Chaotic Regularization and Heavy-Tailed Limits for Deterministic Gradient Descent

Soon Hoe Lim
Nordita
KTH Royal Institute of Technology and Stockholm University
soon.hoe.lim@su.se

Yijun Wan
Département de Mathématiques et Applications
École Normale Supérieure, Université PSL
wan@clipper.ens.fr

Umut Şimşekli
DI ENS, École Normale Supérieure, Université PSL, CNRS, INRIA
umut.simsekli@inria.fr

Abstract

Recent studies have shown that gradient descent (GD) can achieve improved generalization when its dynamics exhibits a chaotic behavior. However, to obtain the desired effect, the step-size should be chosen sufficiently large, a task which is problem dependent and can be difficult in practice. In this study, we incorporate a chaotic component to GD in a controlled manner, and introduce *multiscale perturbed GD* (MPGD), a novel optimization framework where the GD recursion is augmented with chaotic perturbations that evolve via an independent dynamical system. We analyze MPGD from three different angles: (i) By building up on recent advances in rough paths theory, we show that, under appropriate assumptions, as the step-size decreases, the MPGD recursion converges weakly to a stochastic differential equation (SDE) driven by a heavy-tailed Lévy-stable process. (ii) By making connections to recently developed generalization bounds for heavy-tailed processes, we derive a generalization bound for the limiting SDE and relate the worst-case generalization error over the trajectories of the process to the parameters of MPGD. (iii) We analyze the implicit regularization effect brought by the dynamical regularization and show that, in the weak perturbation regime, MPGD introduces terms that penalize the Hessian of the loss function. Empirical results are provided to demonstrate the advantages of MPGD.

1 Introduction

Many important problems in supervised learning can be expressed by the following population risk minimization problem:

$$\min_{x\in\mathbb{R}^d}\Big\{\mathcal{R}(x) := \mathbb{E}_{z\sim\mathcal{D}}[\ell(x,z)]\Big\}, \tag{1}$$

where $x \in \mathbb{R}^d$ denotes the learnable parameter, $z \in \mathcal{Z}$ denotes a data sample coming from an unknown data distribution $\mathcal{D}$, $\mathcal{Z}$ denotes the space of data points, and $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}_+$ is the loss

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

function, measuring the quality of the parameter x. Since the data distribution $\mathcal{D}$ is not known, as a proxy for solving (1), we consider the empirical risk minimization (ERM) problem, given as follows:

$$\min_{x\in\mathbb{R}^d}\Big\{\hat{\mathcal{R}}(x,S_n) := \frac{1}{n}\sum_{i=1}^{n}\ell(x,z_i)\Big\}, \tag{2}$$

where $S_n := \{z_1,\dots,z_n\}$ denotes a *training set* of n points that are independently and identically distributed (i.i.d.) and sampled from $\mathcal{D}$. To solve the ERM problem, typically gradient-based optimization algorithms are being used in practice.

In this study, we will consider the gradient descent (GD) algorithm [C$^+$47], which is based on the following deterministic recursion:

$$x_{k+1} = x_k - \eta_k \nabla\hat{\mathcal{R}}(x_k,S_n), \quad k=0,1,2,\dots. \tag{3}$$

Here, k represents the iteration number, $\eta_k > 0$ is the step-size (learning rate), and $x_0 \in \mathbb{R}^d$ is the initialization. Even though stochastic counterparts of GD (e.g., stochastic GD [BCN18]) have been more popular due to their reduced computational requirements, the GD algorithm has attracted increasing attention in the past few years, as it has been illustrated that GD can achieve similar generalization performance, as measured by the *generalization error* $|\hat{\mathcal{R}}(x_k,S_n) - \mathcal{R}(x_k)|$, as long as certain design choices are implemented [GGP$^+$21]. These design choices mainly include using large and/or oscillating learning rates and using explicit regularization of the empirical risk.

Contrary to the wisdom of optimization theory, which typically suggests using smaller learning rates, the fact that large learning rates often provide better generalization performance in modern non-convex optimization problems has opened up several research directions. In this context, [CKL$^+$21] showed that large learning rates drive the GD dynamics towards the 'edge of stability', meaning that the learning rate should be chosen large enough to make the algorithm diverge. Recently, the large learning rates have been extended to exotic learning rate schedules (i.e., non-trivial evolution of η_k), such as fractal scheduling [AGZ21] and cyclic scheduling [Oym21, Smi17].

It has been shown that the use of large learning rates introduces a *chaotic behavior* in the GD dynamics [KT20]. Hence, a possible explanation for the improvements brought by large learning rates can be partially attributed to such chaotic behavior, since chaotic systems exhibit a stochastic-like evolutions (even though they are fully deterministic), which might be an underlying cause of the improvements in generalization obtained in non-convex settings. However, achieving the desired chaotic dynamics is a problem-dependent task and can be often difficult in practice.

In this paper, we develop a novel optimization framework, coined *multiscale perturbed gradient descent* (MPGD), where a chaotic component is introduced to GD in a controlled manner. Our approach is based on extending the GD recursion via generic regularization terms, where the regularization coefficients are modulated by using external deterministic dynamical systems that exhibit chaotic characteristics. Our main contributions are as follows.

- We build up on recent advances in rough paths theory and homogenization techniques [GM13, CFK$^+$16, CFKM20], and show that, under appropriate assumptions, as the step-size decreases, the MPGD recursion converges weakly to a stochastic differential equation (SDE) driven by a heavy-tailed Lévy-stable process (Theorem 4.1; see also Theorem C.1 in Appendix).
- We draw connections to the recent theoretical links between heavy-tailed processes and generalization [SSG19, ŞSDE20, HŞKM21], and under certain topological regularity assumptions, we derive a generalization bound (Theorem 4.2) for the limiting SDE and relate the worst-case generalization error over its trajectories to the parameters of MPGD.
- We show that MPGD exhibits a natural form of implicit regularization by deriving an appropriate explicit regularizer in the weak perturbation regime (Theorem 4.3). Our analysis shows that, in this regime, the derived regularizer introduces terms that penalize the Hessian of the loss function.
- We empirically demonstrate the advantages of MPGD on different settings and provide further support for the developed framework and the theory (see Section 5).

We emphasize that our primary goal is to understand the behavior of deterministic GD in a chaotic setting. In this respect, we shall underline that, at this stage, our goal is not to develop a competitive algorithm that would outperform SGD, but rather to provide a solid theoretical analysis for how stochastic behavior can emerge from GD with suitable deterministic components, as well as their implication in terms of generalization error.

2 Preliminaries and Technical Background

Notation. For a vector $v \in \mathbb{R}^d$, we denote its ith component by either v^i or $[v]^i$ and the Euclidean norm by $|v|$ (or $\|v\|$). $\mathrm{diag(v)}$ denotes the diagonal matrix with ith diagonal entry of v^i. $\langle \cdot, \cdot \rangle$ denotes the dot product of two vectors. $\mathcal{U}(d)$ denotes uniform distribution on d dimension, tr denotes trace, and the superscript ^{T} denotes transposition. $\circ$ and $\odot$ denote composition and Hadamard product respectively. ∇ and ∇^2 denote gradient and Hessian respectively. $\mathbb{E}$ denotes expectation. $\overset{\text{(d)}}{=}$ and $\overset{\text{(d)}}{\to}$ denote equivalence and convergence in the sense of distribution respectively. I denotes the identity matrix, and 1_A denotes indicator function of the set A.

Stable distributions. Let $\alpha \in (0, 2]$ be a stability parameter and $d \geq 1$. A random variable $X \in \mathbb{R}^d$ is α-*stable* if $X_1, X_2, \ldots$ are independent copies of X, then $n^{-1/\alpha} \sum_{j=1}^{n} X_j \overset{\text{(d)}}{=} X$ for all $n \geq 1$ [Sam17]. Stable distributions appear as the limiting distribution in the generalized central limit theorem (CLT) [GK54]. The case $\alpha = 2$ corresponds to the Gaussian distribution, while $\alpha = 1$ corresponds to the Cauchy distribution. The p-th moment of a stable random variable is finite if and only if $p < \alpha$. Here we are interested in the case $\alpha \in (1, 2)$, such that $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|X|^2] = \infty$. If X is symmetric, then there exists a scale parameter $c > 0$ and the characteristic function of X is given by $\mathbb{E}\left[\exp(i\langle \xi, X \rangle)\right] = \exp(-|c\xi|^\alpha)$ for all $\xi \in \mathbb{R}^d$. If X is a random vector with independent components, then there exists a scale vector $\mathbf{c} = (c^1, \ldots, c^d) \in \mathbb{R}^d_+$ and the characteristic function of X is given by $\mathbb{E}\left[\exp(i\langle \xi, X \rangle)\right] = \exp(-\sum_{j=1}^{d} |c^j \xi^j|^\alpha)$.

Lévy processes. Lévy processes are stochastic processes $(\mathrm{L}_t)_{t \geq 0}$ with independent and stationary increments, and are defined as follows [App09]:

1. For $N \in \mathbb{N}$ and $t_0 < t_1 < \ldots < t_N$, the increments $(\mathrm{L}_{t_i} - \mathrm{L}_{t_{i-1}})$ are independent for all i.
2. For any $t > s > 0$, $(\mathrm{L}_t - \mathrm{L}_s)$ and L_{t-s} have the same distribution.
3. L_t is continuous in probability, i.e., for all $\delta > 0$ and $s \geq 0$, $\mathbb{P}(|\mathrm{L}_t - \mathrm{L}_s| > \delta) \to 0$ as $t \to s$.

Typical examples of Lévy processes include Brownian motion and the α-stable processes. By the Lévy-Khintchine formula, a Lévy process $(\mathrm{L}_t)_{t \geq 0}$ with $\mathrm{L}_0 = 0$ is determined by a triplet (b, Σ, ν) for some $b \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ positive semi-definite and a measure ν on $\mathbb{R}^d \setminus \{0\}$ such that $\int_{x \neq 0} \min\{1, |x|^2\} \nu(dx) < \infty$. The characteristic function of L_t is therefore $\exp(-t\Psi(\xi))$, with the characteristic exponent $\Psi : \mathbb{R}^d \to \mathbb{C}$ defined by

$$\Psi(\xi) = -i\langle b, \xi \rangle + \frac{1}{2}\langle \xi, \Sigma\xi \rangle + \int_{\mathbb{R}^d} \left[1 - e^{i\langle x, \xi \rangle} + i\langle x, \xi \rangle 1_{|x|<1}\right] \nu(dx). \tag{4}$$

In the Lévy–Itô decomposition of a Lévy process, b denotes a constant drift, Σ is the covariance matrix of a Brownian motion and ν characterises a Lévy jump process, which is modified to be càdlàg (right-continuous and has left limits everywhere) with countably many jumps. In particular, if $b = 0_{\mathbb{R}^d}$, $\Sigma = 0_{\mathbb{R}^{d \times d}}$ and $\nu(dx) = |x|^{-\alpha-1} dx$ for some $\alpha \in (0, 2)$, then the triplet $(0_{\mathbb{R}^d}, 0_{\mathbb{R}^{d \times d}}, |x|^{-\alpha-1} dx)$ gives a symmetric α-stable process, denoted by $(\mathrm{L}^\alpha_t)_{t \geq 0}$.

Marcus differential equations. Passing to the limit of the driving processes with jumps in an SDE is not trivial, as one needs to be careful with the meaning of integration [CP14]. Unfortunately, the common Itô and Stratonovich integration (which are based on Riemann-Stieltjes sums with the integrand evaluated at the left end point and the midpoint of the partition intervals respectively) fail to provide the desired convergence. To this end, we resort to Marcus differential equations [Mar81], which in their basic form are given by:

$$\mathrm{d}X_t = b(X_t) \diamond \mathrm{dL}_t,$$

where $\diamond$ denotes integration in the Marcus sense. Marcus integrals involve sums over infinitely many jumps and transform under the usual laws of calculus [App09], and thus play a similar role for Lévy processes as the Stratonovich integral for Brownian motion. The solution map of Marcus differential equations is continuous with respect to the driving process under certain variants of the Skorokhod topology. Note that if $b(X_t)$ is constant in X_t, then there is no difference in the solutions of the SDE in the sense of Marcus, Itô and Stratonovich. We refer to Appendix B for more details on these.

Generating stable laws using the Thaler map. Let $\gamma := 1/\alpha$. It was shown in [GM21] that α-stable laws with $\alpha \in (1, 2)$ can be generated using a deterministic dynamical system whose states

y_k are obtained by iterating the Thaler map $T : [0,1] \to [0,1]$ [Tha80]:

$$T(y) = (y^{1-\gamma} + (1+y)^{1-\gamma} - 1)^{1/(1-\gamma)} \mod 1\,,\; y_{k+1} = T(y_k), \tag{5}$$

with any source of randomness coming solely from the initial condition $y_0 \in [0,1]$. The map has the following properties. Let $y^* \in (0,1)$ be the unique solution to the equation

$$(y^*)^{1-\gamma} + (1+y^*)^{1-\gamma} = 2.$$

There are two increasing branches on the intervals $[0, y^*]$, $[y^*, 1]$. For $\gamma \in [0,1)$, there exists a unique invariant probability density $h(y) = \frac{1-\gamma}{2^{1-\gamma}}(y^{-\gamma} + (1+y)^{-\gamma})$. Moreover, the density defines a finite measure. For $\gamma \in (0,1)$, the map is nonuniformly expanding with a neutral fixed point at $y = 0$ and correlations decay algebraically with rate $1/k^{\alpha-1}$ (which is sharp) [GM21]. Heuristically, the trajectory spends prolonged period of times near $y = 0$ (laminar dynamics near the fixed point), slowing down the decay of correlation as α decreases.

