Journal of Mathematical Modeling
Vol. -, No. -, 2026, pp. -. Research Article

JMM

Uniform stability of noisy gradient dynamics with dataset-dependent diffusion and learning-rate scaling

Fabrizio Morlando[†]

[†] *Independent Researcher, Capua (Ce), Italy*

Email(s): fabriziomorlando84@gmail.com

Abstract. We study the algorithmic stability of the SDE $dw_t = -\nabla \hat{R}_S(w_t)\,dt + \sqrt{\eta}\,\Sigma_S(w_t)\,dB_t$, a continuous time surrogate of SGD whose diffusion depends on the state and on the dataset S. Coupling trajectories on adjacent datasets through a shared Brownian motion, we bound $\mathbb{E}\|w_t - w_t'\|$ by three explicit terms: the contraction rate $\kappa = \mu - \frac{1}{2}\eta L_\Sigma^2$, a gradient mismatch of order $\rho/(n\kappa)$ and a diffusion mismatch of order $\varsigma_n\sqrt{\eta\mu}/\kappa$. If $\mu < 0$, finite-horizon bounds still hold under the remaining assumptions, but they may grow exponentially in time. For the gradient-covariance noise, the sample sensitivity is $\varsigma_n = O(n^{-1/2})$ in general and $O(\sqrt{d}/n)$ under uniform non-degeneracy, which requires $d < n$. The resulting generalization bounds are illustrated on ℓ_2-regularized logistic regression.

Keywords: Algorithmic stability; stochastic gradient descent; generalization bounds.
AMS Subject Classification 2010: 90C31, 60H35.

1 Introduction

Algorithmic stability links the generalization of a learning algorithm to its sensitivity to the replacement of a single training sample [2]. For stochastic gradient descent (SGD), Hardt, Recht and Singer [4] proved uniform-stability bounds for convex and strongly convex losses and, for non-convex losses, for a limited number of iterations. In earlier work [11] we studied the deterministic limit: under the Polyak–Łojasiewicz condition and restricted strong convexity we showed that gradient flow is a contractive dynamical system and derived algorithmic-stability bounds for over-parameterized models. Practical training, however, is stochastic, and the present paper takes the next step: it studies the stability of a *noisy* gradient dynamics whose noise has the structure of the mini-batch gradient noise. A first line of work models SGD by a continuous-time diffusion. Li, Tai and E [7] develop the method of stochastic modified equations (SME), in which SGD-type algorithms are approximated in the weak sense by stochastic differential equations; they also use the continuous-time formulation together with optimal control to derive

*Corresponding author
Received: 09 January 2026/ Revised: 23 November 2026/ Accepted: 24 November 2026
DOI: 10.22124/jmm.2026.35699.3352

© 2026 University of Guilan

Corrected Proof

adaptive learning-rate policies. Mandt, Hoffman and Blei [9] view constant-rate SGD as a stochastic process whose stationary behavior is determined by the gradient noise, and relate it to approximate Bayesian inference. In both cases the diffusion coefficient is state-dependent and proportional to the square root of the learning rate. The validity of such SDE models is itself regime-dependent [8], and the gradient noise of deep networks has been reported to be heavy-tailed rather than Gaussian [15]; we return to this point in Section 4. A second line of work proves generalization bounds for stochastic gradient Langevin dynamics (SGLD), where noise is injected on purpose, in non-convex settings; see Raginsky, Rakhlin and Telgarsky [14] for a non-asymptotic analysis of non-convex learning with SGLD. Mou, Wang, Zhai and Zheng [12] give two non-asymptotic discrete-time analyses of regularized non-convex empirical risk minimization with SGLD, one based on uniform stability and one on PAC-Bayes theory. Their stability bound depends on the Lipschitz parameter, on the inverse temperature and on the sum of the step sizes, so it deteriorates as training proceeds; the PAC-Bayes bound has a slower rate but each step contributes less and it uses the gradient norms along the trajectory. Li, Luo and Qiao [6] introduce the *Bayes-Stability* framework, which combines PAC-Bayesian theory with algorithmic stability, and obtain data-dependent bounds for SGLD and other noisy gradient methods (momentum, mini-batching, acceleration, Entropy-SGD); they recover and often tighten the bounds of [12] and, for ℓ_2-regularized losses, obtain bounds for the continuous Langevin dynamics that remain non-vacuous as the training time grows. Farghly and Rebeschini [3] establish time-independent generalization bounds for SGLD with constant learning rate under dissipativity and smoothness: they work in the framework of uniform stability and exploit the Wasserstein contraction of the Langevin diffusion; their analysis also covers other discretisations, Euclidean projections and non-isotropic noise. In all these works the injected noise is a design choice of the algorithm and its law does not depend on the training sample. In the sequel, we want to carry the stability analysis of [11] from gradient flow to a noisy dynamics with the following three features. Non-isotropic noise is already handled in [3] and data-dependent bounds are given in [6]; the point specific to the present setting is that the law of the noise itself depends on the training sample, and we keep track of this dependence explicitly:

1. the diffusion matrix Σ_S is *state-dependent and depends on the dataset S*, as the gradient covariance of mini-batch SGD does; the mismatch between Σ_S and $\Sigma_{S^{(i)}}$ on adjacent datasets is not assumed away but enters the bound through an explicit sensitivity ς_n;

2. the learning rate η appears explicitly through the scaling $\sqrt{\eta}\,\Sigma_S$ of the stochastic modified equation, so the bound shows how η affects both the contraction rate and the stability plateau;

3. the monotonicity assumption is weakened: the one-sided monotonicity constant μ is allowed to be negative, so that finite-horizon bounds hold without convexity, provided the remaining assumptions on the drift and on the diffusion are satisfied (these bounds may grow exponentially in time when $\kappa < 0$), and global strong monotonicity gives time-uniform bounds. The finite-horizon bound is a Gronwall-type estimate of the same kind as the early-stopping bounds of [4]; its only new element is the diffusion-mismatch term.

For the gradient-covariance diffusion we use an n-dimensional factor of the covariance, for which all the assumptions are verified with explicit, dimension-free constants and no non-degeneracy hypothesis (Proposition 1(i)), and we give a worked example with explicit constants (Remark 4). The technique is elementary (synchronous coupling and Itô's formula for a smoothed norm, which needs only second

Corrected Proof

moments); the point of the paper is the explicit bookkeeping of the two mismatch terms and of the dependence on η. The results are statements about the SDE surrogate (1); the relation with SGD is the heuristic one of the SME method and no discretisation error is estimated. The time-uniform bounds need global strong monotonicity, which is restrictive and is not claimed to hold for over-parameterized networks; for those, only the finite bounds of Corollary 3 apply.

