *arXiv preprint arXiv:1603.05749*, 2016.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

L. Wu, M. Wang, and W. Su. The alignment property of sgd noise and how it helps select flat minima: A stability analysis. In *Advances in Neural Information Processing Systems*, volume 35, pages 4680–4693. Curran Associates, Inc., 2022.

G. Wunder, B. Groß, R. Fritschek, and R. F. Schaefer. A reverse Jensen inequality result with application to mutual information estimation. In *2021 IEEE Information Theory Workshop*, pages 1–6, 2021.

A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

W. H. Young. On classes of summable functions and their Fourier series. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 87(594):225–229, 1912.

Y. Zhang, Ö. D. Akyildiz, T. Damoulas, and S. Sabanis. Nonasymptotic estimates for stochastic gradient Langevin dynamics under local conditions in nonconvex optimization. *arXiv preprint arXiv:1910.02008*, 2019.

L. Zhu, M. Gurbuzbalaban, A. Raj, and U. Simsekli. Uniform-in-time Wasserstein stability bounds for (noisy) stochastic gradient descent, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   : Provided in Section 2.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   : Provided in Section 3.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/**Not Applicable**]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   : Provided in Section 2.
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   : Provided in Section 4 and the supplementary materials.
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]
   : Provided in Section 2.

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/**Not Applicable**]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/**Not Applicable**]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/**Not Applicable**]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/**Not Applicable**]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]

(b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]

(c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]

(d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

Supplementary Materials

7 Technical Backgrounds

7.1 Continuous-time stochastic dynamics modeling

A standard formulation of a SDE for the process $(\Theta_t)_{t=0}^{\infty}$ is given by:

$$d\Theta_t = b(\Theta_t, t)dt + G(\Theta_t, t)dW_t,$$

where W denotes the Wiener process (standard Brownian motion). Here, the term $b(\Theta_t, t)$ is the drift term, which determines the trend or direction of the process, and $G(\Theta_t, t)$ is the noise term, which determines the randomness of the process.

Consider an update rule

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t) + \sqrt{\eta} V_t, \tag{13}$$

where $L : \mathbb{R}^d \to \mathbb{R}$ is an arbitrary function and V_t is a d-dimensional random vector. Let $\Sigma := \frac{1}{\eta}\mathrm{Cov}\left[V_t | \theta_t = \theta\right]$. Then, the update (13) is the Euler-Maruyama discretization of the time-homogeneous SDE

$$d\Theta_t = -\nabla L(\Theta_t)dt + (\eta\Sigma)^{1/2} dW_t,$$

where W_t is a Wiener process.

7.2 Itô calculus

In many papers that analyze diffusion processes of Gaussian noise algorithms, such as SGLD, the noise terms are often in a simple constant scalar form, making calculations relatively straightforward. However, analyzing the diffusion process (4) of label noise SGD with squared loss requires more involved calculations due to a more general noise term: a $d \times k$ matrix that depends on the current value of the process.

Hence, we record below some key lemmas of Itô calculus, which is an extension of calculus to stochastic processes. These lemmas are essential in proving Theorem 3, particularly when extending the proof of Farghly and Rebeschini [2021] to the label noise SGD algorithm.

Lemma 9 (Itô's lemma [Itô, 1944])**.** *Let Θ_t be a $\mathbb{R}^d$-valued Itô process satisfying the SDE*

$$\mathrm{d}\Theta_t = b_t \mathrm{d}t + G_t \mathrm{d}W_t,$$

where $\mu_t \equiv \mu(\Theta_t, t)$ and $G_t \equiv G(\Theta_t, t)$ are adapted processes to the same filtration as the n-dimensional Wiener's process W_t. Here, b_t is $\mathbb{R}^d$-valued and G_t is $\mathbb{R}^{d\times n}$-valued.

Suppose that $\phi \in \mathcal{C}^2$. Then, with probability 1, for all $t \geq 0$,

$$\mathrm{d}\phi(\Theta_t) = \left\{ \frac{\partial \phi}{\partial t} + (\nabla \phi)^\top b_t + \frac{1}{2} Tr[G_t^\top (\nabla^2 \phi) G_t] \right\} \mathrm{d}t + (\nabla \phi)^\top G_t \mathrm{d}W_t.$$

Lemma 10 (Itô isometry [Øksendal, 2003])**.** *If $g(t, w)$ is bounded and elementary then*

$$\mathbb{E}\left[\left(\int_s^t g(t, w) dW_t \right)^2 \right] = \mathbb{E}\left[\int_s^t g(t, w)^2 dt \right].$$

7.3 Relevant inequalities

Throughout our proofs, we employ valuable inequalities, which are elaborated upon as follows.

Lemma 11 (Grönwall's lemma [Gronwall, 1919]). *Assume $\phi : [0,T] \to \mathbb{R}$ is a bounded non-negative measurable function, $C : [0,T] \to \mathbb{R}$ is a non-negative integrable function and $B \geq 0$ is a constant with the property that*

$$\phi(t) \leq B + \int_0^t C(\tau)\phi(\tau)d\tau \quad \forall t \in [0,T].$$

Then,

$$\phi(t) \leq B \exp\left(\int_0^t C(\tau)d\tau\right) \quad \forall t \in [0,T].$$

Theorem 4 (Young's inequality for products [Young, 1912]). *If $a \geq 0$ and $b \geq 0$ are non-negative real numbers and if $p > 1$ and $q > 1$ are real numbers such that $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Equality holds if and only if $a^p = b^q$.

7.4 Properties of the semimetric

To make this paper self-contained, we will include the lemmas from Farghly and Rebeschini [2021] regarding the semimetric (5) and the Wasserstein distance (6) for our future reference.

The convexity of the Wasserstein distance is a crucial property that plays a central role in our results:

Lemma 12 (Convexity of the Wasserstein distance [Farghly and Rebeschini, 2021, Lemma 2.3]). *Suppose that ρ_g is a semimetric and $\mu_1, \mu_2, \nu_1, \nu_2$ are probability measures. Then, for any $r \in [0,1]$,*

$$W_{\rho_g}(\mu,\nu) \leq rW_{\rho_g}(\mu_1,\nu_1) + (1-r)W_{\rho_g}(\mu_2,\nu_2),$$

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

We require the following lemma for computing the divergence bound:

Lemma 13 ([Farghly and Rebeschini, 2021, Lemma D.3]). *Suppose X, Y, Δ_x and Δ_y are random variables on $\mathbb{R}^d$, then*

$$\mathbb{E}\rho_g(X+\Delta_x, Y+\Delta_y) \leq \mathbb{E}\rho_g(X,Y) + \sigma_\Delta^{1/2}(1 + 2\varepsilon + 6\varepsilon\sigma^{1/2}),$$

where we define $\sigma_\Delta := \mathbb{E}\left\|\Delta_x\right\|^2 \vee \mathbb{E}\left\|\Delta_y\right\|^2$ and $\sigma := \mathbb{E}\left\|X\right\|^4 \vee \mathbb{E}\left\|Y\right\|^4 \vee \mathbb{E}\left\|X+\Delta_x\right\|^4 \vee \mathbb{E}\left\|Y+\Delta_y\right\|^4$.

To establish the contraction result and simplify the calculation of discretization error in Lemma 6, we will simplify the computation by using the 2-Wasserstein distance. In order to do so, we need the following lemma:

Lemma 14 (Comparison with the 2-Wasserstein distance [Farghly and Rebeschini, 2021, Lemma D.2]). *For any two probability measures μ and ν on $\mathbb{R}^d$,*

$$W_{\rho_g}(\mu,\nu) \leq W_2(\mu,\nu)(1 + 2\varepsilon + \varepsilon\mu(\|\cdot\|^4)^{1/2} + \varepsilon\nu(\|\cdot\|^4)^{1/2}).$$

Finally, we apply the weak triangle inequality in the following lemma to extend the results from continuous-time dynamics to the discrete-time case:

Lemma 15 (Weak triangle inequality [Farghly and Rebeschini, 2021, Lemma D.1]). *For any $x, y, z \in \mathbb{R}^d$ it holds that,*

$$\rho_g(x,y) \leq \rho_g(x,z) + 2\left(1 + \frac{R}{\varphi}(\varepsilon R \vee 1)\right)\rho_g(z,y).$$

7.5 Regularity assumptions

The assumptions **A2-4** are fairly standard in the literature on non-convex optimization. The uniform dissipativity assumption (**A1**) merit some discussion.

Some may express concerns that the constraint of the dissipativity condition addressed in Lemma A.2 of Farghly and Rebeschini [2021] may confine the process within the absorbing set, potentially simplifying the dynamics of the stochastic gradient flow that we are analyzing. However, it is worth noting that the dissipative assumptions can be enforced through techniques such as weight decay regularization [Krogh and Hertz, 1991; Raginsky et al., 2017]. This regularization method allows for control over the dynamics of the algorithm and can help ensure that the dissipativity condition is satisfied, while still preserving the overall stability and convergence properties of the algorithm. Thus, the seemingly restrictive nature of the dissipativity condition can be effectively managed through appropriate regularization techniques, adding flexibility to the analysis and applicability of the results.

7.6 Derivation of Decay Rates

In Section 3.2, we determine the fastest decay rate for our generalization error bound concerning parameters like n, d, and η. This computation is done by direct optimization as below.

In Remark 3, we aim to control the learning rate η to achieve the fastest decay rate of the bound which scales as $\mathcal{O}(n^{-1}\{\eta + \eta^{1/2} + 1 + \eta^{-1/2} + \eta\})$ in terms of η and n. Therefore, we introduce the variable power of η as $\eta = \mathcal{O}(n^q)$, and we optimize for the best value of q that results in the fastest decay rate of our bound as n increases.

8 Missing Proofs

8.1 Uniform Dissipativity and Dissipativity

In the proof, we leverage the implication of **A1′** from **A1** in Lemma 8, where we set $m := \alpha/4$ and $b := \left(1 + 4/(\alpha^2 - 4M^2)\right)\eta_{\max}\delta\ell_f^2/(2k)$. This simplifies the direct application of existing lemmas pertaining to dissipativity properties.