Let the observable $v : [0,1] \to \mathbb{R}$ be a Hölder function with zero mean with respect to the invariant measure of T, i.e., $\mathrm{d}\nu = h\mathrm{d}y$, and consider the sequence of random variables $(v \circ T^k(y_0))_{k\geq 0}$, with the randomness coming from the initial state y_0. When $\gamma \in (1/2, 1)$, the correlations are not summable and CLT breaks down if $v(0) \neq 0$ (it "sees" the fixed point at $y = 0$). Heuristically, the Birkhoff sum $\sum_{j=0}^{k-1} v \circ T^j$ is ballistic with almost linear behavior near $x = 0$ and the small jumps of size $v(0)$ accumulate into a single large jump, contradicting CLT [GM21]. Instead, if v is properly normalized, then it was shown in [Gou04] that the one-sided stable limit law $\frac{1}{k^\gamma}\sum_{j=0}^{k-1} v \circ T^j \overset{(\mathrm{d})}{\to} X_{\alpha,\beta}$ as $k \to \infty$, with $\beta = \mathrm{sign}(v(0))$, holds instead, where $X_{\alpha,\beta}$ denotes the stable law whose characteristic function is given by:

$$\mathbb{E}[e^{itX_{\alpha,\beta}}] = \exp(-|t|^\alpha(1 - i\beta\,\mathrm{sign}(t)\tan(\alpha\pi/2))). \tag{6}$$

In this paper, we focus on MPGD with chaotic components whose limiting behavior at each time is described by the class of stable laws $X_{\alpha,\beta}$ whose characteristic function is given by (6) with the stability parameter $\alpha \in (1,2)$ (i.e., $\gamma \in (1/2, 1)$) and the skewness parameter $\beta \in [-1, 1]$. The $X_{\alpha,\beta}$ is centered (i.e., $\mathbb{E}X_{\alpha,\beta} = 0$), totally skewed (or one-sided) if $\beta = \pm 1$, and symmetric if $\beta = 0$.

3 Multiscale Perturbed Gradient Descent (MPGD)

In full generality, we start by considering the following family (parametrized by $m > 0$) of deterministic fast-slow dynamical systems on $\mathbb{R}^d \times Y$, with Y a bounded metric space:

$$\begin{cases} x_{k+1}^{(m)} = x_k^{(m)} + \dfrac{1}{m} a_m(x_k^{(m)}) + \displaystyle\sum_{i=1}^{q} \dfrac{1}{m^{1/\alpha_i}} b_m^{(i)}(x_k^{(m)}) v_i(y_k^{(i)}), \\ y_{k+1}^{(i)} = T_i(y_k^{(i)}), \;\; i = 1, 2, \ldots, q, \end{cases} \tag{7}$$

for $k = 0, 1, 2, \ldots$, where the $\alpha_i \in (1,2)$, $a_m : \mathbb{R}^d \to \mathbb{R}^d$, $b_m^{(i)} : \mathbb{R}^d \to \mathbb{R}^{d\times r_i}$, $v_i : Y \to \mathbb{R}^{r_i}$, and $T_i : Y \to Y$. The initial conditions $x_0^{(m)}, y_0^{(i)}$ are independent random variables. One important example of recursions of the form (7) is the GD defined in (3). Other than GD, the above family of recursions (7) includes various variants of GD as special cases upon specifying suitable choices of the a_m, $b_m^{(i)}$, v_i, T_i, and the dimensions.

Of particular interest are new algorithms that can be studied within the above framework. To this end, we introduce *multiscale perturbed gradient descent* (MPGD), whose dynamics are described by the following recursion:

$$\begin{cases} x_{k+1}^{(m)} = x_k^{(m)} - \dfrac{1}{m}\nabla\hat{\mathcal{R}}(x_k^{(m)}, S_n) - \dfrac{\mu}{m^{\frac{1}{\alpha_1}}} v_1(y_k^{(1)}) \odot x_k^{(m)} + \dfrac{\sigma}{m^{\frac{1}{\alpha_2}}} v_2(y_k^{(2)}), \\ y_{k+1}^{(1)} = T(y_k^{(1)}), \;\; y_{k+1}^{(2)} = T(y_k^{(2)}), \;\; k = 0, 1, 2, \ldots, \end{cases} \tag{8}$$

where T is the Thaler map (5) with $\gamma = \alpha^{-1} \in (1/2, 1)$, the v_1, v_2 are observable maps, $\mu \geq 0$ and $\sigma \in \mathbb{R}$ are tunable parameters. The above recursion is of the form (7), with $q = 2$, $r_1 = r_2 = d$,

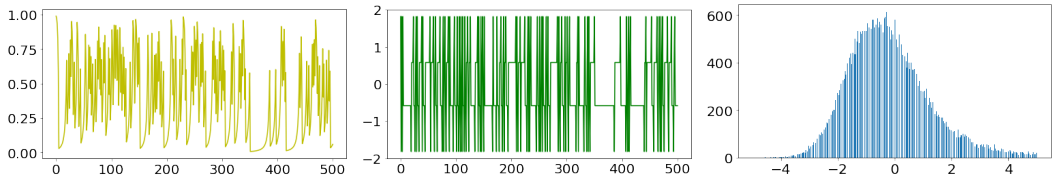

Figure 1: A realization of the Thaler iterates (left) and the observable v (middle) for $\gamma = 0.6$, $\beta = 0.1$. Empirical distribution of the corresponding Birkhoff sum $\frac{1}{k^\gamma}\sum_{j=0}^{k-1} v^{(j)}$ with $k = 1000$ (right).

$a_m(x) = -\nabla\hat{\mathcal{R}}(x, S_n)$, $b_m^{(1)}(x) = -\mu\,\mathrm{diag(x)}$, and $b_m^{(2)}(x) = \sigma I$ independent of m. Another class of algorithms that falls within this framework is when m is set to n, the number of training data points, in which case the empirical risk converges to the population risk as $m = n \to \infty$.

To complete the description of MPGD, it remains to specify the choice for the observables $v_i(y_k^{(i)})$ in (8) whose rescaled version can be shown to generate stable laws $X_{\alpha,\beta}$, with $\alpha \in (1, 2)$ and $\beta \in [-1, 1]$, in the limit $m \to \infty$. Following [GM21], we take them to be independent realizations of the observable $v^{(k)}$ described as follows. Let $d_\alpha = \frac{\alpha^\alpha(1-\gamma)\Gamma(1-\alpha)\cos(\alpha\pi/2)}{2^{1-\gamma}-1}$, and $\delta_0, \delta_1, \ldots$ be independent copies of the random variable δ where $\mathbb{P}(\delta = \pm 1) = \frac{1}{2}(1 \pm \beta)$. The observables $v^{(k)}$ are then defined to be: $v^{(k)} = \chi^{(k)} v \circ T^k$, where $v : [0, 1] \to \mathbb{R}$ is the mean zero observable given by $v(y) = d_\alpha^{-\gamma}(1 - 2^{\gamma-1})^{-\gamma}\tilde{v}(y)$, with

$$\tilde{v}(y) = \begin{cases} 1 & \text{if } y \le y^*, \\ (1 - 2^{1-\gamma})^{-1} & \text{if } y > y^*, \end{cases} \quad \chi^{(k)} = \chi_{k-1}\cdots\chi_0 \in \{\pm 1\} \text{ with } \chi_j = \begin{cases} 1 & \text{if } T^j y \le y^*, \\ \delta_j & \text{if } T^j y > y^*. \end{cases}$$

In particular, the random variables $\chi^{(k)}$ get updated only when the trajectory visits $(y^*, 1]$ and are not changed during the phase in $[0, y^*]$.

The initial condition can be, in theory, equally well chosen using the invariant probability measure ν or the uniform Lebesgue measure. Empirically, convergence of the probability density is faster if it is drawn using ν. Hence, we consider initial conditions drawn using ν. Therefore, for MPGD we propagate uniformly distributed initial conditions $y_0' \in [0, 1]$ under 10,000 iterations of the Thaler map and take the initial condition as $y_0 = T^{10000} y_0'$. Figure 1 illustrates the dynamics of a realization of the Thaler iterates and the observable (perturbations used in MPGD), as well as the asymmetry and heavy tail nature of the distribution of the corresponding Birkhoff sum.

4 Theoretical Analysis

In this section, we analyze MPGD via three different angles. We start by presenting results on superdiffusive limit (homogenization [PS08]) for a class of fast-slow deterministic systems that include MPGD as a special case in Subsection 4.1. We then study generalization properties of MPGD by deriving a generalization bound for the limiting dynamics in Subsection 4.2. Lastly, we study the regularizing effects of MPGD via the lens of implicit regularization in Subsection 4.3.

4.1 Superdiffusive Limits for the Fast-Slow Systems and MPGD

Without loss of generality, we set $q = 1$ in (7) and consider the following family of recursions:

$$\begin{cases} x_{k+1}^{(m)} = x_k^{(m)} + m^{-1} a_m(x_k^{(m)}) + m^{-\frac{1}{\alpha}} b_m(x_k^{(m)}) v(y_k), \\ y_{k+1} = T(y_k), \end{cases} \tag{9}$$

where T is the Thaler map and v is the observable constructed in Section 3.

The following result shows that stochastic dynamics emerges in the above family of deterministic dynamical systems in the limit $m \to \infty$. Moreover, the limiting dynamics are described by a stochastic differential equation (SDE) driven by a heavy-tailed Lévy-stable process.

Theorem 4.1 (Superdiffusive Limit – Informal)**.** *Let $\alpha \in (1, 2)$ and assume that the coefficients a_m, b_m and the initial condition $x_0^{(m)}$ of* (9) *are well-behaved as $m \to \infty$. Then, under certain regularity*

conditions on the a_m, b_m, the process $(x^{(m)}_{\lfloor mt \rfloor})_{t\geq 0}$ converges in distribution to the solution $(X_t)_{t\geq 0}$ of the SDE:

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t + b(X_t) \diamond \mathrm{d}\mathrm{L}^{\alpha,\beta}_t \text{ with } X_0 = x_0 \in \mathbb{R}^d,$$

as $m \to \infty$, where $a = \lim_{m\to\infty} a_m$, $b = \lim_{m\to\infty} b_m$, $x_0 = \lim_{m\to\infty} x^{(m)}_0$, $\diamond$ denotes the Marcus integration, defined in Section B.2.1, and $(\mathrm{L}^{\alpha,\beta}_t)_{t\geq 0}$ is a Lévy process with the characteristic exponent

$$\Psi(\xi) = |\xi|^\alpha (1 - i\beta \operatorname{sign}(\xi) \tan(\alpha\pi/2)). \tag{10}$$

In particular, Theorem 4.1 implies convergence in distribution of $x^{(m)}_{\lfloor mt \rfloor}$ to X_t at fixed t. It is a special case of Theorem C.1, which covers a large class of the maps T and observables v, and extends the results in [GM21, CFKM20] to the case where the coefficients are dependent on m. We provide technical details on the topology used for the convergence and the proof in Appendix B-C.

Specializing Theorem 4.1 (extended to the case of $q = 2$) to the MPGD recursion (8), we see that the rescaled MPGD iterate process, $(x^{(m)}_{\lfloor mt \rfloor})_{t\geq 0}$, converges in distribution to the solution $(X_t)_{t\geq 0}$ of the following SDE:

$$\mathrm{d}X_t = -\nabla\hat{\mathcal{R}}(X_t, S_n)\mathrm{d}t - \mu \operatorname{diag}(X_t) \diamond \mathrm{d}\mathrm{L}^{\alpha_1,\beta_1}_t + \sigma \mathrm{d}\tilde{\mathrm{L}}^{\alpha_2,\beta_2}_t, \;\; X_0 = \lim_{m\to\infty} x^{(m)}_0,$$

in the limit $m \to \infty$, which is equivalent to sending the step-size, $\eta := 1/m$, in the MPGD to zero. Here $(\mathrm{L}^{\alpha_1,\beta_1}_t)_{t\geq 0}$ and $(\tilde{\mathrm{L}}^{\alpha_2,\beta_2}_t)_{t\geq 0}$ are independent Lévy processes with characteristic exponent (10), and β_1, $\beta_2 \in [-1, 1]$ are the parameters appearing in the construction of the observables v_1 and v_2 respectively. Therefore, we see that MPGD converges to a gradient flow driven by heavy-tailed Lévy-stable processes in the considered scaling limit. The driving processes contain both additive and multiplicative components. The multiplicative component can be interpreted as a dynamical version of the weight decay, whereas the additive component can be viewed as gradient noise [NVL+15].

4.2 Generalization Properties of MPGD

In this section, we shift our focus to the generalization properties of MPGD. Our main roadmap will be to consider the limiting SDE arising from Theorem 4.1 (with $\beta = 0$) and relate it to the recently developed generalization bounds for heavy-tailed random processes [ŞSDE20, HŞKM21].

For mathematical convenience, we consider the following special case of (7)[1]:

$$\begin{cases} x^{(m)}_{k+1} = x^{(m)}_k + m^{-1} a_m(x^{(m)}_k) + m^{-\frac{1}{\alpha}} b_m(x^{(m)}_k) \left(v(y^{(1)}_k) - v(y^{(2)}_k) \right), \\ y^{(1)}_{k+1} = T(y^{(1)}_k), \quad y^{(2)}_{k+1} = T(y^{(2)}_k) \end{cases} \tag{11}$$

In the light of Theorem 4.1, the recursion (11) converges weakly to the following SDE:

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t + b(X_t) \diamond \mathrm{d}\mathrm{L}^{\alpha}_t, \tag{12}$$

where L^α_t denotes the *symmetric* α-stable process in $\mathbb{R}^d$ (i.e., $\beta = 0$).

Following the previous work [ŞSDE20, HŞKM21], we are interested in bounding the worst-case generalization error over the trajectories of (12). More precisely, let $(X_t)_{0\leq t\leq 1}$ denote the solution of (12)[2], and define the set $\mathcal{X}$ so that it contains all the points visited by the trajectory:

$$\mathcal{X} := \{x \in \mathbb{R}^d : \exists t \in [0, 1], X_t = x\}. \tag{13}$$

Then, our goal is to analyze $\sup_{x\in\mathcal{X}} |\hat{\mathcal{R}}(x, S_n) - \mathcal{R}(x)|$. Since we are mainly interested in the case when $a(x) = -\nabla\hat{\mathcal{R}}(x, S_n)$, the SDE (12) and hence the trajectory $\mathcal{X}$ will depend on the dataset S_n. Therefore, in the generalization bound, we will need to control the statistical dependence of $\mathcal{X}$ on S_n, through the notion of ρ-mutual information, defined as follows. Let X, Y be two random elements, let $\mathbb{P}_{X,Y}$ denote their joint distribution, and let $\mathbb{P}_X, \mathbb{P}_Y$ denote their respective marginal distributions. Then the ρ-mutual information between X and Y is defined as: $I_\rho(X, Y) = D_\rho(\mathbb{P}_{X,Y} \| \mathbb{P}_X \otimes \mathbb{P}_Y)$,

[1] We believe a similar result would hold for (9) without symmetrization, which we leave for future work.