2 Setting and assumptions

Let $\mathscr{Z}$ be a measurable space, $\mathscr{D}$ a distribution on $\mathscr{Z}$, and $S=(z_1,\dots,z_n)\in\mathscr{Z}^n$ a dataset. For parameters $w\in\mathbb{R}^d$ the loss is $l(w,z)=l_0(w,z)+\frac{\mu_0}{2}\|w\|^2$ with $\mu_0\geq 0$; $\hat{R}_S(w)=\frac{1}{n}\sum_{j=1}^n l(w,z_j)$ is the empirical risk and $R(w)=\mathbb{E}_{z\sim\mathscr{D}}\,l(w,z)$ the population risk. Two datasets are *adjacent*, S and $S^{(i)}$, if they differ only in the i-th sample (z_i replaced by z_i').
For a learning rate $\eta>0$ we consider

$$dw_t=-\nabla\hat{R}_S(w_t)\,dt+\sqrt{\eta}\,\Sigma_S(w_t)\,dB_t, \tag{1}$$

where B is an m-dimensional Brownian motion and $\Sigma_S:\mathbb{R}^d\to\mathbb{R}^{d\times m}$ is the diffusion matrix ($m=d$ for a square factor, $m=n$ for the factor of Proposition 1(i)). If SGD $w_{k+1}=w_k-\eta g_k$ uses an unbiased stochastic gradient g_k with $\mathbb{E}g_k=\nabla\hat{R}_S(w_k)$ and covariance $\Sigma_S(w_k)\Sigma_S(w_k)^\top$, then the first-order stochastic modified equation of [7] is (1), with $t=k\eta$: after K steps the horizon is $T=\eta K$. We use (1) as the model; we do not prove that the SGD iterates inherit the bounds below.

2.1 Assumptions

Assumption 1.

(A1) ***Monotonicity (sign-free).*** *For every $S\in\mathscr{Z}^n$, $\hat{R}_S$ is C^1 with globally Lipschitz gradient and, for some $\mu\in\mathbb{R}$ (the same for all S), $\langle w-w',\nabla\hat{R}_S(w)-\nabla\hat{R}_S(w')\rangle\geq\mu\|w-w'\|^2$ for all w,w'.*

(A2) ***Lipschitz data-fit loss.*** *$l_0(\cdot,z)$ is C^1, $\|\nabla_w l_0(w,z)\|\leq\rho$ for all w,z, and $\mathbb{E}_{z\sim\mathscr{D}}|l_0(0,z)|<\infty$. Hence $|l_0(w,z)-l_0(w',z)|\leq\rho\|w-w'\|$.*

(A3) ***Regular diffusion.*** *Σ_S is measurable and $\|\Sigma_S(w)-\Sigma_S(w')\|_F\leq L_\Sigma\|w-w'\|$ for all w,w' and all $S\in\mathscr{Z}^n$.*

(A4) ***Sample sensitivity of the diffusion.*** *There is $\varsigma_n\geq 0$ such that $\sup_w\|\Sigma_S(w)-\Sigma_{S^{(i)}}(w)\|_F\leq\varsigma_n$ for all adjacent $S,S^{(i)}$.*

We set

$$\kappa:=\mu-\tfrac{1}{2}\eta L_\Sigma^2\in\mathbb{R},\qquad \mu_+:=\max(\mu,0),\qquad \varphi_\kappa(t):=\int_0^t e^{-\kappa u}\,du=\begin{cases}\frac{1-e^{-\kappa t}}{\kappa}, & \kappa\neq 0,\\ t, & \kappa=0.\end{cases}$$

Note that φ_κ is increasing in t, and $\varphi_\kappa(t)\leq 1/\kappa$ if $\kappa>0$.

Corrected Proof

Remark 1 (On (A1))**.** *The Lipschitz gradient already implies (A1) with $\mu = -\ell$, ℓ being the Lipschitz constant, so (A1) is not a restriction when $\mu \le 0$ is allowed; $\mu > 0$ is strong monotonicity (e.g. strong convexity). Because $\hat{R}_S - \hat{R}_{S^{(i)}} = \frac{1}{n}\big(l_0(\cdot, z_i) - l_0(\cdot, z_i')\big)$, the regularizer $\frac{\mu_0}{2}\|w\|^2$ cancels and*

$$\sup_w \|\nabla \hat{R}_S(w) - \nabla \hat{R}_{S^{(i)}}(w)\| \le \frac{2\rho}{n}. \tag{2}$$

This is why ρ bounds only the data-fit part l_0: a strongly monotone function has unbounded gradient on $\mathbb{R}^d$. In fact, under (A2) the data-fit part cannot contribute any positive monotonicity: taking $w' = w + sv$ with $\|v\| = 1$ in (A1) gives $\mu s^2 \le \mu_0 s^2 + 2\rho s$ for all $s > 0$, hence $\mu \le \mu_0$. Thus $\mu > 0$ requires the explicit regularizer, and the time-uniform bounds of Corollary 2 cover essentially ℓ_2-regularized problems with a Lipschitz data-fit loss, such as regularized logistic regression (Remark 4). Assumption (A2) excludes the squared loss and any loss with quadratic growth, as well as networks with unbounded activations or unbounded weights.

Definition 1 (Uniform stochastic stability)**.** *Let w_t, w_t' solve (1) on adjacent datasets $S, S^{(i)}$, driven by the* same *Brownian motion and started from the same deterministic point w_0. We say that (1) is β_n-uniformly stable if, for all adjacent pairs and all $t \ge 0$,*

$$\mathbb{E}\Big[\sup_z |l_0(w_t, z) - l_0(w_t', z)|\Big] \le \beta_n(t).$$

By (A2), $\sup_z |l_0(w_t, z) - l_0(w_t', z)| \le \rho \|w_t - w_t'\|$, so it suffices to bound $\mathbb{E}\|w_t - w_t'\|$.

Assumption (A4) is a property of the model for Σ_S, and it is the place where the dataset dependence of the noise enters. The next proposition shows the sensitivity of the gradient-covariance diffusion for two choices of the factor.

Proposition 1. *Let $g_j(w) = \nabla_w l_0(w, z_j)$, $\bar{g} = \frac{1}{n}\sum_j g_j$ and $C_S(w) = \frac{1}{n}\sum_j g_j g_j^\top - \bar{g}\bar{g}^\top = \frac{1}{n}\sum_j (g_j - \bar{g})(g_j - \bar{g})^\top$, so that $C_S(w)/b$ is the covariance of the average of b gradients drawn with replacement. Assume (A2).*

(i) (*n*-dimensional factor, no non-degeneracy.) *Let*

$$\Sigma_S(w) = (bn)^{-1/2}\big[g_1(w) - \bar{g}(w), \dots, g_n(w) - \bar{g}(w)\big] \in \mathbb{R}^{d \times n},$$

with an n-dimensional Brownian motion, so that $\Sigma_S \Sigma_S^\top = C_S/b$. Then $\|\Sigma_S(w)\|_F \le \rho/\sqrt{b}$ for all S, w, and (A4) holds with

$$\varsigma_n \le 2\rho\sqrt{\frac{1 + 3/n}{bn}} \le \frac{4\rho}{\sqrt{bn}}.$$

If moreover $\nabla_w l_0(\cdot, z)$ is ℓ_0-Lipschitz for all z, (A3) holds with $L_\Sigma \le \ell_0/\sqrt{b}$. These constants do not depend on d.