Proof of Lemma 8. Suppose that the diffusion process 4 is α-uniformly dissipative: $\forall\ \theta, \theta'$,

$$2\langle -\nabla L_S(\theta) + \nabla L_S(\theta'), \theta - \theta'\rangle + \frac{\eta\delta}{k^2}\left\|\nabla \mathbf{f}(\theta) - \nabla \mathbf{f}(\theta')\right\|_F^2 \leq -\alpha\left\|\theta - \theta'\right\|^2.$$

For $\theta' = 0$,

$$\begin{aligned} & 2\langle -\nabla L_S(\theta) + \nabla L_S(0), \theta\rangle + \frac{\eta\delta}{k^2}\left\|\nabla \mathbf{f}(\theta) - \nabla \mathbf{f}(0)\right\|_F^2 \leq -\alpha\left\|\theta\right\|^2 \\ \Rightarrow\ & 2\langle -\nabla L_S(\theta), \theta\rangle + \frac{\eta\delta}{k^2}\left\|\nabla \mathbf{f}(\theta)\right\|_F^2 \leq -\alpha\left\|\theta\right\|^2 - 2\langle \nabla L_S(0), \theta\rangle + \frac{\eta\delta}{k^2}\left(2\langle \nabla \mathbf{f}(\theta), \nabla \mathbf{f}(0)\rangle_F - \left\|\nabla \mathbf{f}(0)\right\|_F^2\right). \end{aligned}$$

By **A3** and Cauchy-Swartz inequality,

$$2\langle -\nabla L_S(\theta), \theta\rangle + \frac{\eta\delta}{k^2}\left\|\nabla \mathbf{f}(\theta)\right\|_F^2 \leq -\alpha\left\|\theta\right\|^2 + 2\left\|\nabla L_S(0)\right\|\left\|\theta\right\| + \frac{2\eta\delta}{k}\ell_f^2.$$

Since $2\left\|\nabla L_S(0)\right\|\left\|\theta\right\| \leq \frac{\alpha}{2}\left\|\theta\right\|^2 + \frac{2}{\alpha}\left\|\nabla L_S(0)\right\|^2$,

$$2\langle -\nabla L_S(\theta), \theta\rangle + \frac{\eta\delta}{k^2}\left\|\nabla \mathbf{f}(\theta)\right\|_F^2 \leq -\frac{\alpha}{2}\left\|\theta\right\|^2 + \frac{2}{\alpha}\left\|\nabla L_S(0)\right\|^2 + \frac{2\eta\delta}{k}\ell_f^2.$$

Then,

$$\begin{aligned}\langle \nabla L_S(\theta), \theta \rangle &\geq \frac{\alpha}{4} \|\theta\|^2 - \frac{1}{\alpha} \|\nabla L_S(0)\|^2 - \frac{\eta\delta}{k}\ell_f^2 + \frac{\eta\delta}{2k^2} \|\nabla \mathbf{f}(\theta)\|_F^2 \\ &\geq \frac{\alpha}{4} \|\theta\|^2 - \frac{1}{\alpha} \|\nabla L_S(0)\|^2 - \frac{\eta_{\max}\delta}{2k}\ell_f^2 \\ &\geq \frac{\alpha}{4} \|\theta\|^2 - \left(\frac{4}{\alpha^2 - 4M^2} + 1\right) \frac{\eta_{\max}}{2k}\delta\ell_f^2.\end{aligned}$$

The last inequality used Lemma A.3. in Farghly and Rebeschini [2021]. Hence, **A1′** is satisfied with $m = \alpha/4$ and $b = \left(\frac{4}{\alpha^2-4M^2} + 1\right) \frac{\eta_{\max}}{2k}\delta\ell_f^2$.

To establish the converse, let's assume, for the sake of contradiction, that for any $\alpha > 0$, $\exists \theta, \theta'$ such that

$$2\langle -\nabla L_S(\theta) + \nabla L_S(\theta'), \theta - \theta' \rangle + \left\| \frac{\sqrt{\eta\delta}}{k} (\nabla \mathbf{f}(\theta) - \nabla \mathbf{f}(\theta'))^\top \right\|_F^2 > -\alpha \|\theta - \theta'\|_2^2 .$$

For $n = 1, 2, 3, \ldots$, let $\alpha = \frac{1}{n}$. Then, for any n, $\exists \theta_n, \theta_n'$ such that

$$2\langle -\nabla L_S(\theta_n) + \nabla L_S(\theta_n'), \theta_n - \theta_n' \rangle + \left\| \frac{\sqrt{\eta\delta}}{k} (\nabla \mathbf{f}(\theta_n) - \nabla \mathbf{f}(\theta_n'))^\top \right\|_F^2 > -\frac{1}{n} \|\theta_n - \theta_n'\|_2^2 .$$

Given that θ is bounded, we can apply the Bolzano-Weierstrass theorem. Consequently, there exists a subsequence (θ'_{n_s}) that converges to the limit point u. As $\|\theta - \theta'\|^2 < 4B^2 \ \ \forall \theta, \theta'$, for sufficiently large s, we have

$$\begin{aligned}&2\langle -\nabla L_S(\theta_{n_s}) + \nabla L_S(u), \theta_{n_s} - u \rangle + \left\| \frac{\sqrt{\eta\delta}}{k} (\nabla \mathbf{f}(\theta_{n_s}) - \nabla \mathbf{f}(u))^\top \right\|_F^2 \geq 0 \\ \Rightarrow\ &\langle -\nabla L_S(\theta_{n_s}), \theta_{n_s} \rangle + \langle \nabla L_S(\theta_{n_s}), u \rangle - \langle \nabla L_S(u), u \rangle + \langle \nabla L_S(u), \theta_{n_s} \rangle \geq -\frac{2\eta\delta}{k}\ell_f^2. \qquad \text{(by } \mathbf{A3}\text{)}\end{aligned}$$

By Cauchy-Swartz,

$$\begin{aligned}&\langle -\nabla L_S(\theta_{n_s}), \theta_{n_s} \rangle + \|\nabla L_S(\theta_{n_s})\| \|u\| - \langle \nabla L_S(u), u \rangle + \|\nabla L_S(u)\| \|\theta_{n_s}\| \geq -\frac{2\eta\delta}{k}\ell_f^2 \\ \Rightarrow\ &\langle -\nabla L_S(\theta_{n_s}), \theta_{n_s} \rangle \geq -\|\nabla L_S(u)\| \|\theta_{n_s}\| - \|\nabla L_S(\theta_{n_s})\| \|u\| + C_u,\end{aligned}$$

where $C_u := \langle \nabla L_S(u), u \rangle - \frac{2\eta\delta}{k}\ell_f^2 \geq m\|u\| - b - \frac{2\eta\delta}{k}\ell_f^2$ by **A1′**.

Then, by **A1′**,

$$\begin{aligned}-m\|\theta_{n_s}\| + b \geq \langle -\nabla L_S(\theta_{n_s}), \theta_{n_s} \rangle &\geq -\|\nabla L_S(u)\| \|\theta_{n_s}\| - \|\nabla L_S(\theta_{n_s})\| \|u\| + C_u \\ \Rightarrow\ \ (\|\nabla L_S(u)\| - m) \|\theta_{n_s}\| &\geq -\|\nabla L_S(\theta_{n_s})\| \|u\| + C_u - b.\end{aligned}$$

Then,

$$\|\nabla L_S(\theta_{n_s})\| \geq -\frac{\|\nabla L_S(u)\| - m}{\|u\|} \|\theta_{n_s}\| + \frac{C_u - b}{\|u\|}.$$

By **A2** and reverse triangle inequality $\|\nabla L_S(\theta_{n_s})\| - \|\nabla L_S(0)\| \leq M\|\theta_{n_s}\|$, so

$$\begin{aligned}M\|\theta_{n_s}\| + \|\nabla L_S(0)\| &\geq \frac{m - \|\nabla L_S(u)\|}{\|u\|} \|\theta_{n_s}\| + \frac{C_u - b}{\|u\|} \\ \Rightarrow\ \ (2M\|u\| + \|\nabla L_S(0)\| - m) \|\theta_{n_s}\| &\geq C_u - b - \|\nabla L_S(0)\| \|u\| .\end{aligned}$$

By Lemma A.3 of Farghly and Rebeschini [2021], from **A1′** and **A2**, above implies

$$(2M\|u\| + M\sqrt{b/m} - m) \|\theta_{n_s}\| \geq C_u - b - M\sqrt{b/m}\,\|u\|$$

$$\Rightarrow \quad (2M\|u\| + M\sqrt{b/m} - m)\|\theta_{n_s}\| \geq (m - M\sqrt{b/m})\|u\| - 2b - \frac{2\eta\delta}{k}\ell_f^2$$
$$\Rightarrow \quad (m - M\sqrt{b/m})(\|u\| + \|\theta_{n_s}\|) \leq 2M\|u\|\|\theta_{n_s}\| + 2b + \frac{2\eta\delta}{k}\ell_k^2.$$

Since $m < M\sqrt{b/m}$ (i.e. $m^3 < M^2 b$),

$$2MB^2 + 2(M\sqrt{b/m} - m)B + (2b + \frac{2\eta\delta}{k}\ell_f^2) \geq 0$$
$$\iff \quad B \notin \left(\frac{1}{2M}\left(m - M\sqrt{b/m} - \sqrt{(M\sqrt{b/m} - m)^2 - 2M(2b + \frac{2\eta\delta}{k}\ell_f^2)}\right)\right.$$
$$\left., \frac{1}{2M}\left(m - M\sqrt{b/m} + \sqrt{(M\sqrt{b/m} - m)^2 - 2M(2b + \frac{2\eta\delta}{k}\ell_f^2)}\right)\right).$$