[2] Here, the range $[0, 1]$ is arbitrary and could be replaced with any range $[0, T]$ for $0 < T < \infty$.

where D_ρ is the ρ-Renyi divergence:[3] $D_\rho(\mu,\nu) = \frac{1}{\rho-1}\log \mathbb{E}\left[\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(Z)^\rho \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(Z)^{1-\rho}\right]$, and $\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ denotes the Radon-Nikodym derivative.

Finally, we require two technical regularity assumptions (Assumptions 2 and 3, given in Appendix D) for our generalization bound. Informally, Assumption 2 is a topological regularity condition over the trajectory $\mathcal{X}$, whereas Assumption 3 is a statistical regularity assumption and imposes that in the neighborhood of a local minimum $x^\star$ of $\hat{\mathcal{R}}(x, S_n)$, the statistical behavior of the process X_t can be approximated by the process solving the SDE $\mathrm{d}\tilde{X}_t = b(x^\star)\diamond \mathrm{dL}_t^\alpha$. This requires b to be regular in the neighborhood of $x^\star$ and further imposes that $a(x^\star) = 0$, which is natural when $a(x) = -\nabla\hat{\mathcal{R}}(x, S_n)$. We now ready to present our generalization bound.

Theorem 4.2 (Generalization bound)**.** *Assume that Assumptions 2 and 3 given in Section D hold, and that ℓ is bounded by $B > 0$ and L-Lipschitz continuous. There exists a constant $K_1 > 0$ such that with probability at least $1-\delta$,*

$$\sup_{x\in\mathcal{X}} |\hat{\mathcal{R}}(x, S_n) - \mathcal{R}(x)| \le K_1 \max\{B, L\}\left(\sqrt{\frac{\alpha}{n}} + \sqrt{\frac{\log(1/\delta) + I_\infty(S_n, \mathcal{X})}{n}}\right), \tag{14}$$

where $I_\infty(X,Y) := \lim_{\rho\to\infty} I_\rho(X,Y)$. Furthermore, there exists $K_2 > 0$ such that

$$\mathbb{E}\sup_{x\in\mathcal{X}} |\hat{\mathcal{R}}(x, S_n) - \mathcal{R}(x)| \le K_2 \max\{B, L\}\left(\sqrt{\frac{\alpha}{n}} + \sqrt{\frac{I_1(S_n, \mathcal{X})}{n}}\right). \tag{15}$$

This result shows that the generalization error is mainly bounded by two terms: the tail exponent of the limiting Lévy process L_t^α, and the statistical dependence of $\mathcal{X}$ on S_n. The bound suggests that, provided the mutual information between $\mathcal{X}$ and S_n is fixed, the generalization error can be reduced by using a smaller α in MPGD. However, it has been illustrated that a smaller α can induce significant error on $\hat{\mathcal{R}}(x, S_n)$ [SZTG20, CWZ+21], hence a reasonable value of α should be chosen to balance both the empirical risk and the generalization error. The algorithm parameters μ, σ interact with the bound via the constant terms K_1, K_2 and the mutual information terms I_1, I_∞. Unfortunately, their dependence is quite implicit and we are not able to provide quantitative estimations.

Finally, the main difference between Theorem 4.2 and [ŞSDE20, HŞKM21] is that, in our bounds we have explicit access to the characteristic exponent of L_t^α since we construct it manually through our dynamics; whereas in the prior work generic processes were used as approximations for SGD, hence their results are not as explicit.

4.3 Implicit Regularization of MPGD

To understand MPGD and its connection to the vanilla GD (3) better, we provide additional analysis for the behavior of loss functions optimized under the recursion (8) via the lens of implicit regularization [Mah12, LEHM21]. By this, we mean regularization imposed implicitly by the learning strategy, without explicitly modifying the loss. We shall achieve this by deriving an appropriate explicit regularizer through a perturbation analysis in the weak perturbation regime.

To this end, we work in the regime where the parameters μ, σ are small, keeping m fixed. We set $\mu = \mu_0\epsilon$ and $\sigma = \sigma_0\epsilon$, where $\epsilon > 0$ is a small parameter, and analyze the loss optimized using the recursion (8) in the small ϵ regime. In the sequel, we let $\overline{x}_k^{(m)}$ denote the states of the unperturbed GD, satisfying the recursion $\overline{x}_{k+1}^{(m)} = \overline{x}_k^{(m)} - \frac{1}{m}\hat{\mathcal{R}}(\overline{x}_k^{(m)}, S_n)$, for $k = 0, 1, \ldots$. To simplify notation, we shall denote $\hat{\mathcal{R}}(x) := \hat{\mathcal{R}}(x, S_n)$.

The following result relates the loss function, averaged over realizations of the injected perturbations, evolved under training with MPGD to that of unperturbed GD in the weak perturbation regime.

Theorem 4.3 (Implicit regularization)**.** *Let $m > 0$, $k \in \mathbb{N}$, and S_n be given. Assume that $r_1 = 1$ and $r_2 = d$ in Eq. (8), and $\mathbb{E}[v_1(y_k^{(1)})] = \mathbb{E}[v_2(y_k^{(2)})] = 0$ for all k. Then, for a scalar-valued loss function $\hat{\mathcal{R}}$ with $\nabla\hat{\mathcal{R}}$ having Lipschitz continuous partial derivatives in each coordinate up to order*

[3] As $\rho \to 1$, D_ρ tends to the Kullback-Leibler divergence.

three (inclusive),

$$\mathbb{E}[\hat{\mathcal{R}}(x_k^{(m)})] = \hat{\mathcal{R}}(\overline{x}_k^{(m)}) + \frac{\epsilon^2}{2}\left(tr\left(C_k^{(m)}\nabla^2\hat{\mathcal{R}}(\overline{x}_k^{(m)})\right) - \nabla\hat{\mathcal{R}}(\overline{x}_k^{(m)})^T\lambda_k^{(m)}\right) + \mathcal{O}(\epsilon^3), \quad (16)$$

as $\epsilon \to 0$, where the expectation is with respect to the realization of the $y^{(i)}$, the $\lambda_k^{(m)}$ are the vector with the lth component:

$$[\lambda_k^{(m)}]^l = \frac{1}{m}\sum_{i=1}^{k}\sum_{j=1}^{d}[\Phi_i^{(m)}]^{lj} tr\left(C_{i-1}^{(m)}\nabla^2[\nabla\hat{\mathcal{R}}(\overline{x}_{i-1}^{(m)})]^j\right),$$

the $\Phi_i^{(m)} := \prod_{j=i}^{k-1}(I - \frac{1}{m}\nabla^2\hat{\mathcal{R}}(\overline{x}_j^{(m)}))$, with the empty product taken to be the identity, and the $C_k^{(m)}$ are covariance matrices with the (p,q)-entry of $\sum_{i_1,i_2=1}^{k}\sum_{r,s=1}^{d}[\Phi_{i_1}^{(m)}]^{pr}[\Phi_{i_2}^{(m)}]^{qs}\theta_{i_1,i_2,r,s}^{(m)}$, where

$$\theta_{i_1,i_2,r,s}^{(m)} = \frac{\mu_0^2}{m^{2/\alpha_1}}\mathbb{E}[v_1(y_{i_1-1}^{(1)})v_1(y_{i_2-1}^{(1)})]\cdot[x_{i_1-1}^{(m)}]^r[x_{i_2-1}^{(m)}]^s + \frac{\sigma_0^2}{m^{2/\alpha_2}}\mathbb{E}[[v_2]^r(y_{i_1-1}^{(2)})[v_2]^s(y_{i_2-1}^{(2)})].$$

Theorem 4.3 implies that the loss function optimized under MPGD at a given iteration and m is, on average, approximately equivalent to a regularized objective functional. Moreover, $\frac{\nabla\hat{\mathcal{R}}(\overline{x}_k^{(m)})^T v_k^{(m)}}{tr\left(C_k^{(m)}\nabla^2\hat{\mathcal{R}}(\overline{x}_k^{(m)})\right)} \leq \frac{C}{m}$ for some constant C independent of m, suggesting that the trace term is the dominant explicit regularizer for large m. This explicit regularizer is solely determined by the discrete-time flow generated by the Jacobians $I - \frac{1}{m}\nabla^2\hat{\mathcal{R}}(\overline{x}_j^{(m)})$, the covariances $C_k^{(m)}$, and the Hessian of the loss function, all evaluated along the dynamics of the unperturbed GD. We can therefore expect the use of the perturbations in MPGD as a regularization mechanism should reduce the Hessian of the loss function according to the perturbation levels (μ_0, σ_0) and the correlations $\frac{1}{m^{2/\alpha_1}}\mathbb{E}[v_1(y_{i_1-1}^{(1)})v_1(y_{i_2-1}^{(1)})]$, $\frac{1}{m^{2/\alpha_2}}\mathbb{E}[v_2(y_{i_1-1}^{(2)})v_2(y_{i_2-1}^{(2)})]$. Reducing the Hessian of the loss function can lead to flatter minima in the loss landscape, which is widely believed to associate with better generalization in deep learning [KMN+16, JNM+19].

Lastly, we observe that in the weak perturbation regime and for non-convex losses, MPGD is more beneficial than GD in the sense that it reduces the Hessian of the loss, thereby promoting flatter minima. Hence, in order to observe the benefits of our scheme, we suspect that the optimization problem should be non-convex and local minima should have different curvatures.

5 Empirical Results

In this section, we illustrate the advantages of MPGD compared to other schemes on the tasks of (1) minimizing the widening valley loss, (2) regression on the Airfoil Self-Noise Dataset, and (3) classification on CIFAR-10. We also provide additional results and details in Appendix F.

5.1 Minimizing the Widening Valley Loss

We consider the problem of minimizing the widening valley loss, given by $\ell(u,v) = v^2\|u\|^2/2$. The gradient of l is given by $\nabla\ell(u,v) = (v^2u, \|u\|^2v)$, and the trace of the Hessian is $dv^2 + \|u\|^2$, where d is the dimension of u. The trace of the Hessian measures the flatness of the minima, which is monotonously changing in this case. The loss has a valley of minima with zero loss for all (u,v) with $v = 0$. The smaller the norm of u, the flatter the minimum. Vanilla GD gets stuck when it first enters this valley, and the intuition is that injecting suitable perturbations should help convergence to a flat (with small $\|u\|$) part of the valley. All (u,v) with $v = 0$ are minima, but we also need $\|u\|$ to be minimized in order to minimize the trace of the Hessian. In high dimension (large d), the GD path is biased towards making v small and not optimizing u since the direction along v is the most curved.

For the experiments, we start optimizing from the point $(u_0, 0)$, where $u_0 \sim 5\cdot\mathcal{U}(d)$ with $d = 10$, and use the learning rate $\eta = 0.01$. We study and compare the behavior of the following schemes: (i) baseline (vanilla GD), (ii) GD with uncorrelated Gaussian noise injection instead, and (iii) MPGD. Figure 2 demonstrates that MPGD can lead to successful optimization of the widening loss whereas the baseline GD and GD with Gaussian perturbations lead to poor solutions. This is in agreement with our analysis of implicit regularization for MPGD, showing that the injected perturbations effectively favor small trace of the loss Hessian, thereby biasing the solution to flatter region of the loss landscape.

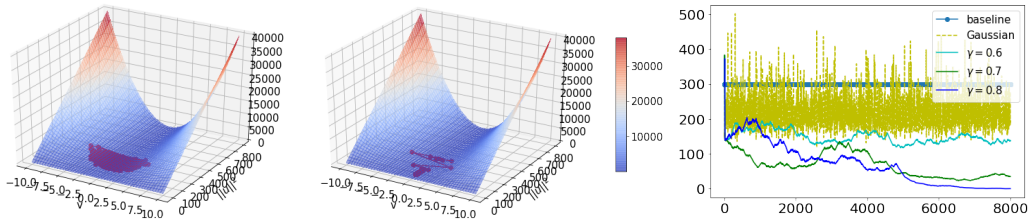

Figure 2: Evolution of GD with Gaussian perturbations (left plot) vs. that of MPGD with $\gamma = 0.7$, $\beta = 0.5$ (middle plot), using $\mu = 0.02$, $\sigma = 0.05$, and $\eta = 0.01$ ($m = 100$). Here we see that MPGD leads to successful optimization of the widening valley loss whereas that with Gaussian perturbations fails to converge. Moreover, MPGD effectively reduces the trace of loss Hessian (see right plot), steering the GD iterates to flatter region of the loss landscape.

5.2 Airfoil Self-Noise Prediction

We consider the Airfoil Self-Noise Dataset [DG17] from the UCI repository. It comprises of differently sized airfoils at various wind tunnel speeds and angles of attack. This is a regression problem aiming to predict scaled sound pressure level, in decibels, of the airfoil based on the the features: frequency, angle of attack, chord length, free-stream velocity, and suction side displacement thickness. We use 1202 samples for training and 301 samples for testing.

Table 1: Shallow neural nets trained on the airfoil data set. The results in parenthesis are achieved with the variant (11). All the results are averaged over 5 models trained with different seed values.