(ii) (Symmetric square root under uniform non-degeneracy.) *Let $\Sigma_S(w) = b^{-1/2} C_S(w)^{1/2} \in \mathbb{R}^{d \times d}$. Then $\|C_S(w) - C_{S^{(i)}}(w)\|_1 \le 10\rho^2/n$ (trace norm). If moreover $C_S(w) \succeq \lambda I$ for all S, w with $\lambda > 0$, then*

$$\varsigma_n \le \frac{5\rho^2}{\sqrt{\lambda b}\, n},$$

Corrected Proof

and if $\nabla_w l_0(\cdot,z)$ is ℓ_0-Lipschitz for all z, (A3) holds with $L_\Sigma \le 2\rho\ell_0/\sqrt{b\lambda}$. This hypothesis forces $d \le n-1$ and $\lambda \le \rho^2/d$; consequently $\frac{5\rho^2}{\sqrt{\lambda bn}} \ge \frac{5\rho\sqrt{d}}{\sqrt{bn}}$, and the sufficient condition $\eta < 2\mu/L_\Sigma^2$ for $\kappa > 0$ reads $\eta < \mu b\lambda/(2\rho^2\ell_0^2) \le \mu b/(2d\ell_0^2)$.

Proof. (i) $\Sigma_S\Sigma_S^\top = \frac{1}{bn}\sum_j (g_j-\bar g)(g_j-\bar g)^\top = C_S/b$, and $\|\Sigma_S\|_F^2 = \operatorname{tr} C_S/b \le \frac{1}{bn}\sum_j \|g_j\|^2 \le \rho^2/b$. For the sensitivity, replacing z_i by z_i' moves the mean to $\bar g + h$ with $h = (g_i' - g_i)/n$, $\|h\| \le 2\rho/n$. Column $j \ne i$ of $\sqrt{bn}\,(\Sigma_S - \Sigma_{S^{(i)}})$ equals h, and column i equals $g_i - g_i' + h$, of norm at most $2\rho(1+1/n)$. Hence

$$bn\,\|\Sigma_S - \Sigma_{S^{(i)}}\|_F^2 \le (n-1)\frac{4\rho^2}{n^2} + 4\rho^2\Big(1+\frac{1}{n}\Big)^2 = 4\rho^2\Big(1+\frac{3}{n}\Big),$$

which gives the bound on ς_n. For (A3), put $u_j = g_j(w) - g_j(w')$, $\|u_j\| \le \ell_0\|w-w'\|$, and $\bar u = \frac{1}{n}\sum_j u_j$; then $bn\,\|\Sigma_S(w) - \Sigma_S(w')\|_F^2 = \sum_j \|u_j - \bar u\|^2 \le \sum_j \|u_j\|^2 \le n\ell_0^2\|w-w'\|^2$.

(ii) Replacing g_i by g_i' changes $\frac{1}{n}\sum g_j g_j^\top$ by $\frac{1}{n}(g_i g_i^\top - g_i' g_i'^\top)$, whose trace norm is at most $2\rho^2/n$. With h as above, $\bar g'\bar g'^\top - \bar g\bar g^\top = \bar g h^\top + h\bar g^\top + hh^\top$ has trace norm at most $4\rho^2/n + 4\rho^2/n^2$. Adding up gives $10\rho^2/n$. Put $X = A^{1/2} - B^{1/2}$ with $A, B \succeq \lambda I$. Then $A^{1/2}X + XB^{1/2} = A - B$, and since $A^{1/2}, B^{1/2} \succeq \sqrt{\lambda} I$ the solution of this Sylvester equation satisfies $\|X\|_F \le \|A-B\|_F/(2\sqrt{\lambda})$ (see [1, Ch. VII]). With $\|\cdot\|_F \le \|\cdot\|_1$ and the trace-norm bound we get $\varsigma_n \le b^{-1/2}\cdot\frac{10\rho^2}{n}\cdot\frac{1}{2\sqrt{\lambda}} = \frac{5\rho^2}{\sqrt{\lambda bn}}$. The same estimate applied to $C_S(w), C_S(w')$, with $\|gg^\top - \tilde g\tilde g^\top\|_F \le 2\rho\|g-\tilde g\|$, gives $\|C_S(w) - C_S(w')\|_F \le 4\rho\ell_0\|w-w'\|$, hence the value of L_Σ. Finally, the vectors $g_j - \bar g$ sum to zero, so $\operatorname{rank} C_S \le n-1$ and $C_S \succeq \lambda I$ with $\lambda > 0$ forces $d \le n-1$; and $\lambda d \le \operatorname{tr} C_S \le \rho^2$. The last two claims follow by inserting $\lambda \le \rho^2/d$. □

Thus the factor (i) satisfies (A3)–(A4) with explicit, dimension-free constants and gives $\varsigma_n = O(n^{-1/2})$ with no non-degeneracy assumption (Proposition 1(i)). The symmetric root gives $\varsigma_n = O(n^{-1})$ only under uniform non-degeneracy (Proposition 1(ii)), which requires $d < n$ and hides a factor at least $\sqrt{d}$ in the constant. The bound of (ii) improves on that of (i) only if $\frac{5\rho^2}{\sqrt{\lambda bn}} < 2\rho\sqrt{\frac{1+3/n}{bn}}$, which, since $\lambda \le \rho^2/d$, is possible only if $d < \frac{4}{25}(n+3)$. Without non-degeneracy the Powers–Størmer inequality $\|A^{1/2} - B^{1/2}\|_F^2 \le \|A-B\|_1$ for $A, B \succeq 0$ [13] still gives $\varsigma_n \le \rho\sqrt{10/(bn)}$ for the symmetric root, but the root is only $\frac{1}{2}$-Hölder in the covariance and (A3) may fail; this is why the factor (i) is used for the general case. In the following, we state the main result of the paper.