Hence, for

$$B \in \left(\frac{1}{2M}\left(m - M\sqrt{b/m} - \sqrt{(M\sqrt{b/m} - m)^2 - 2M(2b + \frac{2\eta\delta}{k}\ell_f^2)}\right)\right.$$
$$\left., \frac{1}{2M}\left(m - M\sqrt{b/m} + \sqrt{(M\sqrt{b/m} - m)^2 - 2M(2b + \frac{2\eta\delta}{k}\ell_f^2)}\right)\right),$$

there is a contradiction hence we have a uniform dissipativity. □

8.2 Proofs for the Wasserstein Contraction

8.2.1 Moment Bound

Proof of Lemma 1. Suppose that θ_t is a solution of the SDE (4). By Ito's Lemma (Lemma 9), for any $\phi \in \mathcal{C}^2$,

$$\mathrm{d}\phi = \left\{\frac{\partial\phi}{\partial t} + (\nabla\phi)^\top b_t + \frac{1}{2}\mathrm{Tr}[G_t^\top H G_t]\right\}\mathrm{d}t + (\nabla\phi)^\top G_t \mathrm{d}W_t,$$

with probability 1, where $G_t = \frac{\sqrt{\delta\eta}}{k}\nabla\mathbf{f}(\theta_t, X)^\top$, $b_t = -\nabla L_S(\theta_t, B)$ and $H = \nabla^2\phi$. Consider $\phi(\theta) = \|\theta\|_2^p$. Then,

$$\nabla\phi(\theta) = p\|\theta\|_2^{p-2}\theta,$$
$$\frac{d\phi}{dt} = \nabla\phi(\theta)^\top\frac{\mathrm{d}\theta}{\mathrm{d}t} = p\|\theta\|^{p-2}\theta\left[-\nabla L_S(\theta, B)dt + \frac{\sqrt{\delta\eta}}{k}\nabla\mathbf{f}(\theta_t, X)^\top dW_t\right],$$
$$H_{ij} = (\nabla^2\|\theta\|_2^p)_{ij} = p\{(p-2)\|\theta\|_2^{p-4}\theta_i\theta_j\} + p\delta_{ij}\|\theta\|^{p-2},$$
$$\mathrm{Tr}(H) = p(p+d-2)\|\theta\|_2^{p-2}.$$

We also bound $\mathrm{Tr}[G_t^\top H G_t]$ using Von Neumann's Trace inequality [Mirsky, 1975] as below:

$$\frac{k^2}{\delta\eta}\mathrm{Tr}[G_t^\top H G_t] = \mathrm{Tr}[\nabla\mathbf{f}H\nabla\mathbf{f}^\top] \leq |\mathrm{Tr}[\nabla\mathbf{f}H\nabla\mathbf{f}^\top]| \leq \sum_{i=1}^d \sigma_i(\nabla\mathbf{f}^\top\nabla\mathbf{f})\sigma_i(H) \quad \text{(by Von Neumann's Trace inequality)}$$
$$\leq \sigma_{\max}(\nabla\mathbf{f}^\top\nabla\mathbf{f})\sum_{i=1}^d\sigma_i(H) = \sigma_{\max}(\nabla\mathbf{f}^\top\nabla\mathbf{f})\mathrm{Tr}(H) \qquad \text{(since } H = H^\top \text{ and } H \succeq 0)$$
$$= p(p+d-2)\|\theta\|_2^{p-2}\sigma_{\max}(\nabla\mathbf{f}^\top\nabla\mathbf{f}) = p(p+d-2)\|\theta\|_2^{p-2}\left\|\nabla\mathbf{f}^\top\nabla\mathbf{f}\right\|_2.$$

Here, H is symmetric since it is Hessian and is positive semi-definite since $\|\theta\|_2^p$ is convex in θ. ($\|\cdot\| : \mathbb{R}^n \to [0,\infty)$ is convex by Δ-inequality and $h : x \mapsto x^p$ is non-decreasing and convex for $[0,\infty)$). Since $\nabla\mathbf{f}^\top\nabla\mathbf{f}$ is symmetric positive semidefinite,

$$\left\|\nabla\mathbf{f}^\top\nabla\mathbf{f}\right\|_2 = \sigma_{\max}(\nabla\mathbf{f}^\top\nabla\mathbf{f}) \leq \sum_i\lambda_i(\nabla\mathbf{f}^\top\nabla\mathbf{f}) = \mathrm{Tr}(\nabla\mathbf{f}^\top\nabla\mathbf{f}) = \sum_{i=1}^k\|\nabla f_i\|^2 < k\ell_f^2,$$

where the last inequality follows from **A3**. Thus, $\mathrm{Tr}[G_t^\top H G_t] \le \frac{\delta\eta}{k} p(p+d-2)\left\|\theta\right\|^{p-2}\ell_f^2$. Then, by Itô's lemma,

$$\mathrm{d}\left\|\theta_t\right\|^p \le -2p\left\|\theta_t\right\|^{p-2}\langle\theta_t, \nabla L_S(\theta_t, B)\rangle dt + \frac{\delta\eta}{2k^2}p(p+d-2)\left\|\theta_t\right\|^{p-2}\ell_f^2\mathrm{d}t + 2\frac{\sqrt{\delta\eta}}{k}p\left\|\theta_t\right\|^{p-2}\langle\theta_t, \nabla\mathbf{f}^\top \mathrm{d}W_t\rangle.$$

By (m,b)-dissipativity of L_S, this can be bounded further as:

$$\begin{aligned}\mathrm{d}\left\|\theta_t\right\|^p \le& -2pm\left\|\theta_t\right\|^p dt + p\left\{2b + \frac{\delta\eta}{2k^2}(p+d-2)\ell_f^2\right\}\left\|\theta_t\right\|^{p-2}\mathrm{d}t + 2\frac{\sqrt{\delta\eta}}{k}p\left\|\theta_t\right\|^{p-2}\langle\theta_t, \nabla\mathbf{f}^\top \mathrm{d}W_t\rangle\\ \le& -\frac{pm}{2}\left\|\theta_t\right\|^p dt + p\left\{b + \frac{\delta\eta}{2k^2}(p+d-2)\ell_f^2\right\}^{p/2}(m/2)^{1-p/2}\mathrm{d}t + \frac{\sqrt{\delta\eta}}{k}p\left\|\theta_t\right\|^{p-2}\langle\theta_t, \nabla\mathbf{f}^\top \mathrm{d}W_t\rangle,\end{aligned}$$

where for the second inequality we used Young's inequality with exponents $p/(p-2)$ and $p/2$ and $t = \left(\frac{p-2}{2}\left(\frac{m}{2}\right)^{-p/2}\right)^{-2(p-2)/p^2}$. Then, by multiplying $e^{pmt/2}$ and using product rule,

$$d(e^{pmt/2}\left\|\theta_t\right\|^p) \le e^{pmt/2}p\left\{b + \frac{\delta\eta}{2k^2}(p+d-2)\ell_f^2 2\right\}^{p/2}(m/2)^{1-p/2}\mathrm{d}t + e^{pmt/2}\frac{\sqrt{\delta\eta}}{k}p\left\|\theta_t\right\|^{p-2}\langle\theta_t, \nabla\mathbf{f}^\top \mathrm{d}W_t\rangle.$$

Integrating from $t=0$ to $t=T$,

$$\begin{aligned}\left\|\theta_T\right\|^p \le& e^{-pmT/2}\left\|\theta_0\right\|^p + (1-e^{-pmT/2})p\left\{b + \frac{\delta\eta}{2k^2}(p+d-2)\ell_f^2\right\}^{p/2}\frac{2}{pm}(\frac{m}{2})^{1-p/2}\\ &+ e^{-pmT/2}\int_{t=0}^{t=T} e^{pmt/2}\frac{\sqrt{\delta\eta}}{k}p\left\|\theta_t\right\|^{p-2}\langle\theta_t, \nabla\mathbf{f}^\top \mathrm{d}W_t\rangle.\end{aligned}$$

Taking expectations,

$$\mathbb{E}\left\|\theta_T\right\|^p \le e^{-pmT/2}\mathbb{E}\left\|\theta_0\right\|^p + (1-e^{-pmT/2})\{\frac{2b}{m} + \frac{\delta\eta}{k^2 m}(p+d-2)\ell_f^2\}^{p/2}.$$

Hence,

$$\begin{aligned}\mu P_t^B(\|\cdot\|^p) \le& \mu(\|\cdot\|^p)e^{-pmt/2} + \left[\frac{2b}{m} + \frac{\delta\eta}{k^2 m}(p+d-2)\ell_f^2\right]^{p/2}(1-e^{-pmt/2})\\ \le& \mu(\|\cdot\|^p) + \left[\frac{2b}{m} + \frac{\delta\eta}{k^2 m}(p+d-2)\ell_f^2\right]^{p/2},\end{aligned}$$

as required. □

8.2.2 Moment Estimate Bound

In order to perform our estimations in a continuous-time setting, we introduce an auxiliary continuous-time process. First, recall the stochastic differential equation (SDE) of label noise gradient descent (LNGD):

$$d\theta_t = -\nabla L_S(\theta_t, B)dt + \frac{\sqrt{\delta\eta}}{k}\left(\nabla\mathbf{f}(\theta_t, X_B)\right)^\top dW_t, \quad \theta_0 \sim \mu_0, \tag{14}$$

where $(W_t)_{t\ge 0}$ is a standard k-dimensional Wiener process.

It is worth noting that the SDE (14) has a unique solution on $\mathbb{R}^+$, since the smoothness assumption (**A2**) holds for L_S. Hence, we define, for each $\eta > 0$, a convenient time-changed version of Θ_t as

$$\theta_t^\eta := \theta_{\eta t}.$$

Then, $\tilde{W}_t^\eta := W_{\eta t}/\sqrt{\eta}$ is also a Wiener process and

$$d\theta_t^\eta = -\eta L_S(\theta_t^\eta, B)dt + \frac{\eta\sqrt{\delta}}{k}\left(\nabla\mathbf{f}(\theta_t^\eta, X_B)\right)^\top d\tilde{W}_t^\eta, \quad \theta_0^\eta \sim \mu_0.$$

We proceed with the required moment estimate that is essential for the derivation of the main result in Theorem 2. These estimates will also enable us to calculate how far the process θ_t^η diverges from its initial condition in one step in the derivation of Theorem 3. To prove these bounds, we will heavily rely on the auxiliary process defined above and perform significant calculations.