Scheme	test RMSE	RMSE gap
Baseline	0.4309	0.2411
Gaussian	0.4279	0.2354
MPGD, $\gamma = 0.55$	0.3916 (0.3865)	0.2256 (0.2305)
MPGD, $\gamma = 0.6$	**0.3810** (0.4092)	0.2298 (**0.2206**)
MPGD, $\gamma = 0.65$	0.3829 (0.3891)	0.2407 (0.2307)
MPGD, $\gamma = 0.7$	0.4600 (0.3754)	0.2315 (0.2311)

For the training, we use a fully connected shallow neural network of width 16 with ReLU activation and train for 3000 epochs with the learning rate $\eta = 0.1$, using mean square error (MSE) as the loss and choosing $\beta = 0.5$. Table 1 reports the average root MSE (RMSE) and the RMSE gap (defined as test RMSE - train RMSE) evaluated for models that are trained with 5 different seed values for this task. We can see that MPGD leads to both lower test RMSE and RMSE gap when compared with vanilla GD (baseline) and GD with uncorrelated Gaussian perturbations (see the results not in parenthesis in Table 1; here $\mu = 0.01$, $\sigma = 0.02$). Using the form of the perturbations in (11) instead can also give lower RMSE gap (see the results in parenthesis in Table 1; here $\sigma = \mu = 0.01$). Overall these results support our generalization theory for MPGD. Lastly, we note that using larger values of γ naively does not guarantee better test performance: one has to fine tune the parameters β, μ, σ, η appropriately to achieve favorable trade-off between training stability and test performance.

5.3 CIFAR-10 Classification

We consider training a ResNet-18 classifier [HZRS16] on the CIFAR-10 data set [Kri09]. We follow the setup used in [GGP+21], where a ResNet-18 is trained using batch size of 50K (the entire training dataset). As in [GGP+21], we consider a standard ResNet-18 with the parameters in the linear layer randomly initialized and the parameters in the batch normalization initialized with mean zero and unit variance, except for the last in each residual branch which is initialized to zero. The default random CIFAR-10 data ordering is kept as is, and every image is randomly augmented by horizontal flips and random crops after padding by 4 pixels.

Using the setup of [GGP+21], the reference mini-batch SGD is trained using a batch size of 128 (sampling without replacement), Nesterov momentum of 0.9 and weight decay of 0.0005. The learning rate is warmed up from 0.0 to 0.1 over the first 5 epochs and then reduced via cosine annealing to 0 over the course of training for 300 epochs (resulting in 390x300=117,000 update steps). The validation accuracy obtained in [GGP+21] is 95.7%. For the full-batch GDs, we replace the mini-batch updates by full batches and accumulate the gradients over all mini-batches. For the Gaussian scheme and MPGD, we inject the perturbations into the GD recursion for parameters of the input layer of ResNet-18, choosing $\mu = 0.03$, $\sigma = 0.01$ and $\beta = 0.5$ (for MPGD). In our experiments,

Table 2: ResNet-18 trained on CIFAR-10 for 1000 epochs. Here, accuracy gap = training accuracy - validation accuracy. The results in parenthesis are achieved with the variant of MPGD (11). All the results are averaged over 5 models trained with different seed values.

Scheme	val. accuracy in %	accuracy gap in %
Baseline full batch GD	73.10	8.99
Gaussian	70.58	8.53
MPGD, $\gamma = 0.55$	**75.97** (74.51)	6.08 (6.58)
MPGD, $\gamma = 0.6$	75.78 (74.60)	**5.64** (6.15)
MPGD, $\gamma = 0.65$	74.39 (72.96)	6.41 (6.85)
MPGD, $\gamma = 0.7$	73.72 (73.27)	6.69 (6.04)

we fix the learning rate to 0.1 throughout training, and do not add momentum and weight decay. This is in constrast to [GGP+21], where the additional tricks applied in the mini-batch SGD are also used for the full-batch GD, leading to higher test accuracies (see Table 2 in [GGP+21]).

Table 2 shows that MPGD can lead to better test performance and lower accuracy gap when compared to the baseline (full batch GD) and Gaussian noise-perturbed GD. Using the perturbations in (11) instead can also achieve lower accuracy gap. Therefore, adding perturbations of MPGD can be potentially a useful trick to improve training of deep architectures on benchmark data sets. Lastly, we note that additional tricks, such as gradient clipping and tuning of optimization hyperparameters, can be applied to improve our test accuracies and close the gap to results obtained with the reference mini-batch SGD. However, our main focus here is not on competitive performance, but rather on demonstrating the effects of the perturbations in MPGD when compared to full batch GDs and Gaussian noise-perturbed full-batch GD to support our theory.

Lastly, we remark that we did not tune the step-size for any of the algorithms. Since our generalization bound applies to the asymptotic SDE, which is obtained when the step-size goes to zero, in order to stay close to the theory, we chose a small enough step-size to be both not far from the continuous dynamics, and large-enough so that the algorithm converges in a reasonable amount of time. That being said, we have tried a range of step-sizes for both algorithms, and we observed that the proposed scheme consistently outperforms vanilla GD for smaller step-sizes. Whereas if we use a large step-size, both algorithms perform similarly. In this regime, vanilla GD as well potentially emits a chaotic behavior, which might also indicate the importance of the "implicit randomness". However, as we indicated in the introduction, this regime is not easily controllable, and our purpose is to introduce a controlled chaotic component, with a clear theoretical understanding.

6 Conclusion

In this paper, we introduce and study a class of slow-fast deterministic dynamical systems which homogenize to a limiting SDE driven by a heavy-tailed Lévy-stable process in an appropriate scaling limit as a rigorous framework for perturbed GDs. Within this framework, we introduce MPGD, a novel version of perturbed GD, which we show to have good generalization and regularization properties. We further demonstrate the advantages of MPGD empirically in various optimization tasks. Our framework can provide useful tools for identifying implicit randomness in deterministic optimization algorithms and inspire other promising algorithms. It would be interesting to investigate the interactions of the chaotic perturbations in MPGD with the noise arising from mini-batch sampling in SGD [WHX+20] and other optimization tricks. Since this is a theoretical paper studying a framework for perturbed GDs, there are no potential negative societal impacts of our work.

Acknowledgements. We are grateful to the computational resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2018-05973. S. H. Lim would like to acknowledge the WINQ Fellowship, the SNIC AI/ML grant, and the Swedish Research Council (VR/2021-03648) for providing support of this work. U. Ş.'s research was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), and the European Research Council Starting Grant DYNASTY – 101039676.

References

[AGZ21] Naman Agarwal, Surbhi Goel, and Cyril Zhang. Acceleration via fractal learning rate schedules. *arXiv preprint arXiv:2103.01338*, 2021.

[App09] David Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge University Press, 2009.

[BCN18] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[C+47] Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

[CFK+16] Ilya Chevyrev, Peter K Friz, Alexey Korepanov, Ian Melbourne, and Huilin Zhang. Multiscale systems, homogenization, and rough paths. In *International Conference in Honor of the 75th Birthday of SRS Varadhan*, pages 17–48. Springer, 2016.

[CFKM20] Ilya Chevyrev, Peter K Friz, Alexey Korepanov, and Ian Melbourne. Superdiffusive limits for deterministic fast–slow dynamical systems. *Probability Theory and Related Fields*, 178(3):735–770, 2020.

[CKL+21] Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *ICLR*, 2021.

[CP14] Alexei Chechkin and Ilya Pavlyukevich. Marcus versus Stratonovich for systems with jump noise. *Journal of Physics A: Mathematical and Theoretical*, 47(34):342001, 2014.

[CWZ+21] Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gürbüzbalaban, and Umut Şimşekli. Asymmetric heavy tails and implicit bias in Gaussian noise injections. *arXiv preprint arXiv:2102.07006*, 2021.

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[FV10] Peter K. Friz and Nicolas B. Victoir. *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2010.

[GAG+00] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[GGP+21] Jonas Geiping, Micah Goldblum, Phillip E Pope, Michael Moeller, and Tom Goldstein. Stochastic training is not necessary for generalization. *arXiv preprint arXiv:2109.14119*, 2021.

[GK54] Boris Vladimirovich Gnedenko and Andrey Nikolaevich Kolmogorov. *Limit Distributions for Sums of Independent Random Variables*. Cambridge, Addison-Wesley, 1954.

[GM13] Georg A Gottwald and Ian Melbourne. Homogenization for deterministic maps and multiplicative noise. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2156):20130201, 2013.

[GM21] Georg A Gottwald and Ian Melbourne. Simulation of non-Lipschitz stochastic differential equations driven by α-stable noise: A method based on deterministic homogenization. *Multiscale Modeling & Simulation*, 19(2):665–687, 2021.

[Gou04] Sébastien Gouëzel. Central limit theorem and stable laws for intermittent maps. *Probability Theory and Related Fields*, 128(1):82–122, 2004.

[HŞKM21] Liam Hodgkinson, Umut Şimşekli, Rajiv Khanna, and Michael W Mahoney. Generalization properties of stochastic optimizers via trajectory analysis. *arXiv preprint arXiv:2108.00781*, 2021.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[JNM+19] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[KMN+16] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[Kri09] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[KT20] Lingkai Kong and Molei Tao. Stochasticity of deterministic gradient descent: Large learning rate for multiscale objective function. *Advances in Neural Information Processing Systems*, 33:2625–2638, 2020.

[LEHM21] Soon Hoe Lim, N Benjamin Erichson, Liam Hodgkinson, and Michael W Mahoney. Noisy recurrent neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[Mah12] M. W. Mahoney. Approximate computation and implicit regularization for very large-scale data analysis. In *Proceedings of the 31st ACM Symposium on Principles of Database Systems*, pages 143–154, 2012.

[Mar81] Steven I Marcus. Modeling and approximation of stochastic differential equations driven by semimartingales. *Stochastics: An International Journal of Probability and Stochastic Processes*, 4(3):223–245, 1981.

[NVL+15] Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

[Oym21] Samet Oymak. Provable super-convergence with a large cyclical learning rate. *IEEE Signal Processing Letters*, 28:1645–1649, 2021.

[PS08] Grigoris Pavliotis and Andrew Stuart. *Multiscale Methods: Averaging and Homogenization*. Springer Science & Business Media, 2008.

[Sam17] G. Samoradnitsky. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. CRC Press, 2017.

[Sch98a] René L Schilling. Feller processes generated by pseudo-differential operators: On the Hausdorff dimension of their sample paths. *Journal of Theoretical Probability*, 11(2):303–330, 1998.

[Sch98b] Réné.L. Schilling. Conservativeness of semigroups generated by pseudo differntial operators. *Potential Analysis*, 9:91–104, 1998.

[Smi17] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.

[ŞSDE20] Umut Şimşekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *arXiv preprint arXiv:2006.09313*, 2020.

[SSG19] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[SZTG20] Umut Simsekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gurbuzbalaban. Fractional underdamped Langevin dynamics: Retargeting SGD with momentum under heavy-tailed gradient noise. In *International Conference on Machine Learning*, pages 8970–8980. PMLR, 2020.

[Tha80] Maximilian Thaler. Estimates of the invariant densities of endomorphisms with indifferent fixed points. *Israel Journal of Mathematics*, 37(4):303–314, 1980.

[WHX+20] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as SGD. In *International Conference on Machine Learning*, pages 10367–10376. PMLR, 2020.

[You36] L. C. Young. An inequality of the Hölder type, connected with Stieltjes integration. *Acta Mathematica*, 67(none):251 – 282, 1936.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We pointed out in Conclusion that our results only cover the setting of full-batch gradient descent, leaving out the studies for the more challenging case of mini-batch SGD. We also do not study the interactions of the chaotic perturbations in MPGD with other optimization tricks such as gradient clipping and adding momentum.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Conclusion.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See the Appendix for all the proofs and details.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the Appendix for such details.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We leave out such details in the main paper due to page constraints, and report them in the appendix instead.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Overview

Organizational Details. This Appendix is organized as follows.

- In Section B, we provide technical background materials that will be needed for proving the homogenization results.
- In Section C, we present a general theorem that includes Theorem 4.1 as a special case and provide the proof.
- In Section D, we provide proof for the generalization bound (Theorem 4.2).
- In Section E, we provide proof for the implicit regularization result (Theorem 4.3).
- In Section F, we provide additional empirical results and details.

In addition to the notations introduced in the main paper, we shall need the following notations.

Notation. Denote the space of càdlàg (right continuous with left limits) functions from $[0,1]$ to $\mathbb{R}^d$ by $D([0,1],\mathbb{R}^d)$, on which the Skorokhod-type metrics are denoted by $\sigma_\cdot$. The left limit of $f \in D([0,1],\mathbb{R}^d)$ at t is written by $f(t-) := \lim_{s\uparrow t} f(s)$. The identity function is denoted by $\mathrm{id} : t \mapsto t$, $t \in [0,1]$. For a continuous function f defined on $\mathbb{R}^d$, denote the uniform norm by $\|f\|_\infty = \sup_{x\in\mathbb{R}^d} |f(x)|$. Denote by $\mathscr{D}([0,1],\mathbb{R}^d)$ the space of admissible path-trajectory pairs in Definition B.5, whose metrics are denoted by $\alpha_\cdot$. Denote by $C^{p\text{-var}}([0,1],\mathbb{R}^d)$ the space of continuous functions with finite p-variation. In Definition B.4, denote by l_d the linear interpolation in $\mathbb{R}^d$.

B Additional Technical Background

B.1 Background on metrics and topologies on the space of càdlàg functions

B.1.1 Skorokhod-type topologies

The Skorokhod $\mathcal{J}_1$ topology on $D([0,1],\mathbb{R}^d)$ is induced by the following metric.

Definition B.1 (Skorokhod distance)**.** The Skorokhod distance on $D([0,1],\mathbb{R}^d)$, the space of càdlàg functions, is defined by

$$\sigma_\infty(X_1,X_2) = \inf \|\lambda - \mathrm{id}\|_\infty \vee \|X_1 \circ \lambda, X_2\|_\infty,$$

where the inf is taken on all increasing bijections λ from $[0,1]$ to itself.

Another important topology on $D([0,1],\mathbb{R}^d)$ is the $\mathcal{SM}_1$ topology defined as follows.