Theorem 1 (Stability of the coupled dynamics)**.** *Let $S, S' \in \mathscr{Z}^n$ be two datasets, with (A1) and (A3) holding for both. Set*

$$\delta := \sup_w \|\nabla\hat R_S(w) - \nabla\hat R_{S'}(w)\|, \qquad \varsigma := \sup_w \|\Sigma_S(w) - \Sigma_{S'}(w)\|_F,$$

and assume $\delta, \varsigma < \infty$. Let w_t, w_t' solve (1) for S, S', driven by the same Brownian motion B, with $\mathscr{F}_0$-measurable initial data $w_0, w_0' \in L^2$. Then for all $t \ge 0$

$$\mathbb{E}\|w_t - w_t'\| \le e^{-\kappa t}\,\mathbb{E}\|w_0 - w_0'\| + \varphi_\kappa(t)\big(\delta + \eta L_\Sigma\varsigma\big) + \varsigma\sqrt{2\eta\,\varphi_\kappa(t)\big(e^{-\kappa t} + \mu_+\varphi_\kappa(t)\big)}. \tag{3}$$

More precisely, for every $\varepsilon > 0$,

$$\mathbb{E}\|w_t - w_t'\| \le e^{-\kappa t}\big(\mathbb{E}\|w_0 - w_0'\| + \varepsilon\big) + \varphi_\kappa(t)\Big(\delta + \eta L_\Sigma\varsigma + \mu_+\varepsilon + \frac{\eta\varsigma^2}{2\varepsilon}\Big), \tag{4}$$

and (3) is (4) minimized over ε (for $\varsigma = 0$, let $\varepsilon \downarrow 0$).

Corrected Proof

Proof. We split the proof into five parts.
Step 0: well-posedness. The drifts are globally Lipschitz by (A1) and the diffusions are Lipschitz by (A3); hence strong solutions exist, are unique, and $\mathbb{E}\sup_{s\le t}\|w_s\|^2<\infty$ for every t when $w_0\in L^2$ [10, Ch. 2].
Step 1: Itô's formula for a smoothed norm. Let $\Delta_t=w_t-w'_t$, $a_t=-(\nabla\hat R_S(w_t)-\nabla\hat R_{S'}(w'_t))$ and $D_t=\Sigma_S(w_t)-\Sigma_{S'}(w'_t)$, so that $d\Delta_t=a_t\,dt+\sqrt{\eta}\,D_t\,dB_t$. For $\varepsilon>0$ let $f(x)=\sqrt{\|x\|^2+\varepsilon^2}$, which is C^2 with $\nabla f=x/f$ and $\nabla^2 f=I/f-xx^\top/f^3\preceq I/f$. Itô's formula gives

$$df(\Delta_t)=\langle\tfrac{\Delta_t}{f(\Delta_t)},a_t\rangle dt+\tfrac{\eta}{2}tr\big(\nabla^2 f(\Delta_t)D_tD_t^\top\big)dt+dM_t,\qquad dM_t=\sqrt{\eta}\,\langle\tfrac{\Delta_t}{f(\Delta_t)},D_t\,dB_t\rangle.$$

Step 2: the martingale. Since $\|D_t^\top\Delta_t\|/f\le\|D_t\|_F$ and $\|D_t\|_F\le L_\Sigma\|\Delta_t\|+\varsigma$ (see below), the quadratic variation of M satisfies $\mathbb{E}\langle M\rangle_t\le\eta\,\mathbb{E}\int_0^t\|D_s\|_F^2ds\le 2\eta\int_0^t\big(L_\Sigma^2\mathbb{E}\|\Delta_s\|^2+\varsigma^2\big)ds<\infty$ by Step 0. Hence M is a true martingale; only second moments are needed.

Step 3: the drift. Write $\nabla\hat R_S(w_t)-\nabla\hat R_{S'}(w'_t)=[\nabla\hat R_S(w_t)-\nabla\hat R_S(w'_t)]+[\nabla\hat R_S(w'_t)-\nabla\hat R_{S'}(w'_t)]$. By (A1) and $\|\Delta\|/f\le 1$,

$$\langle\tfrac{\Delta}{f},a\rangle\le-\mu\frac{\|\Delta\|^2}{f}+\delta\frac{\|\Delta\|}{f}\le-\mu f+\mu_+\varepsilon+\delta,$$

where we used $\|\Delta\|^2/f=f-\varepsilon^2/f\ge f-\varepsilon$ (if $\mu\le 0$ the bound $-\mu\|\Delta\|^2/f\le-\mu f$ is used instead). For the diffusion, $D=[\Sigma_S(w)-\Sigma_S(w')]+[\Sigma_S(w')-\Sigma_{S'}(w')]$ gives $\|D\|_F\le L_\Sigma\|\Delta\|+\varsigma\le L_\Sigma f+\varsigma$, and since $f\ge\varepsilon$,

$$\tfrac{\eta}{2}tr(\nabla^2 f\,DD^\top)\le\frac{\eta}{2f}\|D\|_F^2\le\frac{\eta}{2}\Big(L_\Sigma^2 f+2L_\Sigma\varsigma+\frac{\varsigma^2}{f}\Big)\le\frac{\eta L_\Sigma^2}{2}f+\eta L_\Sigma\varsigma+\frac{\eta\varsigma^2}{2\varepsilon}.$$

Altogether, with $\kappa=\mu-\frac{1}{2}\eta L_\Sigma^2$ and $K_\varepsilon:=\delta+\eta L_\Sigma\varsigma+\mu_+\varepsilon+\eta\varsigma^2/(2\varepsilon)$,

$$df(\Delta_t)\le\big(-\kappa f(\Delta_t)+K_\varepsilon\big)dt+dM_t.$$

Step 4: integration. By the product rule, $d\big(e^{\kappa t}f(\Delta_t)\big)\le K_\varepsilon e^{\kappa t}dt+e^{\kappa t}dM_t$. Taking expectations (Step 2) and multiplying by $e^{-\kappa t}$,

$$\mathbb{E}f(\Delta_t)\le e^{-\kappa t}\mathbb{E}f(\Delta_0)+K_\varepsilon\int_0^t e^{-\kappa(t-s)}ds=e^{-\kappa t}\mathbb{E}f(\Delta_0)+K_\varepsilon\varphi_\kappa(t).$$

Since $\|x\|\le f(x)\le\|x\|+\varepsilon$, this is (4). Finally, the ε-dependent part of the right-hand side is $\varepsilon\alpha+\beta/\varepsilon$ with $\alpha=e^{-\kappa t}+\mu_+\varphi_\kappa(t)$ and $\beta=\eta\varsigma^2\varphi_\kappa(t)/2$, minimized at $\varepsilon=\sqrt{\beta/\alpha}$ with value $2\sqrt{\alpha\beta}=\varsigma\sqrt{2\eta\,\varphi_\kappa(t)\alpha}$, which is (3). □