Proof of Lemma 2. First, note that by Itô's isometry and commutativity of trace operator with expectation and integral,

$$\begin{aligned}\mathrm{Tr}\left(\mathrm{Var}\left(\int_{u=s}^{t}\frac{\eta\sqrt{\delta}}{k}(\nabla\mathbf{f}(\theta_u))^\top dW_u\right)\right) &= \frac{\delta\eta^2}{k^2}\mathrm{Tr}\left(\mathbb{E}\left[\int_s^t \nabla\mathbf{f}(\theta_u)^\top\nabla\mathbf{f}(\theta_u)du\right]\right)\\ &= \frac{\delta\eta^2}{k^2}\mathbb{E}\left[\int_s^t \mathrm{Tr}\left(\nabla\mathbf{f}(\theta_u)^\top\nabla\mathbf{f}(\theta_u)\right)du\right]\\ &= \frac{\delta\eta^2}{k^2}\mathbb{E}\left[\int_s^t\sum_{i=1}^n\|\nabla f_i(\theta_u)\|^2 du\right] \le \frac{\delta\eta^2}{k}\ell_f^2(t-s).\end{aligned}$$

For any $s\in\mathbb{N}$ and $t\in(s,s+1]$, define $\Delta_{s,t}=\theta_s-\eta\nabla L_S(\theta_s,B)(t-s)$. Note that for a vector v, $\mathbb{E}\|v\|^2=\|\mathbb{E}(v)\|^2+\mathrm{Tr}(\mathrm{Var}(v))$. Then, for $t\in(s,s+1]$,

$$\begin{aligned}\mathbb{E}[\|\theta_t^\eta\|^2\mid\theta_s^\eta] &= \mathbb{E}[\theta_t^{\eta\top}\theta_t^\eta\mid\theta_s^\eta] = \mathbb{E}[\mathrm{Tr}(\theta_t^{\eta\top}\theta_t^\eta)\mid\theta_s^\eta] = \mathbb{E}[\mathrm{Tr}(\theta_t^\eta\theta_t^{\eta\top})\mid\theta_s^\eta] = \mathrm{Tr}(\mathbb{E}[\theta_t^\eta\theta_t^{\eta\top}\mid\theta_s^\eta])\\ &= \mathbb{E}[\theta_t^\eta\mid\theta_s^\eta]^\top\mathbb{E}[\theta_t^\eta\mid\theta_s^\eta]+\mathrm{Tr}\left(\mathrm{Var}\left(\int_{u=s}^t\frac{\sqrt{\delta\eta}}{k}(\nabla\mathbf{f}(\theta_u))^\top dW_u\right)\right)\\ &\le \|\mathbb{E}[\theta_t^\eta\mid\theta_s^\eta]\|^2+\frac{\delta\eta^2}{k}\ell_f^2(t-s) = \|\Delta_{s,t}\|^2+\frac{\delta\eta^2}{k}\ell_f^2(t-s)).\end{aligned}$$

Here, $\forall\eta\le\eta_{\max}:=\min\{\frac{1}{m},\frac{m}{2M^2}\}$,

$$\begin{aligned}\|\Delta_{s,t}\|^2 =\ & \|\theta_s^\eta\|^2-2\eta(t-s)\langle\theta_s^\eta,\nabla L_S(\theta_s^\eta,B)\rangle+\eta^2\|\nabla L_S(\theta_s^\eta,B)(t-s)\|^2\\ &\le (1-2m\eta(t-s))\|\theta_s^\eta\|^2+2b\eta(t-s)+\eta^2(t-s)^2\|\nabla L_S(\theta_s^\eta,B)\|^2 && \text{(by } \mathbf{A1}')\\ &\le (1-2m\eta(t-s))\|\theta_s^\eta\|^2+2b\eta(t-s)+2\eta^2(t-s)^2\left\{M^2\|\theta_s^\eta\|^2+\|\nabla L_S(0,B)\|^2\right\} && \text{(by } \mathbf{A2})\\ &\le (1-2m\eta(t-s))\|\theta_s^\eta\|^2+2b\eta(t-s)+2\eta^2(t-s)^2\left\{M^2\|\theta_s^\eta\|^2+\frac{M^2b}{m}\right\}\\ &&& \text{(by Lemma A.3 in Farghly and Rebeschini [2021])}\\ &\le (1-m\eta(t-s))\|\theta_s^\eta\|^2+2b\eta(t-s)\left(1+\eta_{\max}\frac{M^2}{m}\right) && (15)\\ &\le (1-m\eta(t-s))\|\theta_s^\eta\|^2+6b\eta(t-s),\end{aligned}$$

where the fourth inequality is from that $2\eta^2(t-s)^2M^2\le 2\eta^2(t-s)M^2\le m\eta(t-n)$. For higher moments, the computation is more complex. To simplify the calculation, let $U_{s,t}^\eta:=\frac{\eta\sqrt{\delta}}{k}\int_s^t\nabla\mathbf{f}(\theta_r^\eta)^\top d\tilde{W}_r^\eta$ be defined. Then, for $t\in[s,s+1)$,

$$\begin{aligned}\mathbb{E}[|\theta_t^\eta|^{2p}\mid\theta_s^\eta] &= \mathbb{E}\left[\left\|\theta_s^\eta-\eta\nabla L_S(\theta_s^\eta,B)(t-s)+U_{s,t}^\eta\right\|^{2p}\mid\theta_s^\eta\right] = \mathbb{E}\left[\left\|\Delta_{s,t}+U_{s,t}^\eta\right\|^{2p}\mid\theta_s^\eta\right]\\ &\le \|\Delta_{s,t}\|^{2p}+2p\mathbb{E}\left[\|\Delta_{s,t}\|^{2p-2}\langle\Delta_{s,t},U_{s,t}^\eta\rangle\mid\theta_s^\eta\right]+\sum_{k=2}^{2p}\binom{2p}{k}\mathbb{E}\left[\|\Delta_{s,t}\|^{2p-k}\left\|U_{s,t}^\eta\right\|^k\mid\theta_s^\eta\right]\\ &&& \text{(by Lemma A.3 in Chau et al. [2021])}\\ &\le \|\Delta_{s,t}\|^{2p}+p(2p-1)\mathbb{E}\left[(\|\Delta_{s,t}\|+\left\|U_{s,t}^\eta\right\|)^{2p-2}\|U_{s.t}^\eta\|^2\mid\theta_s^\eta\right]\\ &&& \text{(as in the proof of Lemma 3.9 in Chau et al. [2021])}\\ &\le \|\Delta_{s,t}\|^{2p}+p(2p-1)\cdot\frac{\delta\eta^2}{k}\ell_f^2(t-s)\|\Delta_{s,t}\|^{2p-2}+p(2p-1)\mathbb{E}[\left\|U_{s,t}^\eta\right\|^{2p}]\\ &\le \|\Delta_{s,t}\|^{2p}+p(2p-1)\cdot\frac{\delta\eta^2}{k}\ell_f^2(t-s)\|\Delta_{s,t}\|^{2p-2}+\{p(2p-1)\}^{p+1}(t-s)^p\frac{\delta^p\eta^{2p}}{k^p}\ell_f^{2p}.\\ &&& \text{(by Theorem 7.1 in Mao [2011])}\end{aligned}$$

Note the following inequality for further analysis

$$(r+s)^p\le(1+\varepsilon)^{p-1}r^p+(1+\varepsilon^{-1})^{p-1}s^p, \tag{16}$$

where $p \geq 2, r, s, \geq 0$ and $\varepsilon > 0$. Letting $\varepsilon = m\eta(t-s)/2$,

$$\begin{aligned}
\|\Delta_{s,t}\|^{2p} &\leq \left[(1 - m\eta(t-s))\,\|\theta_s^\eta\|^2 + 3b\eta(t-s)\right]^p && \text{(by (15))}\\
&\leq \left(1 + \frac{m\eta(t-s)}{2}\right)^{p-1} (1 - m\eta(t-s))^p\,\|\theta_s^\eta\|^{2p} + \left(1 + \frac{2}{m\eta(t-s)}\right)^{p-1} \eta^p (t-s)^p\,(3b)^p && \text{(by (16))}\\
&\leq a_{s,t}^{\eta,p}\,\|\theta_s^\eta\|^{2p} + b_{s,t}^{\eta,p},
\end{aligned}$$

where $a_{s,t}^{\eta,p} = (1 - m\eta(t-s)/2)^{p-1}(1 - m\eta(t-s))$ and $b_{s,t}^{\eta,p} = (\eta(t-s) + 2/m)^{p-1}\eta(t-s)\,(3b)^p$. Substituting it yields

$$\begin{aligned}
\mathbb{E}[|\theta_t^\eta|^{2p} \mid \theta_s^\eta] \leq a_{s,t}^{\eta,p}\,\|\theta_s^\eta\|^{2p} + b_{s,t}^{\eta,p} + p(2p-1)\cdot\frac{\delta\eta^2}{k}\ell_f^2(t-s)\left[a_{s,t}^{\eta,p-1}\,\|\theta_s^\eta\|^{2(p-1)} + b_{s,t}^{\eta,p-1}\right]\\
+ \{p(2p-1)\}^{p+1}(t-s)^p\frac{\delta^p\eta^{2p}}{k^p}\ell_f^{2p}.
\end{aligned}$$

Define $\widetilde{M}(p) = \sqrt{\frac{4p(2p-1)\delta\eta\ell_f^2}{mk}}$. Then, for $\|\theta_s^\eta\| \geq \widetilde{M}(p)$,

$$\frac{m\eta(t-s)}{4}\,\|\theta_s^\eta\|^{2p} \geq p(2p-1)\frac{\delta\eta^2}{k}\ell_f^2(t-s)\,\|\theta_s^\eta\|^{2(p-1)}.$$