Definition B.2 (Skorokhod $\mathcal{SM}_1$ topology)**.** The Skorokhod $\mathcal{SM}_1$ topology on $D([0,1],\mathbb{R}^d)$, the space of càdlàg functions, is defined by the metric

$$d_{\mathcal{SM}_1}(X_1,X_2) = \inf \|(\lambda_1,\gamma_1) - (\lambda_2,\gamma_2)\|_\infty,$$

where the inf is taken on all pairs $(\lambda_i,\gamma_i) \in C([0,1],[0,1]\times\mathbb{R}^d)$ such that $(\lambda_i,\gamma_i)(0) = (0, X_i(0))$, $(\lambda_i,\gamma_i)(1) = (1, X_i(1))$ and $\gamma_i(t) \in [X_i(\lambda_i(t)_-), X_i(\lambda_i(t))]$, $i = 1,2$. It is equivalent to the metric on the graph of functions in D, with discontinuities connected by the straight segments.

Recall that for $1 \le p < \infty$, the *p-variation* of $u : [0,1] \to \mathbb{R}^d$ is given by

$$\|u\|_{p\text{-var}} = \sup_{0=t_0<t_1<\ldots<t_k=1} \left(\sum_{j=1}^{k} |u(t_j) - u(t_{j-1})|^p \right)^{1/p},$$

and the subspace of $D([0,1],\mathbb{R}^d)$ with finite p-variation is denoted by $D^{p\text{-var}}([0,1],\mathbb{R}^d)$. Let us define the following p-variation generalisations of the Skorokhod topology.

Definition B.3 (Skorokhod-type p-variation)**.**

$$\sigma_{p\text{-var}}(X_1,X_2) = \inf \max\{\|\lambda - \mathrm{id}\|_\infty, \|X_1 \circ \lambda - X_2\|_{p\text{-var}}\},$$

where the inf is taken on all increasing bijections λ from $[0,1]$ to itself.

B.1.2 Generalised Skorokhod topologies with interpolations

For discontinuous càdlàg functions, one can interpolate jumps with path functions.

Definition B.4 (Path function)**.** A *path function* on $\mathbb{R}^d$ is a map $\phi : J \to C([0,1],\mathbb{R}^d)$, where $J \subset \mathbb{R}^d \times \mathbb{R}^d$, for which $\phi(x,y)(0) = x$ and $\phi(x,y)(1) = y$ for all $(x,y) \in J$.

Definition B.5. A pair (X,ϕ) is called admissible if all the jumps of X are in the domain of definiton J of ϕ, i.e. $(X(t-), X(t)) \in J$ for all jump times t of X. Denote by $\mathscr{D}([0,1],\mathbb{R}^d)$ the space of admissible pairs, modulo the equivalence $(X_1,\phi_1) \sim (X_2,\phi_2)$ if $X_1 = X_2$ and $\phi_1(X_1(t-), X_1(t))$ is a reparametrization of $\phi_2(X_1(t-), X_1(t))$ for all jump times t of X_1.

Definition B.6. For admissible pair $(X,\phi) \in \mathscr{D}([0,1],\mathbb{R}^d)$, we now construct a continuous path $X^{\phi,\delta}$ as follows.

- Given a sequence $r_1, r_2, \ldots > 0$ with $r := \sum_j r_j < \infty$,
$$\text{Let } \tau : [0,1] \to [0,1+r] \text{ given by } \tau(t) = t + \sum_k \delta r_k 1_{\{t_k \le t\}}.$$
- Define an intermediate process $\hat{X} \in C([0,1+\delta r],\mathbb{R}^d)$,
$$\hat{X}(t) = \begin{cases} X(s) & \text{if } t = \tau(s) \text{ for some } s \in [0,1] \\ \phi(X(t_k-), X(t_k))\left(\dfrac{s-\tau(t_k-)}{\delta r_k}\right) & \text{if } t \in [\tau(t_k-), \tau(t_k)) \text{ for some } k. \end{cases}$$
- Finally, let $X^{\phi,\delta}(t) = \hat{X}(t(1+\delta r))$. We will drop the superscript δ if $\delta = 1$.

Definition B.7. The (pseudo)metric $\alpha_{p\text{-var}}$ on $\mathscr{D}([0,1],\mathbb{R}^d)$ is defined by
$$\alpha_{p\text{-var}}((X,\phi),(\overline{X},\overline{\phi})) := \lim_{\delta\to 0} \sigma_{p\text{-var}}(X^{\phi,\delta}, \overline{X}^{\overline{\phi},\delta})$$
independent of the choice of the series $\sum_{k=1}^{\infty} r_k$. Denote by $\mathscr{D}^{p\text{-var}}([0,1],\mathbb{R}^d)$ the subspace of $\mathscr{D}([0,1],\mathbb{R}^d)$ with finite $\alpha_{p\text{-var}}$ distance to 0. Conventionally, write
$$\alpha_\infty((X,\phi),(\overline{X},\overline{\phi})) := \lim_{\delta\to 0} \sigma_\infty(X^{\phi,\delta}, \overline{X}^{\overline{\phi},\delta})$$

Remark B.1. If ϕ is the linear path function, then α_∞ induces the $\mathcal{SM}_1$ topology on the space $D([0,1],\mathbb{R}^d)$.

B.2 Background on rough differential equations (RDEs)

Definition B.8 (Forward RDE)**.** We say that X is a solution to the *forward RDE* $\mathrm{d}X_t = b(X)_t^- \mathrm{d}W_t$ if
$$X_t = X_0 + \int_0^t b(X_{s-})\mathrm{d}W_s,$$
where the integral above denotes a limit of Riemann-Stieltjes sums with $b(X(s-))$ evaluated at the left limit points of the partition intervals:
$$\int_0^t b(X_{s-})\mathrm{d}W_s = \lim_{|\mathcal{P}|\to 0} \sum_{[s,s']\in\mathcal{P}} b(X_{s-})(W_{s'} - W_s),$$
where the $\mathcal{P}$ are partitions of $[0,t]$ into intervals, and $|\mathcal{P}|$ is the size of the longest interval.

Remark B.2. We make the following two observations.

- If W has finitely many jumps at times $0 < t_1 < \ldots < t_n \le 1$, then the forward solution of $\mathrm{d}X_t = b(X)_t^- \mathrm{d}W_t, X_0 = x$ can be obtained by solving the canonical RDE on each of the intervals on which W is continuous, i.e., $[0,t_1), [t_1,t_2), \ldots, [t_n,1)$, and requiring that at jump times t_k, $k = 1,\ldots,,n$:
$$X_{t_k} = X_{t_k-} + b(X_{t_k-})(W_{t_k} - W_{t_k-}).$$
- If in (7), we write $X_t^{(m)} := x_{\lfloor mt \rfloor}^{(m)}$ and $V_t^{(m)} = \lfloor tm \rfloor / m$, then the first equation of (7) is nothing but the forward RDE
$$\mathrm{d}X_t^{(m)} = a_m(X_t^{(m)})^- \mathrm{d}V_t^{(m)} + b_m(X_t^{(m)})^- \mathrm{d}W_t^{(m)}, \qquad X_0^{(m)} = x_0^{(m)} \in \mathbb{R}^d.$$

B.2.1 Young's integral and Marcus SDE

For the purpose of defining Young's integral for driving functions with finite q-variation, $q \geq 1$, we fix $x \in C^{p\text{-var}}([0,1],\mathbb{R}^d)$ and $y \in C^{q\text{-var}}([0,1],\mathbb{R}^{e\times d})$ with $\theta = \frac{1}{p} + \frac{1}{q} > 1$.

Definition B.9 (Young's integral [You36])**.** There exists a sequence $x^n \in C^{1\text{-var}}([0,T],\mathbb{R}^d)$ such that $x^n \to x$ in $C^{p\text{-var}}([0,T],\mathbb{R}^d)$ and a sequence $y_n \in C^{1\text{-var}}([0,T],\mathbb{R}^{e\times d})$ such that $y_n \to y$ in $C^{q\text{-var}}([0,T],\mathbb{R}^{e\times d})$. For every $s < t$, the limit of $\int_s^t y^n(u)dx^n(u)$ exists, which we denote by $\int_s^t y(u)dx(u)$ the *Young's integral* of y against x on the interval $[s,t]$. The Young's integral does not depend on the choices of sequences x^n and y^n, and we have

$$\left|\int_s^t y(u)dx(u) - y(s)(x(t)-x(s))\right| \leq \frac{1}{1-2^{1-\theta}}\|x\|_{p\text{-var};[s,t]}\|y\|_{q\text{-var};[s,t]}.$$

We record the following Young-Lóeve estimate, which will be useful for our purposes.

Proposition B.1 ([FV10] Theorem 6.8)**.** *Let $x \in C^{p\text{-var}}([0,T],\mathbb{R}^d)$ and $y \in C^{q\text{-var}}([0,T],\mathbb{R}^{e\times d})$ with $\frac{1}{p} + \frac{1}{q} > 1$. The integral path $t \to \int_0^t y(u)dx(u)$ is continuous with a finite p-variation and*

$$\left\|\int_0^{\cdot} y(u)dx(u)\right\|_{p\text{-var},[s,t]} \leq C\|x\|_{p\text{-var};[s,t]}(\|y\|_{q\text{-var};[s,t]} + \|y\|_{\infty;[s,t]})$$

Solutions to Marcus rough differential equation (RDE) are defined via Young's integrals and the linear path function defined in Definition B.6.

Definition B.10. The solution of a Marcus RDE

$$\mathrm{d}X = b(X) \diamond \mathrm{d}W, \qquad X(0) = x_0$$

is obtained by

- first solving the continuous RDE $\mathrm{d}\tilde{X} = b(\tilde{X})\mathrm{d}W^{\phi,1}$, with $\phi = l_d$ being the linear path function on $\mathbb{R}^d$;
- then the càdlàg solution path $X[0,1] \to \mathbb{R}^d$ is given by $X(t) = \tilde{X}(\tau(t))$. Recall that $\tau : [0,1] \to [0,1+r]$ given by $\tau(t) = t + \sum_k \delta r_k 1_{\{t_k \leq t\}}$.

B.3 Example of a Marcus RDE

Here is an example of the solution of a rough differential equation in the Marcus sense [CFKM20]. Let $\theta > 0$ and $W^{(m)} : [0,1] \to \mathbb{R}$ be the deterministic process which are 0 on $[0,1/2]$, θ on $[1/2+1/m,1]$ and linear on $[1/2,1/2+1/m]$. Denote by $X_t^{(m)} = (x_t^{(m)}, y_t^{(m)})$ the solution to the ordinary differential equation

$$\begin{pmatrix}\mathrm{d}x_t^{(m)}\\ \mathrm{d}y_t^{(m)}\end{pmatrix} = \begin{pmatrix}-y_t^{(m)}\\ x_t^{(m)}\end{pmatrix}\mathrm{d}W_t^{(m)}, \quad \begin{pmatrix}x_0^{(m)}\\ y_0^{(m)}\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix}$$

It is not hard to see that that $X_t^{(m)} = (\cos W_t^{(m)}, \sin W_t^{(m)})$. Therefore, we have that pointwise

$$W_t^{(m)} \to W_t = \theta 1_{t\in(\frac{1}{2},1]} \text{ and } X_t^{(m)} \to X_t = \begin{pmatrix}x_t\\ y_t\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix}1_{t\in[0,\frac{1}{2}]} + \begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix})1_{t\in(\frac{1}{2},1]}.$$

If $\theta \neq 0$, $(X_t)_{t\in[0,1]}$ fails to be the right-hand side of the limiting forward RDE since

$$\int_0^1 \begin{pmatrix}-y_{s-}\\ x_{s-}\end{pmatrix}\mathrm{d}W_s = \lim_{|\mathcal{P}|\to 0}\sum_{[s,s']\in\mathcal{P}}\begin{pmatrix}-y_{s-}\\ x_{s-}\end{pmatrix}(W_{s'} - W_s) = \begin{pmatrix}-0\\ 1\end{pmatrix} \times \theta = \begin{pmatrix}0\\ \theta\end{pmatrix} \neq X_1 - X_0.$$

Effectively, $(X_t)_{t\in[0,1]}$ is the solution to the Marcus RDE in the limit.

C Homogenization: The Rigorous Version and the Proof of Theorem 4.1

When talking about convergence of solutions to RDEs, we need to specify the sense of integration. In view of the example in Subsection B.3, it is not enough to look at the solution X as an element of $D([0,1],\mathbb{R}^d)$ – it has to be coupled with jumps of the driving function. Let us consider the driver-solution space $D([0,1],\mathbb{R}^{r+d})$ and introduce a new path function on $\mathbb{R}^{r+d}$.

Definition C.1. Consider $b \in C(\mathbb{R}^d, \mathbb{R}^{d\times r})$, $a \in C(\mathbb{R}^d, \mathbb{R}^d)$. For $x \in \mathbb{R}^d$ and $\Phi \in C^{1\text{-var}}([0,1],\mathbb{R}^r)$, let $\pi_{a,b}[x;\Phi] \in C^{1\text{-var}}([0,1],\mathbb{R}^d)$ denote the solution Π of the equation

$$\mathrm{d}\Pi = a(\Pi)\mathrm{d}t + b(\Pi)\mathrm{d}\Phi, \qquad \Pi(0) = x.$$

For the coefficient b, a pair of admissible points $(w_1,x_1),(w_2,x_2) \in J_{a,b} \subset \mathbb{R}^{r+d}$ is in the space

$$J_{a,b} = \{\big((w_1,x_1),(w_2,x_2)\big) : w_1,w_2 \in \mathbb{R}^d, \pi_b[x_1; l_d(w_1,w_2)](1) = x_2\}.$$

We define the path function $\phi_{a,b}$ on $J_{a,b}$ by

$$\phi_{a,b}((w_1,x_1),(w_2,x_2)) = \big(l_d(w_1,w_2)(t), \pi_b[x_1; l_d(w_1,w_2)](t)\big).$$

Next, we first state the precise assumptions for Theorem 4.1 in the main paper. The required topology for the coefficients a_m, b_m in (9) is as follows. For $\tilde{\gamma} > 0$, $n_1, n_2 \in \mathbb{N}_+$, denote by $C^{\tilde{\gamma}}(\mathbb{R}^{n_1},\mathbb{R}^{n_2})$ the space of functions $f : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$ such that [4]

$$\|f\|_{C^{\tilde{\gamma}}} = \max_{|\mathbf{j}|=0,\dots,\lfloor\gamma'\rfloor} \|\partial^{\mathbf{j}} f\|_\infty + \sup_{x,y\in\mathbb{R}^m} \max_{|\mathbf{j}|=\lfloor\tilde{\gamma}\rfloor} \frac{|\partial^{\mathbf{j}} f(x) - \partial^{\mathbf{j}} f(y)|}{|x-y|^{\tilde{\gamma}-\lfloor\tilde{\gamma}\rfloor}} < \infty.$$

In particular, if $\tilde{\gamma} > 1$, we have $\|f\|_\infty \le \|f\|_{C^{\tilde{\gamma}}}$ and $\|f\|_{\mathrm{Lip}} \le \|f\|_{C^{\tilde{\gamma}}}$.