Remark 2 (Special cases and sharpness)**.**

1. Same dataset *($S'=S$, so $\delta=\varsigma=0$):* $\mathbb{E}\|w_t-w'_t\|\le e^{-\kappa t}\mathbb{E}\|w_0-w'_0\|$. *For $\kappa>0$ the flow is a W_1-contraction: choose (w_0,w'_0) as an optimal coupling of two initial laws, independent of B.*
2. Dataset-independent diffusion *($\varsigma=0$):* $\mathbb{E}\|w_t-w'_t\|\le e^{-\kappa t}\mathbb{E}\|w_0-w'_0\|+\delta\varphi_\kappa(t)$*; for $\eta\to 0$ this is the gradient-flow bound with $\kappa\to\mu$.*
3. Order of the diffusion-mismatch term. *For $\delta=L_\Sigma=0$ and $\nabla\hat R_S=\mu\,\mathrm{id}$, Δ is an Ornstein–Uhlenbeck process with stationary $\mathbb{E}\|\Delta\|^2=\eta\varsigma^2/(2\mu)$, i.e. $\mathbb{E}\|\Delta\|\lesssim\varsigma\sqrt{\eta/(2\mu)}$; the plateau of (3) is $\varsigma\sqrt{2\eta\mu}/\mu=2\varsigma\sqrt{\eta/(2\mu)}$, so, since $\mathbb{E}\|\Delta\|\le\sqrt{\mathbb{E}\|\Delta\|^2}$ and $\mathbb{E}\|\Delta\|\ge\sqrt{2/\pi}\sqrt{\mathbb{E}\|\Delta\|^2}$ in dimension one, it is sharp up to a factor of at most about 2.5 in this case.*

Corrected Proof

4. Sharpness of the loss $\frac{1}{2}\eta L_\Sigma^2$ in κ. *In dimension $d \geq 2$ this loss is attained. Let $d = 2$, $m = 1$, $\nabla \hat{R}_S(w) = \mu w$ and $\Sigma_S(w) = L_\Sigma J w$ for every S, with J the rotation by $\pi/2$; then $\delta = \varsigma = 0$, (A3) holds with constant L_Σ, and the noise acting on Δ_t is orthogonal to Δ_t. Itô's formula for $\|\cdot\|$ (valid since $\Delta_t \neq 0$ for all t when $\Delta_0 \neq 0$) gives $d\|\Delta_t\| = -(\mu - \frac{1}{2}\eta L_\Sigma^2)\|\Delta_t\|\,dt$, with no martingale part, so $\|\Delta_t\| = e^{-\kappa t}\|\Delta_0\|$ exactly. In dimension one the loss is not present: for $d\Delta_t = -\mu\Delta_t\,dt + \sqrt{\eta}L_\Sigma\Delta_t\,dB_t$ one has $\mathbb{E}|\Delta_t| = e^{-\mu t}|\Delta_0|$.*

3 Consequences for stability and generalization

For adjacent datasets, (A2), (2) and (A4) give $\delta \leq 2\rho/n$ and $\varsigma \leq \varsigma_n$, and the right-hand side of (3) is nondecreasing in δ and ς. Define, for $w_0 = w_0'$,

$$\Lambda_n(t) := \varphi_\kappa(t)\Big(\frac{2\rho}{n} + \eta L_\Sigma \varsigma_n\Big) + \varsigma_n\sqrt{2\eta\,\varphi_\kappa(t)\big(e^{-\kappa t} + \mu_+\varphi_\kappa(t)\big)}.$$

Corollary 1 (Uniform stability)**.** *Under Assumption 1, (1) is β_n-uniformly stable with $\beta_n(t) \leq \rho\,\Lambda_n(t)$ for all $t \geq 0$.*

Corollary 2 (Time-uniform bound, $\kappa > 0$)**.** *If $\kappa > 0$, i.e. $\eta < 2\mu/L_\Sigma^2$ (so that $\mu > 0$), then for all $t \geq 0$*

$$\beta_n(t) \leq \frac{\rho}{\kappa}\Big(\frac{2\rho}{n} + \varsigma_n\big(\eta L_\Sigma + \sqrt{\tfrac{5}{2}\eta\mu}\big)\Big),$$

and $\limsup_{t\to\infty}\beta_n(t) \leq \frac{\rho}{\kappa}\big(\frac{2\rho}{n} + \varsigma_n(\eta L_\Sigma + \sqrt{2\eta\mu})\big)$.

Proof. $\varphi_\kappa \leq 1/\kappa$. With $x = e^{-\kappa t} \in (0,1]$, $\varphi_\kappa e^{-\kappa t} = x(1-x)/\kappa \leq 1/(4\kappa) \leq \mu/(4\kappa^2)$ because $\kappa \leq \mu$, and $\mu\varphi_\kappa^2 \leq \mu/\kappa^2$. Hence $\varphi_\kappa(e^{-\kappa t} + \mu\varphi_\kappa) \leq \frac{5}{4}\mu/\kappa^2$ and $\sqrt{2\eta\cdot\frac{5}{4}\mu/\kappa^2} = \sqrt{\frac{5}{2}\eta\mu}/\kappa$. For the limit, $e^{-\kappa t} \to 0$ and $\varphi_\kappa \to 1/\kappa$. □

Corollary 3 (Finite horizon, no strong monotonicity)**.** *Let $\kappa \in \mathbb{R}$ be arbitrary and $\kappa_- = \max(-\kappa, 0)$. For every $T \geq 0$,*

$$\beta_n(T) \leq \rho\,e^{\kappa_- T}\Big[\Big(\frac{2\rho}{n} + \eta L_\Sigma\varsigma_n\Big)T + \varsigma_n\sqrt{2\eta T(1+\mu_+ T)}\Big],$$

where the factor T multiplies only the gradient-mismatch term $\frac{2\rho}{n} + \eta L_\Sigma\varsigma_n$.

Proof. Since $e^{-\kappa u} \leq e^{\kappa_- T}$ for $u \in [0,T]$, we have $\varphi_\kappa(T) \leq Te^{\kappa_- T}$; moreover $e^{-\kappa T} \leq e^{\kappa_- T}$. Hence

$$\varphi_\kappa(T)\big(e^{-\kappa T} + \mu_+\varphi_\kappa(T)\big) \leq Te^{\kappa_- T}\big(e^{\kappa_- T} + \mu_+ Te^{\kappa_- T}\big) = Te^{2\kappa_- T}(1+\mu_+ T),$$

so that

$$\varphi_\kappa(T)\Big(\frac{2\rho}{n} + \eta L_\Sigma\varsigma_n\Big) \leq e^{\kappa_- T}\,T\Big(\frac{2\rho}{n} + \eta L_\Sigma\varsigma_n\Big),$$
$$\varsigma_n\sqrt{2\eta\,\varphi_\kappa(T)\big(e^{-\kappa T} + \mu_+\varphi_\kappa(T)\big)} \leq e^{\kappa_- T}\,\varsigma_n\sqrt{2\eta T(1+\mu_+ T)}.$$

Inserting these two inequalities in $\beta_n(T) \leq \rho\,\Lambda_n(T)$ (Corollary 1) gives the claim, for every $T \geq 0$. □

Corrected Proof

If $\kappa \geq 0$ the bound is polynomial in T; if $\kappa < 0$ it grows like $e^{|\kappa|T}$ and is informative only for $T \lesssim \log n/|\kappa|$, in the spirit of the early-stopping bounds for non-convex losses of [4]. Corollary 3 is a Gronwall-type estimate: the factor $e^{|\kappa|T}$ and the horizon $T \lesssim \log n/|\kappa|$ are those of [4], and its only new element is the diffusion-mismatch term $\varsigma_n\sqrt{2\eta T(1+\mu_+ T)}$.