Hence,

$$\begin{aligned}
\mathbb{E}[|\theta_t^\eta|^{2p} \mid \theta_s^\eta] &\leq (1 - m\eta(t-s)/4)a_{s,t}^{\eta,p-1}\,\|\theta_s^\eta\|^{2p} + b_{s,t}^{\eta,p} + \eta(t-s)p(2p-1)\frac{\delta\eta}{k}\ell_f^2 b_{s,t}^{\eta,p-1}\\
&\quad + \eta^p(t-s)^p\{p(2p-1)\}^{p+1}\frac{\delta^p\eta^p}{k^p}\ell_f^{2p}\\
&\leq (1 - m\eta(t-s))\,\|\theta_s^\eta\|^{2p} + \eta(t-s)M(p,\eta,k) \leq \|\theta_s^\eta\|^{2p} + \eta M(p,\eta,k),
\end{aligned}$$

where

$$\begin{aligned}
M(p,\eta,k) &:= (\eta(t-s) + 2/m)^{p-1}\,(3b)^p + \eta(t-s)p(2p-1)\frac{\delta\eta}{k}\ell_f^2(\eta(t-s) + 2/m)^{p-2}\,(3b)^{p-1}\\
&\quad + \eta^{p-1}(t-s)^{p-1}\{p(2p-1)\}^{p+1}\frac{\delta^p\eta^p}{k^p}\ell_f^{2p}\\
&\leq (\eta + 2/m)^{p-1}\,(3b)^p + \eta^2 p(2p-1)\frac{\delta}{k}\ell_f^2(\eta + 2/m)^{p-2}\,(3b)^{p-1} + \eta^{2p-1}\{p(2p-1)\}^{p+1}\frac{\delta^p}{k^p}\ell_f^{2p}\\
&=: \frac{1}{\eta}\tilde{c}(p).
\end{aligned}$$

Similarly, for $\|\theta_s^\eta\| < \widetilde{M}(p)$ we attain

$$\mathbb{E}[|\theta_t^\eta|^{2p} \mid \theta_s^\eta] \leq \|\theta_s^\eta\|^{2p} + \tilde{c}(p).$$

Hence, we have

$$\mathbb{E}[|\theta_t^\eta|^{2p} \mid \theta_s^\eta] \leq \|\theta_s^\eta\|^{2p} + \tilde{c}(p),$$

as required. □

8.2.3 ρ_g-Wasserstein Distance and 2-Wasserstein Distance

Proof of Lemma 3. Let $f : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ be a function such that $f(x) = \sqrt{x}$ in its domain. Then, since f is concave and $f(0) = 0$, we can apply Lemma 1 in Wunder et al. [2021]. Let

$$r := r(p) := \frac{\mathbb{E}[g(\|X - Y\|^p)]}{\mathbb{E}[g(\|X - Y\|)]^p} \geq 1$$

be the ratio of first and p-th non-centralized moment and

$$\zeta_r(a) := \sup_{\frac{1}{p}+\frac{1}{q}=1} \frac{[1 - r(p)^{\frac{1}{p}} a^{-\frac{1}{q}}]^+}{a}.$$

Then, $\forall \pi \in \mathcal{C}(\mu,\nu)\ \ \forall 1 \le r < a$,

$$\begin{aligned}\left(\int \|x-y\|^2 \pi(dx,dy)\right)^{1/2} &\le \frac{1}{\varphi}\left(\int g(\|x-y\|)^2 \pi(dx,dy)\right) = \frac{1}{\varphi a \zeta_b(a)} \cdot \left(\int a g(\|x-y\|)^2 \pi(dx,dy)\right)^{1/2} \cdot \zeta_b(a) \\ &\le \frac{1}{\varphi a \zeta_b(a)} \cdot \sup_{s>t}\left(\int s g(\|x-y\|)^2 \pi(dx,dy)\right)^{1/2} \cdot \zeta_t(s) \\ &\le \frac{1}{\varphi a \zeta_b(a)} \cdot \int g(\|x-y\|)\pi(dx,dy) \le \frac{1}{\varphi a \zeta_b(a)} \cdot \int \rho_g(x,y)\pi(dx,dy).\end{aligned}$$

Hence,

$$W_2(\mu,\nu) := \inf_{\pi\in\mathcal{C}(\mu,\nu)}\left(\int \|x-y\|^2 \pi(dx,dy)\right)^{1/2} \le \frac{1}{\varphi a \zeta_b(a)} \cdot \inf_{\pi\in\mathcal{C}(\mu,\nu)} \int \rho_g(x.y)\pi(dx,dy) = \frac{1}{\varphi a \zeta_b(a)} \cdot W_{\rho_g}(\mu,\nu).$$

Note that since $r(p) \ge 1$, we have

$$\inf_{\frac{1}{p}+\frac{1}{q}=1} r(p)^{1/p} a^{-1/q} \ge \inf_{\frac{1}{p}+\frac{1}{q}=1} a^{-1/q} \ge \inf_q a^{-1/q} \ge 0.$$

Then, we can obtain an upper bound for $\zeta_r(a)$ as below:

$$\zeta_r(a) = \frac{1}{a}\sup_{\frac{1}{p}+\frac{1}{q}=1}\left[1 - r(p)^{1/p}a^{-1/q}\right]^+ = \frac{1}{a}\max\left\{0, 1 - \inf_{\frac{1}{p}+\frac{1}{q}=1} r(p)^{1/p}a^{-1/q}\right\} \le \frac{1}{a}.$$

Similarly, we can obtain a lower bound as:

$$\zeta_r(a) = \frac{1}{a}\sup_{\frac{1}{p}+\frac{1}{q}=1}\left[1 - r(p)^{1/p}a^{-1/q}\right]^+ > \frac{1}{a}\left[1 - r(p)^0 a^{-1}\right]^+ = \frac{a-1}{a^2}.$$

□

8.3 Proofs for the Generalization Error Bound

8.3.1 Divergence Bound

Proof of Lemma 5. Integrating the SDE from 0 to t,

$$\theta_t - \theta_0 = -\int_0^t \nabla L_S(\theta_s, B)ds + \int_0^t \frac{\sqrt{\delta\eta}}{k}(\nabla \mathbf{f}(\theta_s))^\top dW_s.$$

Then, by Jensen's inequality of integrals,

$$\|\theta_t - \theta_0\|^2 \le 2\int_0^t \|\nabla L_S(\theta_s, B)\|^2 ds + 2\left\|\int_{s=0}^t \frac{\sqrt{\delta\eta}}{k}(\nabla \mathbf{f}(\theta_s))^\top dW_s\right\|^2.$$

Note that for a vector v, $\mathbb{E}\|v\|^2 = \|\mathbb{E}(v)\|^2 + \mathrm{Tr}(\mathrm{Var}(v))$. Then, by Itô's isometry and commutativity of trace operator with expectation and integral, the expectation is

$$\begin{aligned}\mathbb{E}\|\theta_t - \theta_0\|^2 &\le 2\int_0^t \mathbb{E}\|\nabla L_S(\theta_s, B)\|^2 ds + 2\mathrm{Tr}\left(\mathrm{Var}\left(\int_{s=0}^t \frac{\sqrt{\delta\eta}}{k}(\nabla \mathbf{f}(\theta_s))^\top dW_s\right)\right) \\ &= 2\int_0^t \mathbb{E}\|\nabla L_S(\theta_s, B)\|^2 ds + \frac{2\delta\eta}{k^2}\mathrm{Tr}\left(\mathbb{E}\left[\int_0^t \nabla \mathbf{f}(\theta_s)^\top \nabla \mathbf{f}(\theta_s) ds\right]\right) \\ &= 2\int_0^t \mathbb{E}\|\nabla L_S(\theta_s, B)\|^2 ds + \frac{2\delta\eta}{k^2}\mathbb{E}\left[\int_0^t \mathrm{Tr}\left(\nabla \mathbf{f}(\theta_s)^\top \nabla \mathbf{f}(\theta_s)\right) ds\right] \\ &= 2\int_0^t \mathbb{E}\|\nabla L_S(\theta_s, B)\|^2 ds + \frac{2\delta\eta}{k^2}\mathbb{E}\left[\int_0^t \sum_{i=1}^n \|\nabla f_i(\theta_s)\|^2 ds\right],\end{aligned}$$

where the last equality is from:

$$\mathrm{Tr}(\nabla \mathbf{f}^\top \nabla \mathbf{f}) = \mathrm{Tr}(\sum_{i=1}^n \nabla f_i \nabla f_i^\top) = \sum_{i=1}^n \mathrm{Tr}(\nabla f_i \nabla f_i^\top) = \sum_{i=1}^n \|\nabla f_i\|^2 .$$

Using Lemma A.3 in Farghly and Rebeschini [2021], we can find the upper bound of the first term,

$$\begin{aligned}
\mathbb{E}\,\|\nabla L_S(\theta_s, B)\|^2 &\le 2\mathbb{E}\,\|\nabla L_S(\theta_s, B) - \nabla L_S(0, B)\|^2 + 2\mathbb{E}\,\|\nabla L_S(0, B)\|^2 \\
&\le 2M^2\,\|\theta_s\|^2 + 2M^2\frac{b}{m} && \text{(by Lemma A.3 in Farghly and Rebeschini [2021])} \\
&\le 2M^2(\mathbb{E}\,\|\theta_0\|^2 + 2M^2\left[\frac{3b}{m} + \frac{\delta\eta d}{k^2 m}\left\|\nabla \mathbf{f}(\theta_s)^\top \nabla \mathbf{f}(\theta_s)\right\|\right]. && \text{(by Lemma 1)}
\end{aligned}$$

Hence, we get

$$\mathbb{E}\,\|\theta_t - \theta_0\|^2 \le 4M^2\left[\mathbb{E}\,\|\theta_0\|^2 + \frac{3b}{m} + \frac{\delta\eta d}{k^2 m}\left\|\nabla \mathbf{f}(\theta_t)^\top \nabla \mathbf{f}(\theta_t)\right\|\right] t + \frac{2\delta\eta}{k^2}\mathbb{E}\left[\int_0^t \sum_{i=1}^n \|\nabla f_i(\theta_s)\|^2\, ds\right].$$

□

8.3.2 Discretization Error Bound

This section aims to derive the discretization error bounds using synchronous-type coupling. Synchronous coupling is a way to pair samples from two probability distributions that is commonly used to estimate the error between continuous-time and discrete-time stochastic processes. Given two probability measures μ and ν on a common space $\mathcal{W}$, a synchronous coupling of μ and ν is a joint probability measure π on $\mathcal{W} \times \mathcal{W}$ such that the marginals of π are μ and ν, respectively, and $\pi(w, w') = 0$ whenever $w \neq w'$. This means that in a synchronous coupling, each sample drawn from μ is always paired with a sample drawn from ν in a one-to-one manner.