Assumption 1. Suppose that $a \in C^{\gamma_1}(\mathbb{R}^d,\mathbb{R}^d)$, $b \in C^{\gamma_2}(\mathbb{R}^d,\mathbb{R}^{d\times r})$ for some $\gamma_1 > 1, \gamma_2 > \alpha$. In addition, we assume that

$$\lim_{m\to\infty} a_m = a \text{ in } C^{\gamma_1}(\mathbb{R}^d,\mathbb{R}^d), \ \lim_{m\to\infty} b_m = b \text{ in } C^{\gamma_2}(\mathbb{R}^d,\mathbb{R}^{d\times r}) \ \text{ and } \ \lim_{m\to\infty} x_0^{(m)} = x_0.$$

Assumption 1 requires that the difference between the gradients of coefficients $\|\nabla(a_m - a)\|_\infty$ goes to zero and also their Hölder constant $\frac{|\nabla(a_m-a)(x)-\nabla(a_m-a)(y)|}{|x-y|^\epsilon}$ goes to zero uniformly on $x, y \in \mathbb{R}^m$ for some $\epsilon > 0$. This translates to the assumptions that the loss is second differentiable and its Hessian is ϵ-Hölder, which are reasonable assumptions. Similarly, for the coefficients appearing before the α-stable process (for $\alpha \in (1,2)$), it is required that $\|\nabla(b_m - b)\|_\infty$ goes to zero and $\sup_{x,y\in\mathbb{R}^m} \frac{|\nabla(b_m-b)(x)-\nabla(b_m-b)(y)|}{|x-y|^{\epsilon'}}$ goes to zero for some $\epsilon' > \alpha - 1$. In the setup for MPGD (8) where $a_m = -\hat{\mathcal{R}}(x, S_n) = a$ (independent of m) and the b_m is $-\mu\,\mathrm{diag(x)}$ or σI, it is easy to see that these assumptions are satisfied.

In terms of how general the assumption is, we note that Assumption 1 can cover more general situations; e.g., the case where is an empirical loss (taking $m = n$, the number of training samples) and a is the population loss, and therefore our framework can be adapted for analysis in other settings as well.

Theorem 4.1 is an extension of Theorem 4.1 in [GM21] (with the ϵ there equal $1/m$) to the case where the coefficients a, b are dependent on the scaling parameter m. In particular, the proof of Theorem 4.1 in [GM21] relies on verifying the hypotheses of Theorem 2.6 in [CFKM20] for the particular example of the observable and Thaler map constructed for the MPGD (8).

We now prove a more general theorem that extends Theorem 2.6 in [CFKM20] and includes Theorem 4.1 as a special case, stating all the needed assumptions. Applying this theorem together with the hypotheses verification in the proof of Theorem 4.1 in [GM21] completes the proof of Theorem 4.1.

Theorem C.1 (A more general version of Theorem 4.1)**.** *Let $\alpha \in (1,2)$, $\gamma_1 > 1$ and $\gamma_2 > \alpha$. Let $v \in L^\infty(Y,\mathbb{R}^d)$ be Hölder such that $\int v\mathrm{d}\mu = 0$, where μ is the unique ergodic invariant probability*

[4] For any $n \in \mathbb{N}$, $\mathbf{j} = (j_1,\dots,j_n) \in \mathbb{N}_0^n$ is a multi-index with $|\mathbf{j}| = \sum_{i=1}^n j_i$ and the higher-order partial derivative is defined by $\partial^{\mathbf{j}} = (\partial/\partial_{x_1}^{j_1})\cdots(\partial/\partial_{x_n}^{j_n})$.

measure of $T : Y \to Y$. We focus on the case where T exhibits superdiffusive behavior, which means that there exists a r-dimensional Lévy process L *and*

$$W_n(t) = n^{-1/\alpha} \sum_{j=0}^{\lfloor nt \rfloor - 1} v \circ T^j \overset{(\mathrm{d})}{\to} \mathrm{L}_t \text{ in } D([0,1], \mathbb{R}^r) \text{ under the } \mathcal{SM}_1 \text{ topology (Definition B.2)}$$

as $n \to \infty$.

If, in addition,

- $\lim_{n\to\infty} a_n = a$ *in* $C^{\gamma_1}(\mathbb{R}^d, \mathbb{R}^d)$ *for some* $a \in C^{\gamma_1}(\mathbb{R}^d, \mathbb{R}^d)$*,*
- $\lim_{n\to\infty} b_n = b$ *in* $C^{\gamma_2}(\mathbb{R}^d, \mathbb{R}^{d\times r})$ *for some* $b \in C^{\gamma_2}(\mathbb{R}^d, \mathbb{R}^{d\times r})$*;*
- $\lim_{n\to\infty} x_n = x$*,*
- $\|W_n\|_{p\text{-var}}$ *is tight for all* $p > \alpha$ *and* $\sum_t |W_n(t) - W_n(t-)|^2 \to 0$ *a.s., where the sum is taken over all jump times of* W_n*,*

then for the forward RDE $\mathrm{d}X_n = a_n(X_n)^- \mathrm{d}V_n + b_n(X_n)^- \mathrm{d}W_n, X_n(0) = x_n$ *where* $V_n(t) = \frac{\lfloor tn \rfloor}{n}$*, we have:*

$$((W_n, X_n), l_{r+d}) \overset{(\mathrm{d})}{\to} ((\mathrm{L}, X), \phi_{a,b}) \text{ in } (\mathscr{D}^{p\text{-var}}([0,1], \mathbb{R}^{r+d}), \alpha_{p\text{-var}}) \text{ as } n \to \infty$$

for all $p > \alpha$, where $(X(t))_{t\geq 0}$ is the solution of the Marcus differential equation

$$\mathrm{d}X = a(X)\mathrm{d}t + b(X) \diamond \mathrm{dL} \text{ with } X(0) = x.$$

Remark C.1. With the choices of v and T in Section 3, the assumptions in Theorem C.1 are satisfied (see the proof in [GM21]).

Before proving Theorem C.1, we start by proving the following lemma.

Lemma C.1. *Let $X \in D([0,1], \mathbb{R}^d)$, $p \geq 1$ such that $\|X\|_{p\text{-var}} < \infty$. Then for any path function ϕ,*

$$\|X^{\phi,\delta}\|_{p\text{-var}} = \|X^{\phi}\|_{p\text{-var}} \geq \|X\|_{p\text{-var}} = \|X^{l_d}\|_{p\text{-var}}.$$

Proof. It is obvious from the scaling invariance of of the p-variation that $\|X^{\phi,\delta}\|_{p\text{-var}} = \|X^{\phi}\|_{p\text{-var}}$. For the second inequality, using again the definition,

$$\begin{aligned}
&\|X^{\phi}\|_{p\text{-var}} \\
&= \|\hat{X}(\cdot(1+r))\|_{p\text{-var}} \\
&= \sup_{0=t_0<t_1<\ldots<t_k=1} \left(\sum_{j=1}^{k} |\hat{X}(t_j(1+r)) - \hat{X}(t_{j-1}(1+r))|^p \right)^{1/p} \\
&\geq \sup_{0=t_0<t_1<\ldots<t_k=1, t_j(1+r)=\tau(s_j) \text{ for some } s_j\in[0,1]} \left(\sum_{j=1}^{k} |\hat{X}(t_j(1+r)) - \hat{X}(t_{j-1}(1+r))|^p \right)^{1/p} \\
&= \sup_{0=s_0<s_1<\ldots<s_k=1} \left(\sum_{j=1}^{k} |X(s_j) - X(s_{j-1})|^p \right)^{1/p} = \|X\|_{p\text{-var}},
\end{aligned}$$

where the last equality follows from the fact that if c lies in an interval $[a, b]$ then $(c-a)^p + (b-c)^p \leq (b-a)^p$ for $p \geq 1$. This completes the proof. □

As a first step for proving Theorem C.1, we aim to prove a deterministic variant of Theorem C.1. In order to make the arguments more concise, let us neglect the first time-derivative term (with coefficients a and a_n, for which the treatment is the same by considering the jump process $V_n : t \mapsto \lfloor nt \rfloor / n$) and consider the following.

Theorem C.2. *Assume that $p \in (1,2)$ and $\{W_n\}_{n\geq 1}$ is a sequence in $D^{p\text{-var}}([0,1],\mathbb{R}^r)$ with finitely many jumps. Suppose that $\gamma > p$ and*

- $\lim_{n\to\infty} b_n = b$ *in* $C^\gamma(\mathbb{R}^d,\mathbb{R}^{d\times r})$ *for some* $b \in C^\gamma(\mathbb{R}^d,\mathbb{R}^{d\times r})$*;*
- $d_{\mathcal{SM}_1}(W_n, W) \to 0$ *and* $\sum |W_n(t) - W_n(t-)|^2 \to 0$*, where the sum is taken over all jump times of* W_n*.*

Let X_n be the solution of the forward RDE $\mathrm{d}X_n = b_n(X_n)^- \mathrm{d}W_n$ with $X(0) = x_n$. Let X be the solution of the Marcus RDE $\mathrm{d}X = b(X) \diamond \mathrm{d}W, X(0) = x$. Then it holds that

$$((W_n, X_n), l_{r+d}) \to ((W,X),\phi_b) \text{ in } (\mathscr{D}([0,1],\mathbb{R}^{r+d}),\alpha_{p\text{-var}}) \text{ as } n\to\infty.$$

C.1 Proof of Theorem C.2

We will use the following lemma in the proof of Theorem C.2.

Lemma C.2 ([CFKM20] Lemma 3.6)**.** *Suppose that $W \in D([0,1],\mathbb{R}^r)$ has finitely many jumps. Let $X, \tilde{X} \in D([0,1],\mathbb{R}^d)$ be given by*

$$\mathrm{d}X = b(X)^- \mathrm{d}W, \quad \mathrm{d}\tilde{X} = b(\tilde{X}) \diamond \mathrm{d}W, \quad X(0) = \tilde{X}(0) = x.$$

Then

$$\|X - \tilde{X}\|_{p\text{-var}} \leq K\|b\|_{\mathrm{Lip}}\|b\|_\infty \sum_t |W(t) - W(t-)|^2,$$

where K depends on $\|b\|_{C^\gamma}, \|W\|_{p\text{-var}}, \gamma, p$, and the sum is over all jump times t of W.

We are now ready to prove Theorem C.2.

Proof of Theorem C.2. Let $\tilde{X}_n$ be the solution to the Marcus RDE

$$\mathrm{d}\tilde{X}_n = b(\tilde{X}_n) \diamond \mathrm{d}W_n, \tilde{X}_n(0) = x_n.$$

The continuity of the solution map for generalised geometric RDEs gives that

$$\alpha_{p\text{-var}}(((W_n, \tilde{X}_n), \phi_b), ((W,X),\phi_b)) = 0. \tag{17}$$

Then let $\overline{X}_n$ be the solution to the Marcus RDE

$$\mathrm{d}\overline{X}_n = b_n(\overline{X}_n) \diamond \mathrm{d}W_n, \quad \overline{X}_n(0) = x_n.$$

On any subinterval $[s_1, s_2]$, we can compare solutions to the Marcus RDE with b_n and b by using Proposition B.1:

$$\begin{aligned}
&\|\overline{X}_n - \tilde{X}_n\|_{p\text{-var}} \\
=&\left\| \int_0^\cdot (b_n(\overline{X}_n) - b(\tilde{X}))\mathrm{d}W^{l_r} \right\|_{p\text{-var}} \\
\leq&\left\| \int_0^\cdot (b_n(\overline{X}_n) - b(\overline{X}_n))dW^{l_r} \right\|_{p\text{-var}} + \left\| \int_0^\cdot (b(\overline{X}_n) - b(\tilde{X}_n))dW^{l_r} \right\|_{p\text{-var}} \\
\leq&\|b_n - b\|_\infty \|W^{l_r}\|_{p\text{-var}} + C_1(\|b\|_{\mathrm{Lip}}\|\overline{X}_n - \tilde{X}_n\|_{p\text{-var}} + |b(\overline{X}_n) - b(\tilde{X}_n)|(s_1))\|W^{l_r}\|_{p\text{-var}} \\
=&\|b_n - b\|_\infty \|W\|_{p\text{-var}} + C_1(\|b\|_{\mathrm{Lip}}\|\overline{X}_n - \tilde{X}_n\|_{p\text{-var}} + |b(\overline{X}_n) - b(\tilde{X}_n)|(s_1))\|W\|_{p\text{-var}},
\end{aligned}$$

where the last inequality is due to Lemma C.1. If we choose a subdivision $0 = s_0 < s_1 < \ldots < s_{n-1} < s_n = 1$ of the interval $[0,t]$ such that for any $i = 0, 1, \ldots, n-1$,

$$C_1\|b\|_{\mathrm{Lip}}\|W\|_{p\text{-var},[s_i,s_{i+1}]} \leq c < 1,$$

then on the interval $[0, s_1]$,

$$\begin{aligned}
\|\overline{X}_n - \tilde{X}_n\|_{p\text{-var},[0,s_1]} &\leq \frac{\|b_n - b\|_\infty \|W\|_{p\text{-var},[0,s_1]} + C_1\|W\|_{p\text{-var},[0,s_1]}|b(\overline{X}_n) - b(\tilde{X}_n)|(0)}{1 - C_1\|b\|_{\mathrm{Lip}}\|W\|_{p\text{-var},[0,s_1]}} \\
&\leq \|b_n - b\|_\infty \|W\|_{p\text{-var},[0,s_1]}/(1-c).
\end{aligned}$$