Corollary 4 (Expected generalization gap)**.** *Let $S \sim \mathscr{D}^n$ be i.i.d., B independent of S, and w_T the solution of (1) on S from a deterministic w_0. Assume Assumption 1 and, in addition, that the Lipschitz constant ℓ of $\nabla\hat{R}_S$ and $\|\Sigma_S(0)\|_F$ are bounded uniformly in $S \in \mathscr{Z}^n$ (both hold for the factor of Proposition 1(i) when $\nabla_w l_0(\cdot,z)$ is ℓ_0-Lipschitz, with $\ell \leq \ell_0 + \mu_0$ and $\|\Sigma_S(0)\|_F \leq \rho/\sqrt{b}$). Then*

$$\left|\mathbb{E}_{S,B}\left[R(w_T) - \hat{R}_S(w_T)\right]\right| \leq \rho\,\Lambda_n(T).$$

Proof. The term $\frac{\mu_0}{2}\|w\|^2$ does not depend on z and cancels, so l may be replaced by l_0. All expectations are finite: $\|\nabla\hat{R}_S(w)\| \leq \rho + \ell\|w\|$ (since $\nabla\hat{R}_S(0) = \frac{1}{n}\sum_j \nabla_w l_0(0,z_j)$) and $\|\Sigma_S(w)\|_F \leq \|\Sigma_S(0)\|_F + L_\Sigma\|w\|$, with constants independent of S, so the standard moment estimate [10, Ch. 2] gives $\sup_S \mathbb{E}\sup_{t\leq T}\|w_t\|^2 < \infty$; together with $|l_0(w,z)| \leq |l_0(0,z)| + \rho\|w\|$ and (A2) this yields integrability over (S,B). Let $z'_1,\dots,z'_n$ be an independent copy of the sample, and let $w_T^{(i)}$ be the solution on $S^{(i)}$ driven by the same B from the same w_0. It is a function of $(S^{(i)},B)$, hence independent of z_i, so $\mathbb{E}R(w_T) = \mathbb{E}\,l_0(w_T,z'_i)$; by exchangeability of $(S,z'_i) \leftrightarrow (S^{(i)},z_i)$, $\mathbb{E}\,l_0(w_T,z'_i) = \mathbb{E}\,l_0(w_T^{(i)},z_i)$. Therefore $\mathbb{E}[R(w_T) - \hat{R}_S(w_T)] = \frac{1}{n}\sum_i \mathbb{E}[l_0(w_T^{(i)},z_i) - l_0(w_T,z_i)]$ and each term is bounded in absolute value by $\mathbb{E}\sup_z |l_0(w_T^{(i)},z) - l_0(w_T,z)| \leq \beta_n(T)$. □

Remark 3 (The generalization gap does not depend on the choice of the factor)**.** *The stability quantity $\mathbb{E}\|w_t - w'_t\|$ depends on the factor Σ_S through the coupling, but the generalization gap $\mathbb{E}[R(w_T) - \hat{R}_S(w_T)]$ depends only on the conditional law of w_T given S. If Σ_S is Lipschitz and $\tilde{\Sigma}_S$ is any measurable factor with $\tilde{\Sigma}_S\tilde{\Sigma}_S^\top = \Sigma_S\Sigma_S^\top$, every weak solution of either SDE solves the martingale problem for the generator $\mathscr{L}\phi = -\langle\nabla\hat{R}_S, \nabla\phi\rangle + \frac{\eta}{2}\operatorname{tr}(\Sigma_S\Sigma_S^\top\nabla^2\phi)$, and every solution of this martingale problem is a weak solution of the SDE with the Lipschitz factor Σ_S; pathwise uniqueness for the latter then gives uniqueness in law [5, Sections 5.3–5.4]. Hence Corollary 4 holds for the gradient-covariance diffusion with Λ_n computed from any Lipschitz factor of C_S/b. In particular the bound obtained from Proposition 1(i) also applies to the model with the symmetric root $b^{-1/2}C_S^{1/2}$, without any non-degeneracy assumption, and under the hypotheses of Proposition 1(ii) one may take the smaller of the two bounds.*

Corollary 5 (Sample complexity)**.** *Let $\kappa > 0$, $\varsigma_n \leq c_\Sigma n^{-\gamma}$ with $\gamma \in (0,1]$, and $c_\eta := \eta L_\Sigma + \sqrt{\frac{5}{2}\eta\mu}$. Then for every $T \geq 0$ the generalization gap of Corollary 4 is at most ε as soon as*

$$n \geq \max\left\{\frac{4\rho^2}{\kappa\varepsilon}, \left(\frac{2\rho\, c_\Sigma c_\eta}{\kappa\varepsilon}\right)^{1/\gamma}\right\}.$$

For $\gamma = \frac{1}{2}$ (general gradient-covariance diffusion with the factor of Proposition 1(i), where $c_\Sigma = 4\rho/\sqrt{b}$ and $L_\Sigma \leq \ell_0/\sqrt{b}$, so that all assumptions are verified and no constant depends on d) the diffusion mismatch dominates and $n = O(1/\varepsilon^2)$ is needed. For $\gamma = 1$ (symmetric root under uniform non-degeneracy, Proposition 1(ii), where $c_\Sigma = 5\rho^2/\sqrt{\lambda b} \geq 5\rho\sqrt{d/b}$) one gets $n = O(\sqrt{d}/\varepsilon)$; this regime requires $d < n$ and the step-size restriction $\eta < \mu b\lambda/(2\rho^2\ell_0^2) \leq \mu b/(2d\ell_0^2)$.