Recall from Section 2.2 that we denote R_θ to be a Markov kernel of our algorithm $(\theta_t)_{t=0}^\infty$ and R_Θ to be a Markov kernel of our discrete-time process $(\Theta_{t\eta})_{t=0}^\infty$. Our main objective here is to estimate the Wasserstein distance between μR_θ and μR_Θ, where μ represents an arbitrary probability measure. We focus on obtaining bounds on the Wasserstein distance between two distributions: μR_θ^B, which is the distribution of one step of a label noise SGD with fixed mini-batch B, and μP_η^B. To do this, we define a coupling $(\tilde{\theta}_t, \lambda_{\eta t})$ for $t \in [0, 1]$,

$$\begin{aligned}
d\lambda_t &= -\nabla L_S(\lambda_t, B)dt + \frac{\sqrt{\delta\eta}}{k}\left(\nabla \mathbf{f}(\lambda_t, X_B)\right)^\top dW_t, \qquad \lambda_0 \sim \mu, \\
\tilde{\theta}_t &= \tilde{\theta}_0 - \nabla L_S(\tilde{\theta}_0, B)\eta t + \int_0^{\eta t} \frac{\sqrt{\delta\eta}}{k}\left(\nabla \mathbf{f}(\tilde{\theta}_0, X_B)\right)^\top dW_s, \qquad \tilde{\theta}_0 = \lambda_0.
\end{aligned}$$

Then, by the convexity of the Wasserstein distance (Lemma 12),

$$W_{\rho_g}(\mu R_\theta, \mu R_\Theta) = \binom{n}{k}^{-1} \sum_{B \subset [n], |B| = k} W_{\rho_g}(\mu R_\theta^B, \mu P_\eta^B),$$

so we derive a bound for Wasserstein distance $W(\mu R_\theta, \mu R_\Theta)$ as in the following lemma.

Proof of Lemma 6. Integrating from $s = 0$ to ηt,

$$\lambda_{\eta t} = \lambda_0 - \int_0^{\eta t} \nabla L_S(\lambda_s, B)ds + \int_0^{\eta t} \frac{\sqrt{\delta\eta}}{k}\left(\nabla \mathbf{f}(\lambda_s, X_B)\right)^\top dW_s.$$

Then, by change of variable,

$$\lambda_{\eta t} - \tilde{\theta}_t = -\eta \int_0^t \nabla L_S(\lambda_{\eta s}, B) - \nabla L_S(\tilde{\theta}_0, B)ds + \int_0^{\eta t} \frac{\sqrt{\delta\eta}}{k}\left(\nabla \mathbf{f}(\lambda_s, X_B) - \nabla \mathbf{f}(\tilde{\theta}_0, X_B)\right)^\top dW_s.$$

So, by Jensen's inequality and Itô's isometry as in the proof of Lemma 5,

$$\begin{aligned}\mathbb{E}\left\|\lambda_{\eta t}-\tilde{\theta}_t\right\|^2 &\leq 2\eta^2\int_0^t \mathbb{E}\left\|\nabla L_S(\lambda_{\eta s},B)-\nabla L_S(\tilde{\theta}_0,B)\right\|^2 ds \\ &\quad + 2\frac{\delta\eta}{k^2}\mathbb{E}\int_0^{\eta t}\operatorname{Tr}\left[\left(\nabla\mathbf{f}(\lambda_s,X_B)-\nabla\mathbf{f}(\tilde{\theta}_0,X_B)\right)^\top\left(\nabla\mathbf{f}(\lambda_s,X_B)-\nabla\mathbf{f}(\tilde{\theta}_0,X_B)\right)\right]ds \\ &\leq 2\eta^2M^2\int_0^t\mathbb{E}\left\|\lambda_{\eta s}-(\tilde{\theta}_s-\tilde{\theta}_s)-\tilde{\theta}_0\right\|^2 ds + 2\frac{\delta\eta}{k^2}\mathbb{E}\left[\int_0^{\eta t}4k\ell_f^2 ds\right] \\ &\leq 4\eta^2M^2\int_0^t\mathbb{E}\left\|\lambda_{\eta s}-\tilde{\theta}_s\right\|^2 ds + 4\eta^2M^2\int_0^t\mathbb{E}\left\|\tilde{\theta}_s-\tilde{\theta}_0\right\|^2 ds + \frac{8\delta\eta^2 t}{k}\ell_f^2,\end{aligned}$$

where the second inequality is from **A2, A3**. We can also bound the second term by **A2, A3** and Lemma A.3 in Farghly and Rebeschini [2021] as

$$\begin{aligned}\mathbb{E}\left\|\tilde{\theta}_s-\tilde{\theta}_0\right\|^2 &\leq 2\eta^2s^2\mathbb{E}\left\|\nabla L_S(\tilde{\theta}_0,B)\right\|^2 + 2\frac{\delta\eta}{k^2}\mathbb{E}\left\|\int_0^{\eta s}\left(\nabla\mathbf{f}(\tilde{\theta}_u,X_B)\right)^\top dW_u\right\|^2 \\ &\leq 2\eta^2s^2\mathbb{E}\left\{2M^2\left\|\tilde{x}_0\right\|^2+2M^2\frac{b}{m}\right\} + 2\frac{\delta\eta}{k^2}\mathbb{E}\left[\int_0^{\eta s}\operatorname{Tr}\left(\nabla\mathbf{f}(\tilde{\theta}_u,X_B)^\top\nabla\mathbf{f}(\tilde{\theta}_u,X_B)\right)du\right] \\ &\leq 4\eta^2s^2M^2\left\{\mathbb{E}\left\|\tilde{x}_0\right\|^2+\frac{b}{m}\right\}+\frac{2\delta\eta^2 s}{k}\ell_f^2.\end{aligned}$$

Applying Grönwall's inequality with $\phi(t)=\mathbb{E}\left\|\lambda_{\eta t}-\tilde{\theta}_t\right\|^2$, we have

$$\begin{aligned}\mathbb{E}\left\|\lambda_{\eta t}-\tilde{\theta}_t\right\|^2 &\leq 8\eta^2\exp\left(4\eta^2M^2t\right)\left[\eta^2M^2\left\{\frac{2}{3}t^3M^2\left(\mu(\|\cdot\|^2)+\frac{b}{m}\right)+\frac{\delta t^2}{2k}\ell_f^2\right\}+\frac{\delta\eta^2 t}{k}\ell_f^2\right] \\ &= 8\eta^4\exp\left(4\eta^2M^2t\right)\left[\frac{2}{3}M^4t^3\left(\mu(\|\cdot\|^2)+\frac{b}{m}\right)+(M^2t^2+2t)\frac{\delta}{2k}\ell_f^2\right].\end{aligned}$$

For $t=1$,

$$\mathbb{E}\left\|\lambda_\eta-\tilde{\theta}_1\right\|^2 \leq 8\eta^4\exp\left(4\eta^2M^2\right)\left[\frac{2}{3}M^4\left(\mu(\|\cdot\|^2)+\frac{b}{m}\right)+(M^2+2)\frac{\delta}{2k}\ell_f^2\right].$$

□

Remark 7. *The bound for the discrete-time algorithm is obtained by adding the one-step discretization error to the bound for the continuous-time algorithm.*

8.3.3 Completing the Proof of Theorem 3

With all the necessary components established, we are now able to finalize the proof of Theorem 3.

Proof of Theorem 3. Using Lemma 13, we have the following inequality:

$$W_{\rho_g}(\mu P_\eta^B,\nu\widehat{P}_\eta^B)\leq W_{\rho_g}(\mu,\nu)+\tau_\Delta^{1/2}(1+2\varepsilon+6\varepsilon\tau^{1/2}).$$

Here, τ_Δ and τ are calculated using the divergence bound (Lemma 5) and the moment bound (Lemma 1) with $p=4$, which are given by:

$$\begin{aligned}\tau_\Delta &:= \mathbb{E}\left\|\theta_\eta-\theta_0\right\|^2\vee\mathbb{E}\left\|\widehat{\theta}_\eta-\widehat{\theta}_0\right\|^2 = 4M^2\left[\sigma_\Delta^{1/2}+\frac{3b}{m}+\frac{\delta\eta d}{km}\ell_f^2\right]\eta+\frac{2\delta}{k}\ell_f^2\eta^2, \\ \tau &:= \mathbb{E}\left\|\theta_0\right\|^4\vee\mathbb{E}\left\|\widehat{\theta}_0\right\|^4\vee\mathbb{E}\left\|\theta_\eta\right\|^4\vee\mathbb{E}\left\|\widehat{\theta}_\eta\right\|^4 = \sigma_\Delta+\left[\frac{2b}{m}+\frac{\delta\eta}{km}(d+2)\ell_f^2\right]^2,\end{aligned}$$

where $\sigma_\Delta = \mu(\|\cdot\|^4) \vee \nu(\|\cdot\|^4)$. Then we can further proceed as:

$$
\begin{aligned}
&W_{\rho_g}(\mu P_\eta^B, \nu \widehat{P}_\eta^B) \\
&\le W_{\rho_g}(\mu,\nu) + 2\eta^{1/2}\left[M^2\left(\sigma_\Delta^{1/2} + \frac{3b}{m} + \frac{\delta\eta d}{km}\ell_f^2\right) + \frac{\delta\eta}{2k}\ell_f^2\right]^{1/2} \cdot \left[1 + 2\varepsilon + 6\varepsilon\sigma_\Delta^{1/2} + 6\varepsilon\left(\frac{2b}{m} + \frac{\delta\eta}{km}(d+2)\ell_f^2\right)\right] \\
&\le W_{\rho_g}(\mu,\nu) + \left[\eta^{1/2}\left\{4M^2\left(\sigma_\Delta^{1/2} + \frac{3b}{m}\right)\right\}^{1/2} + \frac{\eta}{k^{1/2}}\left\{\left(\frac{2d}{m}M^2 + 1\right)\ell_f^2\delta\right\}^{1/2}\right] \\
&\quad \cdot \left[1 + 2\varepsilon + 6\varepsilon\left(\sigma_\Delta^{1/2} + \frac{2b}{m}\right) + 6\varepsilon\frac{\eta}{k}\left(\frac{\delta(d+2)}{m}\ell_f^2\right)\right] \\
&\le W_{\rho_g}(\mu,\nu) + \frac{\eta^2}{k^{3/2}}\left[6\varepsilon\frac{\delta(d+2)\ell_f^2}{m}\left\{\left(\frac{2d}{m}M^2+1\right)\ell_f^2\delta\right\}^{1/2}\right] + \frac{\eta^{3/2}}{k}\left[6\varepsilon\frac{\delta(d+2)\ell_f^2}{m}\left\{4M^2\left(\sigma_\Delta^{1/2} + \frac{3b}{m}\right)\right\}^{1/2}\right] \\
&\quad + \frac{\eta}{k^{1/2}}\left[\left\{\left(\frac{2d}{m}M^2+1\right)\ell_f^2\delta\right\}^{1/2}\left\{1 + 2\varepsilon + 6\varepsilon\left(\sigma_\Delta^{1/2} + \frac{2b}{m}\right)\right\}\right] \\
&\quad + \eta^{1/2}\left[\left\{4M^2\left(\sigma_\Delta^{1/2} + \frac{3b}{m}\right)\right\}^{1/2}\left\{1 + 2\varepsilon + 6\varepsilon\left(\sigma_\Delta^{1/2} + \frac{2b}{m}\right)\right\}\right].
\end{aligned}
$$

Assuming $\mu = \mu_0 R_\Theta^t$ and $\nu = \mu_0 \widehat{R}_\Theta^t$ for some t, we can use the moment estimate bound (Lemma 2) to obtain $\sigma_\Delta \le \mu_0(\|\cdot\|^4) + \tilde{c}(2) = \sigma_4 + \tilde{c}(2)$. Therefore, we have:

$$
W_{\rho_g}(\mu P_\eta^B, \nu \widehat{P}_\eta^B) \le W_{\rho_g}(\mu,\nu) + \tilde{c}_1\frac{\eta^2}{k^{3/2}} + \tilde{c}_2\frac{\eta^{3/2}}{k} + \tilde{c}_3\frac{\eta}{k^{1/2}} + \tilde{c}_4\eta^{1/2},
$$

with parameters

$$
\begin{aligned}
\tilde{c}_1 &:= 6\varepsilon\frac{\delta(d+2)\ell_f^2}{m}\left\{\left(\frac{2d}{m}M^2+1\right)\ell_f^2\delta\right\}^{1/2}, \\
\tilde{c}_2 &:= 6\varepsilon\frac{\delta(d+2)\ell_f^2}{m}\left\{4M^2\left(\sigma_4^{1/2} + \tilde{c}(2)^{1/2} + \frac{3b}{m}\right)\right\}^{1/2}, \\
\tilde{c}_3 &:= \left\{\left(\frac{2d}{m}M^2+1\right)\ell_f^2\delta\right\}^{1/2}\left\{1 + 2\varepsilon + 6\varepsilon\left(\sigma_4^{1/2} + \tilde{c}(2)^{1/2} + \frac{2b}{m}\right)\right\}, \\
\tilde{c}_4 &:= \left\{4M^2\left(\sigma_4^{1/2} + \tilde{c}(2)^{1/2} + \frac{3b}{m}\right)\right\}^{1/2}\left\{1 + 2\varepsilon + 6\varepsilon\left(\sigma_4^{1/2} + \tilde{c}(2)^{1/2} + \frac{2b}{m}\right)\right\},
\end{aligned}
$$

where

$$
\tilde{c}(2) = \left\{\frac{18b^2}{m} + 9b^2\eta_{\max} + 18b\delta\ell_f^2\eta_{\max}^2 + 216\delta^2\ell_f^4\eta_{\max}^3\right\}\eta_{\max}.
$$

Without loss of generality, assume that the datasets S and $\widehat{S}$ differs only at i^{th} element. Considering that $\mathbb{P}(i \in B) = k/n$, the convexity of ρ_g-Wasserstein distance (Lemma 12) gives the following inequality:

$$
\begin{aligned}
W_{\rho_g}(\mu R_\Theta, \nu \widehat{R}_\Theta) &\le \frac{k}{n}\sup_{B:n\in B} W_{\rho_g}(\mu P_\eta^B, \nu \widehat{P}_\eta^B) + \left(1 - \frac{k}{n}\right)\sup_{B:n\notin B} W_{\rho_g}(\mu P_\eta^B, \nu \widehat{P}_\eta^B) \\
&\le \tilde{c}_5 W_{\rho_g}(\mu,\nu) + \frac{1}{n}\left[\tilde{c}_1\frac{\eta^2}{k^{1/2}} + \tilde{c}_2\eta^{3/2} + \tilde{c}_3\eta k^{1/2} + \tilde{c}_4\eta^{1/2}k\right],
\end{aligned}
$$

where $\tilde{c}_5 := \frac{k}{n} + \left(1 - \frac{k}{n}\right)C_1 e^{-\alpha t}$. In the second inequality, the second term is bounded by the contraction result in Theorem 2. By appropriate choice of ε as denoted in Section 8.3.4, we have $\tilde{c}_5 < 1$, thus by induction,

$$
W_{\rho_g}(\mu_0 R_\Theta^t, \mu_0 \widehat{R}_\Theta^t) \le \frac{1 - \tilde{c}_5^t}{1 - \tilde{c}_5} \cdot \frac{1}{n}\left[\tilde{c}_1\frac{\eta^2}{k^{1/2}} + \tilde{c}_2\eta^{3/2} + \tilde{c}_3\eta k^{1/2} + \tilde{c}_4\eta^{1/2}k\right]. \tag{17}
$$

Applying Lemma 4, we obtain a bound for uniform stability as shown below:

$$\begin{aligned}\varepsilon_{stab}(\Theta_{\eta t}) &\leq \frac{M(b/m+1)}{\varphi\varepsilon(R\vee 1)}\cdot\frac{1-\tilde{c}_5^t}{1-\tilde{c}_5}\cdot\frac{1}{n}\left[\tilde{c}_1\frac{\eta^2}{k^{1/2}}+\tilde{c}_2\eta^{3/2}+\tilde{c}_3\eta k^{1/2}+\tilde{c}_4\eta^{1/2}k\right]\\ &\leq C_2\frac{1-\tilde{c}_5^t}{1-\tilde{c}_5}\cdot\frac{1}{n}\left[\frac{\eta^2}{k^{1/2}}+\eta^{3/2}+k^{1/2}\eta+\eta^{1/2}k\right],\end{aligned}$$

where

$$C_2 := \frac{M(b/m+1)}{\varphi\widetilde{\varepsilon}(R\vee 1)}(\tilde{c}_1\vee\tilde{c}_2\vee\tilde{c}_3\vee\tilde{c}_4). \tag{18}$$

Here, we define ε to be independent of η by bounding $1/\varepsilon$. Recall the choice of ε from (22), then we have

$$\frac{1}{\varepsilon}\leq\frac{\left(2+2\sigma_4^{1/2}+2\tilde{c}(2)^{1/2}+\frac{4b}{m}+\frac{2\delta\eta_{\max}}{km}(d+2)\ell_f^2\right)}{(1+s)e^{-\alpha\eta/4}-1}\leq\frac{\left(2+2\sigma_4^{1/2}+2\tilde{c}(2)^{1/2}+\frac{4b}{m}+\frac{2\delta\eta_{\max}}{km}(d+2)\ell_f^2\right)}{e^{\alpha\eta_{\max}/4}-1}:=\frac{1}{\widetilde{\varepsilon}}.$$

By approximation $1-\tilde{c}_5^t\leq 1\wedge(1-\tilde{c}_5)t$ and using the bound $(1-e^{-x})^{-1}\leq 1+1/x$ (since $e^x\geq 1+x$),

$$\frac{1-\tilde{c}_5^t}{1-\tilde{c}_5}\leq\left(\frac{1}{1-\tilde{c}_5}\wedge t\right)=\left(\frac{1}{(1-k/n)(1-C_1e^{-\alpha\eta})}\wedge t\right)\leq\left(\frac{1}{(1-k/n)(1-e^{-\alpha\eta/2})}\wedge t\right)\leq\left(\frac{n(1+2/\alpha\eta)}{(n-k)}\wedge t\right),$$

where the second inequality holds because $C_1\leq e^{\alpha\eta/2}$ as determined by the selections of ϕ, a, and ε in (22).