Similarly on the interval $[s_1, s_2]$,

$$\begin{aligned}
&\|\overline{X}_n - \tilde{X}_n\|_{p\text{-var},[s_1,s_2]} \\
\leq &\frac{\|b_n - b\|_\infty \|W\|_{p\text{-var},[s_1,s_2]} + C_1 \|W\|_{p\text{-var},[s_1,s_2]} |b(\overline{X}_n) - b(\tilde{X}_n)|(s_1)}{1 - C_1 \|b\|_{\text{Lip}} \|W\|_{p\text{-var},[s_1,s_2]}} \\
\leq &\left(\|b_n - b\|_\infty \|W\|_{p\text{-var},[s_1,s_2]} + C_1 \|W\|_{p\text{-var},[s_1,s_2]} \|b\|_{\text{Lip}} \|\overline{X}_n - \tilde{X}_n\|_{p\text{-var},[0,s_1]} \right) / (1-c) \\
\leq &\|b_n - b\|_\infty \|W\|_{p\text{-var},[s_1,s_2]} / (1-c) + C_1 \|b\|_{\text{Lip}} \|b_n - b\|_\infty \|W\|_{p\text{-var},[0,s_1]} \|W\|_{p\text{-var},[s_1,s_2]} / (1-c)^2.
\end{aligned}$$

It is not hard to obtain by induction that there exists $C(b, W) > 0$ such that

$$\|\overline{X}_n - \tilde{X}_n\|_{p\text{-var},[0,1]} \leq C(b, W) \|b_n - b\|_\infty.$$

Therefore

$$\alpha_{p\text{-var}}(((W_n, \tilde{X}_n), \phi_b), ((W_n, \overline{X}_n), \phi_{b_n})) = 0. \tag{18}$$

Since $b_n \to b$ in $C^\gamma(\mathbb{R}^d, \mathbb{R}^{d\times r})$, $\|b_n\|_{\text{Lip}}$ and $\|b_n\|_\infty$ are uniformly bounded independent of n. Then it follows from Lemma C.2 that

$$\|X_n - \overline{X}_n\|_{p\text{-var}} \leq K \|b_n\|_{\text{Lip}} \|b_n\|_\infty \sum_t \|W_n(t) - W_n(t-)\| \to 0 \text{ as } n \to \infty.$$

Then the same argument in [CFKM20] gives that for any $p' > p$,

$$\lim_{n\to\infty} \alpha_{p'\text{-var}}(((W_n, \overline{X}_n), l_{r+d}), ((W_n, X_n), l_{r+d})) = 0. \tag{19}$$

Recall also from the [CFKM20] that

$$\lim_{n\to\infty} \alpha_{p'\text{-var}}(((W_n, \overline{X}_n), \phi_{b_n}), (W_n, \overline{X}_n), l_{r+d})) = 0. \tag{20}$$

Finally Theorem C.2 follows from combining (17), (18), (19) and (20). This completes the proof. □

C.2 Proof of Theorem C.1

Proof of Theorem C.1. If we write $V_n(t) = \lfloor tn \rfloor / n$, the assumptions of Theorem C.1 and the same treatment in [CFKM20] implies that up to subtracting a subsequence, $(V_n, W_n) \to (\text{id}, \text{L})$ a.s. in $\alpha_{p\text{-var}}$ for any $p > \alpha$. Then it follows from Theorem C.2 that for $p' > p$ and along each subsequencial limit of (V_n, W_n) as $n \to \infty$,

$$((W_n, X_n), l_{r+d}) \to ((\text{L}, X), \phi_{a,b}) \text{ in } (\mathcal{D}([0,1], \mathbb{R}^{r+d}), \alpha_{p'\text{-var}}).$$

Therefore $((W_n, X_n), l_{r+d}) \overset{(\text{d})}{\to} ((\text{L}, X), \phi_{a,b})$ in $(\mathcal{D}([0,1], \mathbb{R}^{r+d}), \alpha_{p\text{-var}})$ for any $p > \alpha$. This completes the proof.

□

D Generalization Bound: Proof of Theorem 4.2

To begin with, we need a geometric regularity assumption over the trajectory of the multiscale perturbed gradient flow, which is common for random fractal processes given as solutions to stochastic differential equations; see [HŞKM21]. This assumption ensures that the box-counting (Minkowski) dimension coincides with the Hausdorff dimension of the trajectory.

For any $x \in \mathbb{R}^d$, let $(X_t)_{t\in[0,1]}$ be a solution to the stochastic differential equation (12) started from x, and let $\mathcal{A}$ be the infinitesimal generator of $(X_t)_{t\in[0,1]}$ (let us take it for granted that it exists) defined by

$$\mathcal{A}u(x) = \lim_{t\to 0} \frac{\mathbb{E}_x[u(X_t)] - u(x)}{t} \text{ for any } u \in C_c^\infty(\mathbb{R}^d).$$

Let $q : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{C}$ be the *symbol* of $\mathcal{A}$ such that $\mathcal{A}u(x) = -\int_{\mathbb{R}^m} e^{i\langle \xi, x\rangle} q(x, \xi) \hat{u}(\xi) \mathrm{d}\xi$, (informally saying, $-q(x, D)u(x)$), where $\hat{u}(\xi) = \int_{\mathbb{R}^m} e^{-i\langle \xi, x\rangle} u(x) \mathrm{d}x$ is the Fourier transform of u.

Assumption 2. For almost every $\mathcal{X}$, there exists a finite Borel measure μ on $\mathcal{X}$ and $\rho > 0$ such that $C_\rho := \inf_{0<r<\rho, x\in\mathcal{X}} \frac{\mu(B_r(x))}{r^\alpha \mu(\mathcal{X})} > 0$ for μ-almost every x.

If $(X_t)_{t\geq 0}$ is a solution with $X_0 = 0$ to $\mathrm{d}X_t = b \diamond \mathrm{dL}_t^\alpha$ for fixed $b \in \mathbb{R}^{d\times d}$ and a symmetric d-dimensional Lévy process L^α, then $(X_t)_{t\geq 0}$ is also a α-stable Lévy process with symbol independent of x:

$$\psi_X(\xi) = |\xi^T b|^\alpha.$$

We also need the following statistical regularity assumption, as discussed in the paragraph before Theorem 4.2.

Assumption 3. There exists $x^\star \in \mathbb{R}^d$ such that $a(x^*) = 0$. Let $(X_t^*)_{t\geq 0}$ to the solution of $\mathrm{d}X_t = b(x^*) \diamond \mathrm{dL}_t^\alpha$ and write $\tilde{q}(x,\xi) := q(x,\xi) - \psi_{X^*}(\xi)$. For almost-every S_n:

- $b(x^\star)$ is positive definite;
- $|\partial_x^{\mathbf{j}} \tilde{q}(x,\xi)| \leq \Phi_{\mathbf{j}}(x)(1+\kappa_0|\xi|^\alpha)$ with $\mathbf{j} \in \mathbb{N}_0^m$, $|\mathbf{j}| \leq m+1$ for some $\Phi_{\mathbf{j}} \in L^1(\mathbb{R}^m)$;
- $q(x,0) = 0$ and $\|\Phi_0\|_\infty < \infty$.

Some remarks on Assumption 3 are now in place. It is natural to assume the existence of x^* such that $a(x^*) = 0$ (since $a = -\nabla\hat{\mathcal{R}}(x, S_n)$ in our MPGD setup). For the first point, the positive definiteness of $b(x^*)$ can be satisfied by simply choosing b to be the identity map. For the second point, recall that $\tilde{q}(x,\xi) = q(x,\xi) - \psi_{x^*}(\xi)$ and we require that: $|\partial_x^{\mathbf{j}} \tilde{q}(x,\xi)| \leq \Phi_{\mathbf{j}}(x)(1+\kappa_0|\xi|^\alpha)$ with $\mathbf{j} \in \mathbb{N}_0^m$, $|\mathbf{j}| \leq m+1$ for some $\Phi_{\mathbf{j}} \in L^1(\mathbb{R}^m)$.

Here X^* is an α-stable Lévy process, and its characteristic function $\psi_{x^*}(\xi)$ is given by $|\xi^T b(x^*)|^\alpha$ (see Eq. (4)). Moreover, $q(x,\xi)$ is nothing but (4) with the b, Σ, μ depending on x. It is not hard to see that the above assumption holds if the b, Σ, μ depend smoothly on x. The third point is equivalent to saying that the solution to the SDE exists almost surely on infinite time interval (see [Sch98b]), which is the case for the perturbed gradient flow with respect to second differentiable loss.

With all these ingredients in place, we now prove Theorem 4.2.

Proof. We observe that the ellipticity condition holds since $b(x^\star)$ is positive definite.:

$$1 + \psi_{X^*}(\xi) \geq \gamma_1(1+\kappa_0|\xi|^\alpha) \text{ for some } \gamma_1,\ \kappa_0 > 0.$$

One can also check that $\alpha := \inf\{\lambda \geq 0 : \lim_{|\xi|\to\infty} \frac{|\psi_{X^*}(\xi)|}{|\xi|^\lambda} = 0\}$. Under Assumption 3, it follows from [Sch98a, Theorem 4] that

$$\dim_{\mathrm{H}}(\mathcal{X}) \leq \alpha \quad \text{almost surely,}$$

where $\dim_{\mathrm{H}}$ denotes the Hausdorff dimension.

Then, under Assumption 2, the result follows from [HŞKM21, Theorem 1] and [HŞKM21, Corollary 1]. This completes the proof. □

E Implicit Regularization: Proof of Theorem 4.3

In this section, we provide proof to Theorem E.

Let $r_1 = 1$, $r_2 = d$ as assumed in Theorem 4.3. Upon the rescalings $\mu = \mu_0\epsilon$ and $\sigma = \sigma_0\epsilon$ with $\epsilon > 0$ a small parameter, the MPGD recursion Eq. (8) becomes:

$$x_{k+1}^{(m)} = x_k^{(m)} + f(x_k^{(m)}) + \epsilon g(x_k^{(m)}, y_k^{(1)}, y_k^{(2)}), \tag{21}$$

for $k = 0, 1, 2, \ldots$, where

$$f(x_k^{(m)}) = -\frac{1}{m}\nabla\hat{\mathcal{R}}(x_k^{(m)}) =: f_k^{(m)}, \tag{22}$$

$$g(x_k^{(m)}, y_k^{(1)}, y_k^{(2)}) = -\frac{\mu_0}{m^{\frac{1}{\alpha_1}}} v_1(y_k^{(1)}) x_k^{(m)} + \frac{\sigma_0}{m^{\frac{1}{\alpha_2}}} v_2(y_k^{(2)}) =: g_k^{(m)}, \tag{23}$$

with the $y^{(i)}$ satisfying $y_{k+1}^{(i)} = f^{(i)}(y_k^{(i)})$ for $i = 1, 2$.

Consider the following hierarchy of recursive equations. For $k = 0, 1, 2, \ldots$:

$$\overline{x}_{k+1}^{(m)} = \overline{x}_k^{(m)} - \frac{1}{m}\nabla\hat{\mathcal{R}}(\overline{x}_k^{(m)}), \tag{24}$$

$$\phi_{k+1}^{(m)} = J_k^{(m)}\phi_k^{(m)} + g_k^{(m)}, \tag{25}$$

$$\varphi_{k+1}^{(m)} = J_k^{(m)}\varphi_k^{(m)} + \frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2 f_k^{(m)}}{\partial x^i \partial x^j}(\overline{x}_k^{(m)}, \phi_k^{(m)})[\phi_k^{(m)}]^i[\phi_k^{(m)}]^j + \sum_{i=1}^{d}\frac{\partial g_k^{(m)}}{\partial x^i}(\overline{x}_k^{(m)})[\phi_k^{(m)}]^i, \tag{26}$$

with the initial conditions $\overline{x}_0^{(m)} = x_0^{(m)}$, $\phi_0^{(m)} = \varphi_0^{(m)} = 0$, and

$$J_k^{(m)} := I + f'(\overline{x}_k^{(m)}) = I - \frac{1}{m}\nabla^2\hat{\mathcal{R}}(\overline{x}_k^{(m)}). \tag{27}$$

Informally, the processes $\overline{x}^{(m)}$, $\phi^{(m)}$ and $\varphi^{(m)}$ constitute the zeroth-, first- and second-order terms in a pathwise Taylor expansion of $x^{(m)}$ about $\epsilon = 0$, i.e., for $k = 0, 1, 2, \ldots$, $x_k^{(m)} = \overline{x}_k^{(m)} + \epsilon\phi_k^{(m)} + \epsilon^2\varphi_k^{(m)} + \mathcal{O}(\epsilon^3)$, as $\epsilon \to 0$.

Now we are going to compute a second-order Taylor expansion of $\hat{\mathcal{R}}(x_k^{(m)})$ about $\epsilon = 0$. This is the content of the following lemma, which is adapted from Theorem 9 in [LEHM21] to our setting.

Lemma E.1. *Under the assumptions of Theorem 4.3, we have:*

$$\hat{\mathcal{R}}(x_k^{(m)}) = \hat{\mathcal{R}}(\overline{x}_k^{(m)}) + \epsilon\nabla\hat{\mathcal{R}}(\overline{x}_k^{(m)})^T\phi_k^{(m)} + \epsilon^2\left(\hat{\mathcal{R}}(\overline{x}_k^{(m)})^T\varphi_k^{(m)} + \frac{1}{2}(\phi_k^{(m)})^T\nabla^2\hat{\mathcal{R}}(\overline{x}_k^{(m)})\phi_k^{(m)}\right) + \mathcal{O}(\epsilon^3), \tag{28}$$

as $\epsilon \to 0$, where the $(\overline{x}_k^{(m)}, \phi^{(m)}, \varphi^{(m)})$ satisfy Eq. (24)-(26).

We now prove Theorem 4.3.