Proof. By Corollary 2 the gap is at most $\frac{2\rho^2}{\kappa n} + \frac{\rho c_\Sigma c_\eta}{\kappa} n^{-\gamma}$; require each term to be $\leq \varepsilon/2$. □

Corrected Proof

Remark 4 (Worked example: ℓ_2-regularized logistic regression)**.** *Let $z=(x,y)$ with $\|x\|\le X$ almost surely and $y\in\{-1,1\}$, $l_0(w,z)=\log\bigl(1+e^{-y\langle x,w\rangle}\bigr)$ and $\mu_0>0$. Then $\|\nabla_w l_0\|\le X$ and $0\preceq\nabla_w^2 l_0\preceq\frac{X^2}{4}I$, so (A2) holds with $\rho=X$ (and $\mathbb{E}|l_0(0,z)|=\log 2$), $\nabla_w l_0(\cdot,z)$ is ℓ_0-Lipschitz with $\ell_0=X^2/4$, and (A1) holds with $\mu=\mu_0$ and Lipschitz constant $X^2/4+\mu_0$, uniformly in S. With the factor of Proposition 1(i),*

$$L_\Sigma\le\frac{X^2}{4\sqrt{b}},\qquad \varsigma_n\le 2X\sqrt{\frac{1+3/n}{bn}},\qquad \|\Sigma_S(0)\|_F\le\frac{X}{\sqrt{b}},$$

so all the assumptions of Corollaries 2 and 4 hold, with no dependence on d. Now $\kappa=\mu_0-\eta X^4/(32b)>0$ if and only if $\eta<32b\mu_0/X^4$. Take $\eta=16\theta b\mu_0/X^4$ with $\theta\in(0,1]$; then $\kappa=\mu_0(1-\theta/2)\ge\mu_0/2$, $\eta L_\Sigma\le 4\theta\sqrt{b}\,\mu_0/X^2$ and $\sqrt{\frac{5}{2}\eta\mu_0}=2\sqrt{10\theta}\,\sqrt{b}\,\mu_0/X^2$. Corollary 2 gives, for all $t\ge 0$,

$$\beta_n(t)\le\frac{4X^2}{\mu_0 n}+\frac{4\bigl(4\theta+2\sqrt{10\theta}\bigr)\sqrt{1+3/n}}{\sqrt{n}}\le\frac{4X^2}{\mu_0 n}+\frac{59\sqrt{\theta}}{\sqrt{n}}\qquad(n\ge 3),$$

and the same bound holds for the generalization gap of Corollary 4. In particular the gap is at most ε for every T as soon as $n\ge\max\{8X^2/(\mu_0\varepsilon),\,13924\,\theta/\varepsilon^2\}$. The first term has the same form as the bound $2L^2/(\gamma n)$ of [4] for SGD with $\gamma=\mu_0$, up to a factor 2 (there L is the Lipschitz constant of the full regularized loss on the bounded domain of projected SGD, whereas here only the data-fit part enters, with $\rho=X$). The second term is independent of d and b (at fixed θ, since η is proportional to b) and scales as $\sqrt{\eta/n}$. For large n it dominates the first one, and it has no counterpart in the discrete-SGD bound (see Section 4).

4 Discussion and limitations

The learning rate enters in three places. (i) The contraction rate $\kappa=\mu-\frac{1}{2}\eta L_\Sigma^2$ is positive iff $\eta<2\mu/L_\Sigma^2$: a large learning rate with a state-dependent noise destroys the time-uniform guarantee. (ii) The gradient-mismatch term $2\rho/(n\kappa)$ depends on η only through κ. (iii) The diffusion-mismatch term scales as $\sqrt{\eta}$ and is multiplied by the sensitivity ς_n of the noise to one sample: it vanishes for $\eta\to 0$ (the deterministic gradient flow) and for a dataset-independent noise. In terms of the number of SGD steps, $T=\eta K$. Relative to a setting with a common diffusion and global strong monotonicity, (a) the diffusion may depend on S, at the price of the sensitivity ς_n, whose size ($n^{-1/2}$ in general, Proposition 1(i), versus n^{-1} with a constant of order at least $\sqrt{d}$ under uniform non-degeneracy, which requires $d<n$, Proposition 1(ii)) matters for the rate (Corollary 5); (b) the one-sided monotonicity constant μ may be negative, so finite-horizon bounds hold without convexity, provided the remaining assumptions on the drift and on the diffusion ((A1)–(A4)) are satisfied; when $\kappa<0$ these bounds grow like $e^{|\kappa|T}$ (Corollary 3).

Relation with discrete SGD

Strong versus weak approximation. The stability quantity $\mathbb{E}\|w_t-w_t'\|$ refers to a coupled pair of trajectories, whereas the SME approximation of [7] holds in the weak sense, for expectations of test functions. The pathwise distance between the SGD iterates and a Gaussian SDE is in general not small, because the gradient noise is not Gaussian; it may even be heavy-tailed [15], and the validity of SDE models of SGD depends on the regime [8]. The diffusion-mismatch term is of order $\varsigma_n\sqrt{\eta}$, which is the order one

Corrected Proof

should expect for the strong error between SGD and the surrogate. Therefore the bound on $\mathbb{E}\|w_t - w'_t\|$, and in particular this term, does not transfer to the SGD iterates, and it may lie below the resolution of the surrogate. The generalization gap of Corollary 4 is a different matter, because it depends only on the law of w_T given S (Remark 3). A weak-error estimate $|\mathbb{E}[\phi(w_K^{\mathrm{SGD}}) \,|\, S] - \mathbb{E}[\phi(w_{\eta K}) \,|\, S]| \le C(T)\eta$ for the test functions $\phi = l_0(\cdot, z)$, with $C(T)$ uniform in S and z, would give $|\mathrm{gap}_{\mathrm{SGD}} - \mathrm{gap}_{\mathrm{SDE}}| \le 2C(T)\eta$. We do not verify the hypotheses of [7] under which such an estimate holds, $C(T)$ typically grows with T, and a quantitative statement for SGD is left open.

Comparison with the discrete-SGD bounds of [4]. For γ-strongly convex, L-Lipschitz and β-smooth losses, SGD with constant step size at most $1/\beta$ is uniformly stable with constant $2L^2/(\gamma n)$ [4]. Their coupling uses the same sample indices on S and $S^{(i)}$, so the two runs differ only at the steps where the differing sample is drawn, which happens with probability $1/n$ per step (b/n for mini-batches of size b). In our bound the gradient-mismatch term $2\rho^2/(\kappa n)$ reproduces this rate (with $\kappa \to \mu$ as $\eta \to 0$). The diffusion-mismatch term, however, is generically of order $\rho\varsigma_n\sqrt{\eta\mu}/\kappa = O(\sqrt{\eta/n})$ (Proposition 1(i), Remark 4), which is larger than the SGD rate for large n. We do not claim that this term reflects a property of SGD. Heuristically, in SGD the influence of the differing sample is sparse in time, and its contribution to $\mathbb{E}\|\Delta\|$ accumulates at rate $O(1/n)$. In the Gaussian surrogate the same second moment is spread over a continuous martingale, whose first absolute moment is of the order of the square root of its variance, that is $O(n^{-1/2})$. The $n^{-1/2}$ term is therefore a feature of the Gaussian surrogate under the synchronous coupling, for which it is sharp (Remark 2(3)). Whether a different coupling of the SDE gives an $O(n^{-1})$ generalization gap is left open.