So, if $\eta\leq\min\{\frac{1}{m},\frac{m}{2M^2}\}$ then for any $t\in\mathbb{N}$, the continuous-time algorithm attains the generalization error bound

$$|\mathbb{E}\text{gen}(\Theta_{\eta t})|\leq C_2\min\left\{\eta t,\frac{n(\eta+2/\alpha)}{(n-k)}\right\}\frac{1}{n}\left[\frac{\eta}{k^{1/2}}+\eta^{1/2}+k^{1/2}+\frac{k}{\eta^{1/2}}\right].$$

The result is extended to the discrete-time generalization error bound using the weak triangle inequality (Lemma 15),

$$W_{\rho_g}(\mu_0R_\theta,\mu_0\widehat{R}_\theta)\leq\tilde{c}_6W_{\rho_g}(\mu_0R_\theta,\mu_0R_\Theta)+W_{\rho_g}(\mu_0R_\Theta,\mu_0\widehat{R}_\Theta)+\tilde{c}_6W_{\rho_g}(\mu_0\widehat{R}_\Theta,\mu_0\widehat{R}_\theta), \tag{19}$$

with

$$\tilde{c}_6 := 1+\frac{2g(R)}{\varphi}(\varepsilon R\vee 1). \tag{20}$$

To bound the first and third terms in terms of the 2-Wasserstein distance, we utilize the discretization error bound (Lemma 14).

$$\begin{aligned}W_{\rho_g}(\mu_0R_\theta,\mu_0R_\Theta)^2 &\leq W_2(\mu_0R_\theta,\mu_0R_\Theta)^2(1+2\varepsilon+\varepsilon\mu_0R_\theta(\|\cdot\|^4)^{1/2}+\varepsilon\mu_0R_\Theta(\|\cdot\|^4)^{1/2})\\ &\leq W_2(\mu_0R_\theta,\mu_0R_\Theta)^2(1+2\varepsilon(1+(\sigma_4+\tilde{c}(2))^{1/2})) &&\text{(Lemma 2)}\\ &\leq 8\eta^4\exp(4\eta^2M^2)\left[\frac{2}{3}M^4\left(\sigma_4^{\frac{1}{2}}+\tilde{c}(2)^{\frac{1}{2}}+\frac{b}{m}\right)+\frac{(M^2+2)\delta}{2k}\ell_f^2\right]\cdot\left(1+2\varepsilon\left(1+\sigma_4^{\frac{1}{2}}+\tilde{c}(2)^{\frac{1}{2}}\right)\right)^2, &&\text{(Lemma 6)}\end{aligned}$$

so that

$$\begin{aligned}&W_{\rho_g}(\mu_0R_\theta,\mu_0R_\Theta)\\ &\leq 2\sqrt{2}\eta^2\exp(2\eta^2M^2)\left[\frac{2}{3}M^4(\sigma_4^{1/2}+\tilde{c}(2)^{1/2}+\frac{b}{m})+(M^2+2)\frac{\delta}{2k}\ell_f^2\right]^{1/2}\cdot\left(1+2\varepsilon\left(1+\sigma_4^{\frac{1}{2}}+\tilde{c}(2)^{\frac{1}{2}}\right)\right)\\ &\leq 2\sqrt{2}\eta^2\exp(2\eta^2M^2)\left[\frac{2}{3}M^4\sigma_4^{1/2}+\frac{2}{3}M^4\tilde{c}(2)^{1/2}+\frac{2bM^4}{3m}+\frac{\delta M^2}{2k}\ell_f^2+\frac{\delta}{k}\ell_f^2\right]^{1/2}\cdot\left(1+2\varepsilon\left(1+\sigma_4^{\frac{1}{2}}+\tilde{c}(2)^{\frac{1}{2}}\right)\right)\\ &\leq\ \exp(2\eta^2M^2)\left[\tilde{c}_7\eta^2+\tilde{c}_8\frac{\eta^2}{\sqrt{k}}\right],\end{aligned}$$

with parameters

$$\tilde{c}_7 := 2\sqrt{2}M\left[\frac{2}{3}M^2\sigma_4^{1/2} + \frac{2}{3}M^2\tilde{c}(2)^{1/2} + \frac{2bM^2}{3m}\right]^{1/2} \cdot \left(1 + 2\varepsilon\left(1 + \sigma_4^{\frac{1}{2}} + \tilde{c}(2)^{\frac{1}{2}}\right)\right),$$
$$\tilde{c}_8 := 2\sqrt{\delta}\left(M + \sqrt{2}\right)\ell_f \cdot \left(1 + 2\varepsilon\left(1 + \sigma_4^{\frac{1}{2}} + \tilde{c}(2)^{\frac{1}{2}}\right)\right).$$

Therefore, the inequality (19) can be rewritten as

$$\begin{aligned} W_{\rho_g}(\mu_0 R_\theta, \mu_0 \widehat{R}_\theta) &\le 2\tilde{c}_6 \exp(2\eta^2 M^2)\left[\tilde{c}_7\eta^2 + \tilde{c}_8\frac{\eta^2}{\sqrt{k}}\right] + W_{\rho_g}(\mu_0 R_\Theta, \mu_0 \widehat{R}_\Theta) \\ &\le \tilde{c}_5 W_{\rho_g}(\mu,\nu) + \frac{1}{n}\left[\tilde{c}_1\frac{\eta^2}{k^{1/2}} + \tilde{c}_2\eta^{3/2} + \tilde{c}_3\eta k^{1/2} + \tilde{c}_4\eta^{1/2}k\right] + 2\tilde{c}_6\exp(2\eta^2M^2)\left[\tilde{c}_7\eta^2 + \tilde{c}_8\frac{\eta^2}{\sqrt{k}}\right]. \end{aligned}$$

Now, applying the same arguments as above,

$$\varepsilon_{stab}(x_t) \le C_3 \min\left\{\eta t, \frac{n(\eta + 2/\alpha)}{(n-k)}\right\} \cdot \left[\frac{1}{n}\left\{\frac{\eta}{k^{1/2}} + \eta^{1/2} + k^{1/2} + \frac{k}{\eta^{1/2}}\right\} + \left\{\eta + \frac{\eta}{k^{1/2}}\right\}\right],$$

where

$$C_3 := \frac{M(b/m+1)}{\varphi\widetilde{\varepsilon}(R \vee 1)}(\tilde{c}_1 \vee \tilde{c}_2 \vee \tilde{c}_3 \vee \tilde{c}_4 \vee 2\tilde{c}_6\tilde{c}_7 \vee 2\tilde{c}_6\tilde{c}_8)(1 \vee 2\tilde{c_6}\exp(2\eta_{\max}^2 M^2)). \tag{21}$$

So, if $\eta \le \min\{\frac{1}{m}, \frac{m}{2M^2}\}$ then for any $t \in \mathbb{N}$, the discrete-time algorithm attains the generalization error bound

$$|\mathbb{E}\mathrm{gen}(\theta_t)| \le C_3 \min\left\{\eta t, \frac{n(\eta + 2/\alpha)}{(n-k)}\right\} \cdot \left[\frac{1}{n}\left\{\frac{\eta}{k^{1/2}} + \eta^{1/2} + k^{1/2} + \frac{k}{\eta^{1/2}}\right\} + \left\{\eta + \frac{\eta}{k^{1/2}}\right\}\right],$$

as required. □

8.3.4 Convergence of the Bound

As indicated in Remark 1, to ensure the convergence of our generalization error bound, we need $\eta \ge \frac{1}{\alpha}\ln C_1$, as dictated by the induction process in (17). We can achieve this through an appropriate choice of ε, as outlined below.

By the definition of $\zeta_r(a)$ in the proof of Lemma 3, $\forall q > 1$,

$$a\zeta_r(a) \ge \left[1 - r(\frac{q}{q-1})^{\frac{q-1}{q}}a^{-\frac{1}{q}}\right]^+.$$

Therefore, $\forall r < a$,

$$C_1 \le \frac{1}{\varphi(1-1/a)}\left(1 + \varepsilon\left\{2 + 2\sigma_4^{1/2} + 2\tilde{c}(2)^{1/2} + \frac{4b}{m} + \frac{2\delta\eta}{km}(d+2)\ell_f^2\right\}\right).$$

For any $0 \le \exp(\alpha\eta_{\max}/2) - 1 \le s < 1$, let $\varphi = 1 - s$ and choose a and ε as below:

$$a = \left(1 - \frac{e^{-\frac{\alpha\eta}{2}}}{\varphi(1-s)}\right)^{-1} \text{ and } \quad \varepsilon = \frac{(1+s)e^{-\frac{\alpha\eta}{4}} - 1}{2 + 2\sigma_4^{1/2} + 2\tilde{c}(2)^{1/2} + \frac{4b}{m} + \frac{2\delta\eta_{\max}}{km}(d+2)\ell_f^2}. \tag{22}$$

Then, $C_1 \le (1-s^2)e^{\alpha\eta/4} \le e^{\alpha\eta/4} \le e^{\alpha\eta}$ as required.

9 SGLD Bounds

Below, we provide the bounds outlined in Farghly and Rebeschini [2021], which are subsequently compared with our findings in Table 1.

Lemma 16 (Moment bound [Farghly and Rebeschini, 2021, Lemma A.1])**.**

$$\mu P_t^B(\|\cdot\|^p) \leq \mu(\|\cdot\|^p) + \left[\frac{2b}{m} + 2(p+d-2)/\beta m\right]^{p/2}.$$

Lemma 17 (Moments estimate bound [Farghly and Rebeschini, 2021, Lemma B.2])**.**

$$\mu R_\theta^t(\|\cdot\|^{2p}) \leq \mu(\|\cdot\|^{2p}) + \tilde{c}(p),$$

$$\tilde{c}(p) = \frac{1}{m}\left(\frac{6}{m}\right)^{p-1}\left(1 + \frac{2^{2p}p(2p-1)d}{m\beta}\right)\left[\left(2b + 8\frac{M^2}{m^2}b\right)^p + 1 + 2\left(\frac{d}{\beta}\right)^{p-1}(2p-1)^p\right].$$

Lemma 18 (Divergence bound [Farghly and Rebeschini, 2021, Lemma B.3])**.**

$$\mathbb{E}\left\|\theta_t - \theta_0\right\|^2 \leq 4M^2\left[\mathbb{E}\left\|\theta_0\right\|^2 + \frac{3b + 2d/\beta}{m}\right]t^2 + 4d\beta^{-1}t.$$

Lemma 19 (Discretization error bound [Farghly and Rebeschini, 2021, Lemma B.4])**.**

$$W_2(\mu R_\theta, \mu R_\Theta)^2 \leq 8\eta^3 \exp\left(2\eta^2 M^2\right) M^2(M^2\mu(\|\cdot\|^2) + M^2 b/m + \beta^{-1}d).$$