Comparison of the bounds. Table 1 summarizes the dependence of the bounds on n, η, d and the horizon, compared with [4] and [3].

Table 1: Dependence of the stability and generalization bounds on n, η, d and the horizon. For SGD the horizon is the number of steps K; for the SDE it is $T = \eta K$.

Bound	rate in n	η	d	horizon
[4], SGD, strongly convex	$2L^2/(\gamma n)$	step $\le 1/\beta$	none	uniform in K
[4], SGD, non-convex	$O(K^{\beta c/(\beta c+1)}/n)$	steps $\le c/k$	none	polynomial in K
[3], SGLD, Lipschitz loss with weight decay	$O(k/n)$, k batch size	discrete time: extra $(\eta d)^{1/2}$	none in continuous time	uniform in time
This paper, $\kappa > 0$, Prop. 1(i)	$O(1/n) + O(\sqrt{\eta/n})$	$\kappa > 0$ iff $\eta < 2\mu/L_\Sigma^2$	none	uniform in T
This paper, $\kappa > 0$, Prop. 1(ii)	$O(1/n) + O(\sqrt{\eta d}/n)$	$\eta < \mu b/(2d\ell_0^2)$	$\ge \sqrt{d}$; needs $d < n$	uniform in T
This paper, $\kappa \le 0$	$e^{\lvert\kappa\rvert T}\, O\big(T/n + \sqrt{\eta T(1+\mu_+ T)/n}\big)$	through κ	none (factor (i))	exponential in T

The following questions are not covered in this paper:

- Time-uniform bounds under dissipativity only (strong convexity outside a ball). Farghly and

Corrected Proof

Rebeschini [3] obtain them for SGLD with injected noise, through the Wasserstein contraction of the Langevin diffusion; extending this to a state- and dataset-dependent diffusion would need a coupling more refined than the synchronous one used here and is not attempted.

- The discretisation error between SGD and (1): the statements are about the surrogate (see [7]).
- An independent justification of (A4): it is a property of the chosen model for Σ_S; Proposition 1 checks it for the gradient-covariance diffusion, with explicit constants for the factor (i).
- Over-parameterized networks: there the Hessian typically has many (near-)zero eigenvalues, so $\mu \leq 0$ and only Corollary 3 applies. We make no claim that these bounds explain the generalization of over-parameterized models, which the strongly monotone setting does not model; the question was the motivation of [11].

Conclusions

We proved an explicit stability bound for a noisy gradient dynamics with state- and dataset-dependent diffusion and explicit learning-rate scaling. The bound separates the contraction rate $\kappa = \mu - \frac{1}{2}\eta L_\Sigma^2$, the gradient mismatch $2\rho/n$ and the diffusion mismatch $\varsigma_n\sqrt{\eta}$; it gives time-uniform generalization bounds when $\kappa > 0$ and finite-horizon ones otherwise. For the gradient-covariance diffusion the sensitivity ς_n is $O(n^{-1/2})$ in general, with the n-dimensional factor and dimension-free constants (Proposition 1(i)), and $O(n^{-1})$ for uniformly non-degenerate noise, with a constant of order at least $\sqrt{d}$ and only when $d < n$ (Proposition 1(ii)); this determines whether the diffusion mismatch or the gradient mismatch governs the sample complexity. The worked example of ℓ_2-regularized logistic regression shows that all the hypotheses can be met with explicit constants. The gradient-mismatch term matches the discrete-SGD rate of [4], while the diffusion-mismatch term is a property of the Gaussian surrogate under synchronous coupling and is not claimed to describe SGD. Natural next steps are the dissipative (non-globally monotone) case with a non-synchronous coupling, a coupling of the SDE that reflects the sparse dependence of SGD on the differing sample, and a quantitative comparison between the SDE and the SGD iterates.

Conflict of interest

The author declare that there is no conflict of interest in the publication of this paper.

Funding

There are no funders to report for this submission.

Acknowledgments

The author is grateful to the anonymous referees for their careful reading of the manuscript and for their valuable comments and suggestions, which considerably improved the presentation and the precision of the results.

Corrected Proof

References

[1] R. Bhatia, *Matrix Analysis*, Graduate Texts in Mathematics 169, Springer-Verlag, New York, 1997.

[2] O. Bousquet, A. Elisseeff, *Stability and generalization*, *J. Mach. Learn. Res.* **2** (2002) 499–526.

[3] T. Farghly, P. Rebeschini, *Time-independent generalization bounds for SGLD in non-convex settings*, *Adv. Neural Inf. Process. Syst.* **34** (2021).

[4] M. Hardt, B. Recht, Y. Singer, *Train faster, generalize better: Stability of stochastic gradient descent*, *Proc. 33rd Int. Conf. Mach. Learn.* PMLR **48** (2016) 1225–1234.

[5] I. Karatzas, S.E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Graduate Texts in Mathematics 113, Springer-Verlag (ISBN 0-387-97655-8), 1991.

[6] J. Li, X. Luo, M. Qiao, *On generalization error bounds of noisy gradient methods for non-convex learning*, *Int. Conf. Learn. Represent. (ICLR)* (2020). arXiv:1902.00621.

[7] Q. Li, C. Tai, W. E, *Stochastic modified equations and adaptive stochastic gradient algorithms*, *Proc. 34th Int. Conf. Mach. Learn.* PMLR **70** (2017).

[8] Z. Li, S. Malladi, S. Arora, *On the validity of modeling SGD with stochastic differential equations (SDEs)*, *Adv. Neural Inf. Process. Syst.* **34** (2021) 12712–12725.

[9] S. Mandt, M.D. Hoffman, D.M. Blei, *Stochastic gradient descent as approximate Bayesian inference*, *J. Mach. Learn. Res.* **18**(1) (2017) 4873–4907.

[10] X. Mao, *Stochastic Differential Equations and Applications*, 2nd ed., Horwood Publishing, 2007.

[11] F. Morlando, *The paradox of over-parameterization in learning algorithm*, *Ann. Math. Comput. Sci.* **34** (2026) 1–13.

[12] W. Mou, L. Wang, X. Zhai, K. Zheng, *Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints*, *Proc. 31st Conf. Learn. Theory* PMLR **75** (2018) 605–638.

[13] R.T. Powers, E. Størmer, *Free states of the canonical anticommutation relations*, *Commun. Math. Phys.* **16** (1970) 1–33.

[14] M. Raginsky, A. Rakhlin, M. Telgarsky, *Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis*, *Proc. 2017 Conf. Learn. Theory* PMLR **65** (2017) 1674–1703.

[15] U. Şimşekli, L. Sagun, M. Gürbüzbalaban, *A tail-index analysis of stochastic gradient noise in deep neural networks*, *Proc. 36th Int. Conf. Mach. Learn.* PMLR **97** (2019) 5827–5837.

Corrected Proof