Proof of Theorem 4.3. Let us fix the m and S_n. To prove Theorem 4.3, we need to compute $\mathbb{E}\hat{\mathcal{R}}(x_k^{(m)})$, where the expectation is with respect to the randomness in the $y_0^{(i)}$. Since, $\mathbb{E}[v_1(y_k^{(1)})] = \mathbb{E}[v_2(y_k^{(2)})] = 0$ for all k by assumption, we have $\mathbb{E}\phi_k^{(m)} = 0$ for all k, and applying Lemma E.1:

$$\mathbb{E}\hat{\mathcal{R}}(x_k^{(m)}) = \hat{\mathcal{R}}(\overline{x}_k^{(m)}) + \epsilon^2\left(\hat{\mathcal{R}}(\overline{x}_k^{(m)})^T\mathbb{E}\varphi_k^{(m)} + \frac{1}{2}\mathbb{E}\left[(\phi_k^{(m)})^T\nabla^2\hat{\mathcal{R}}(\overline{x}_k^{(m)})\phi_k^{(m)}\right]\right) + \mathcal{O}(\epsilon^3). \tag{29}$$

It remains to compute the $\mathbb{E}\varphi_k^{(m)}$ and $\mathbb{E}\left[(\phi_k^{(m)})^T\nabla^2\hat{\mathcal{R}}(\overline{x}_k^{(m)})\phi_k^{(m)}\right]$ in the above expansion.

Iterating Eq. 25, we obtain

$$\phi_k^{(m)} = \sum_{i=1}^{k}\left(\prod_{j=i}^{k-1} J_j^{(m)}\right) g_{i-1}^{(m)}, \tag{30}$$

for $k = 1, 2, \ldots$. Similarly, iterating Eq. 26, we obtain:

$$\varphi_k^{(m)} = \sum_{i=1}^{k}\left(\prod_{j=i}^{k-1} J_j^{(m)}\right)\left[\frac{1}{2}\sum_{p,q=1}^{d}\frac{\partial^2 f_{i-1}^{(m)}}{\partial x^p \partial x^q}(\overline{x}_{i-1}^{(m)})[\phi_{i-1}^{(m)}]^p[\phi_{i-1}^{(m)}]^q + \sum_{l=1}^{d}\frac{\partial g_{i-1}^{(m)}}{\partial x^l}(\overline{x}_{i-1}^{(m)})[\phi_{i-1}^{(m)}]^l\right]. \tag{31}$$

Now, we compute: $\frac{\partial g_{i-1}^{(m)}}{\partial x^l}(\overline{x}_{i-1}^{(m)}) = -\frac{\mu_0}{m^{\frac{1}{\alpha_1}}}v_1(y_{i-1}^{(1)})$ for all l, and

$$\frac{\partial^2 f_{i-1}^{(m)}}{\partial x^p \partial x^q}(\overline{x}_{i-1}^{(m)}) = -\frac{1}{m}\frac{\partial^2\nabla\hat{\mathcal{R}}}{\partial x^p \partial x^q}(\overline{x}_{i-1}^{(m)}). \tag{32}$$

Substituting the above formula into Eq. (31) and then taking expectation, we obtain, using the fact that $\mathbb{E}v_1(y_{i-1}^{(1)}) = 0$ for all i:

$$\mathbb{E}\varphi_k^{(m)} = -\frac{1}{2m}\sum_{i=1}^{k}\left(\prod_{j=i}^{k-1} J_j^{(m)}\right)\sum_{p,q=1}^{d}\frac{\partial^2\nabla\hat{\mathcal{R}}}{\partial x^p\partial x^q}(\overline{x}_{i-1}^{(m)})\cdot\mathbb{E}[[\phi_{i-1}^{(m)}]^p[\phi_{i-1}^{(m)}]^q] =: -\frac{1}{2}\lambda_k^{(m)}, \quad (33)$$

with the lth component of:

$$-\frac{1}{2m}\sum_{i=1}^{k}\sum_{r=1}^{d}\left[\prod_{j=i}^{k-1} J_j^{(m)}\right]^{lr}\sum_{p,q=1}^{d}\frac{\partial^2\nabla[\hat{\mathcal{R}}]^r}{\partial x^p\partial x^q}(\overline{x}_{i-1}^{(m)})\cdot\mathbb{E}[[\phi_{i-1}^{(m)}]^p[\phi_{i-1}^{(m)}]^q]. \quad (34)$$

Now, for $k = 0, 1, 2, \ldots$, one can compute, using the assumption that $y_0^{(1)}$ and $y_0^{(2)}$ are independent (and thus $\mathbb{E}[v_1(y_{i_1-1}^{(1)})[v_2]^r(y_{i_2-1}^{(2)})] = 0$ for all r, i_1, i_2),

$$\begin{aligned}\mathbb{E}[[\phi_k^{(m)}]^p[\phi_k^{(m)}]^q] = \sum_{i_1,i_2=1}^{k}\sum_{r,s=1}^{d}[\Phi_{i_1}^{(m)}]^{pr}[\Phi_{i_2}^{(m)}]^{qs}\Bigg(&\frac{\mu_0^2}{m^{2/\alpha_1}}\mathbb{E}[v_1(y_{i_1-1}^{(1)})v_1(y_{i_2-1}^{(1)})]\cdot[x_{i_1-1}^{(m)}]^r[x_{i_2-1}^{(m)}]^s \\ &+\frac{\sigma_0^2}{m^{2/\alpha_2}}\mathbb{E}[[v_2]^r(y_{i_1-1}^{(2)})[v_2]^s(y_{i_2-1}^{(2)})]\Bigg), \quad (35)\end{aligned}$$

where the $\Phi_i^{(m)} := \prod_{j=i}^{k-1}(I - \frac{1}{m}\nabla^2\hat{\mathcal{R}}(\overline{x}_j^{(m)}))$, with the empty product taken to be the identity by convention.

The lth component of $\lambda_k^{(m)}$ can thus be written as:

$$[\lambda_k^{(m)}]^l = \frac{1}{m}\sum_{i=1}^{k}\sum_{j=1}^{d}[\Phi_i^{(m)}]^{lj}tr\left(C_{i-1}^{(m)}\nabla^2[\nabla\hat{\mathcal{R}}(\overline{x}_{i-1}^{(m)})]^j\right), \quad (36)$$

where the $C_{i-1}^{(m)}$ are covariance matrices with the (p, q)-entry of $\mathbb{E}[[\phi_{i-1}^{(m)}]^p[\phi_{i-1}^{(m)}]^q]$ whose expression is given in Eq. 35.

Lastly, we compute:

$$\begin{aligned}\mathbb{E}\left[(\phi_k^{(m)})^T\nabla^2\hat{\mathcal{R}}(\overline{x}_k^{(m)})\phi_k^{(m)}\right] &= \mathbb{E}\left[\sum_{p,q=1}^{d}[\phi_k^{(m)}]^p[\nabla^2\hat{\mathcal{R}}(\overline{x}_k^{(m)})]^{pq}[\phi_k^{(m)}]^q\right] \quad (37)\\ &= \sum_{p,q=1}^{d}\mathbb{E}[[\phi_k^{(m)}]^p[\phi_k^{(m)}]^q]\cdot[\nabla^2\hat{\mathcal{R}}(\overline{x}_k^{(m)})]^{pq} \quad (38)\\ &= tr\left(C_k^{(m)}\nabla^2\hat{\mathcal{R}}(\overline{x}_k^{(m)})\right). \quad (39)\end{aligned}$$

Substituting (36) and (39) into Eq. (29), we arrive at Eq. (16) in Theorem 4.3. □

F Additional Empirical Results and Details

In this section, we provide additional empirical results, as well as the details and additional results for the experiments considered in the main paper, to strengthen the support for our theory.

F.1 Electrocardiogram (ECG) Classification

We consider the Electrocardiogram (ECG) binary classification task that aims to discriminate between normal and abnormal heart beats of a patient that has severe congestive heart failure [GAG+00]. We use 500 sequences of length 140 for training, and 4000 sequences for testing. We use a fully connected shallow neural network of width 32 with sigmoid activation and train for 1000 epochs, with the binary cross-entropy as the loss. We choose $\eta = 0.05$. For the MPGD and the Gaussian

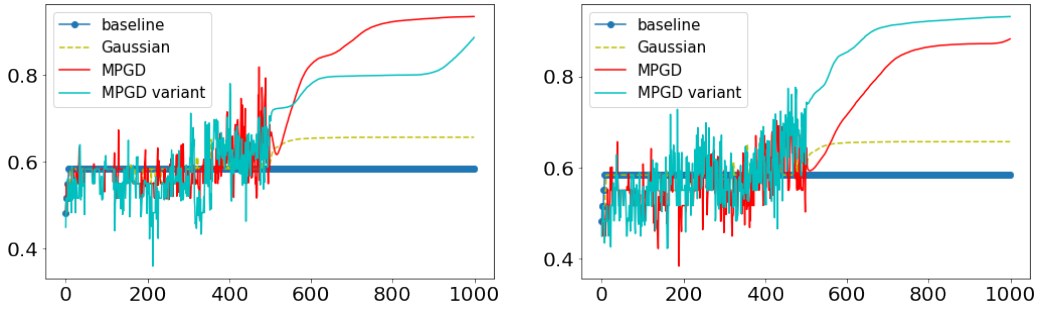

Figure 3: Mean test accuracy obtained under various GD schemes over 1000 epochs for the ECG classification task. Here we choose the learning rate $\eta = 0.05$. MPGD refers to the scheme (8), whereas MPGD variant refers to (11), both using $\mu = \sigma = 0.2$. The MPGDs considered in the left plot use $\gamma = 0.55, \beta = 0$, whereas those in the right plot use $\gamma = 0.55, \beta = 0.5$. We see that applying the MPGD schemes in the first 500 epochs helps to boost the test accuracy significantly when compared to the other two schemes.

Table 3: Shallow neural networks trained on the ECG5000 Data Set for 1000 epochs. The results in parenthesis are achieved with the variant of MPGD (11). All the results are averaged over 5 models trained with different seed values. Here std denotes sample standard deviation. We use $\eta = 0.05$, and $\mu = \sigma = 0.2$ for the perturbation schemes.

Scheme	mean test accuracy in %	std(test accuracy) in %
Baseline full batch GD	58.38	0.0
Gaussian	65.68	16.34
MPGD, $\gamma = 0.55$, $\beta = 0$	93.48 (88.65)	1.17 (9.16)
MPGD, $\gamma = 0.6$, $\beta = 0$	**94.87** (90.52)	0.78 (6.69)
MPGD, $\gamma = 0.65$, $\beta = 0$	92.61 (91.72)	0.99 (4.28)
MPGD, $\gamma = 0.55$, $\beta = 0.5$	88.22 (93.13)	13.99 (2.20)
MPGD, $\gamma = 0.6$, $\beta = 0.5$	94.27 (91.77)	1.06 (2.85)
MPGD, $\gamma = 0.65$, $\beta = 0.5$	94.63 (83.87)	1.13 (23.69)

scheme, we inject the perturbations in the first 500 epochs instead, and stop injecting after that to allow the algorithm to converge. We use the perturbation level $\mu = \sigma = 0.2$. The experimental setup is implemented in PyTorch, and all experiments are run on Google Colab.

Table 3 shows the average test accuracy (evaluated for 5 models that are trained with different seed values) for this task. We see that overall MPGD improves the test accuracy when compared to the vanilla GD and Gaussian perturbations. In fact, the vanilla GD gets stuck in the loss landscape. While Gaussian perturbations can help to mobilize the GD iterates in the landscape, MPGD is more effective in steering the iterates to achieve higher test accuracy (see Figure 3), which is the main goal of the learning task. This illustrates that adding the perturbations of MPGD can help improve the outcome of the optimization process in situations where the vanilla GD and Gaussian perturbations fail to make meaningful progress.

Note that the higher values of the perturbation levels used in the MPGD schemes lead to instabilities and fluctuations of large magnitudes in the earlier epochs. These are necessary to boost the test accuracy; see the jump in the improvement of the test accuracy after epoch 500 in Figure 3. The instabilities and fluctuations could be reduced if lower perturbation levels were used, but this would lower the test accuracy (which would still be higher than the baseline and the Gaussian results) obtained. Again, we emphasize that we are not going after competitive test performance here, but rather demonstrating the effectiveness of MPGD in improving the test performance in situations when other training schemes fail to make substantial progress in optimization.

F.2 Details and Additional Results for the Airfoil Self-Noise Prediction Task

The experimental setup is implemented in PyTorch, and all experiments are run on 4x NVIDIA Tesla T4 GPUs with 16 GB VRAM belonging to an internal SLURM cluster. Table 4 reports the

Table 4: Statistics for the results obtained in Table 1. We report the sample standard deviation, denoted std, of the test RMSE and RMSE gap. The values in parenthesis refer to the sample standard deviation results for the MPGD (11).

Scheme	std(test RMSE)	std(RMSE gap)
Baseline	0.0595	0.0110
Gaussian	0.0980	0.0136
MPGD, $\gamma = 0.55$	0.1035 (0.0629)	0.0201 (0.0162)
MPGD, $\gamma = 0.6$	0.0375 (0.1266)	0.0083 (0.0092)
MPGD, $\gamma = 0.65$	0.0720 (0.0340)	0.0295 (0.0083)
MPGD, $\gamma = 0.7$	0.1091 (0.0094)	0.0162 (0.0076)

Table 5: Mean and standard deviation of RMSE gap under the setting used for obtaining the results in Table 1 but with $\beta = 0$ (left) and $\beta = 0.25$ (right) for the MPGDs instead. The values in parenthesis refer to the corresponding results for the MPGD (11). Here std denotes sample standard deviation, and η, μ and σ are the same as the ones used for obtaining the results in Table 1.

Scheme	mean	std
MPGD, $\gamma = 0.55$	0.2308 (0.2325)	0.0091 (0.0107)
MPGD, $\gamma = 0.6$	0.2274 (0.2339)	0.0151 (0.0167)
MPGD, $\gamma = 0.65$	**0.2273** (0.2373)	0.0113 (0.0173)
MPGD, $\gamma = 0.7$	0.2362 (0.2333)	0.0092 (0.0093)

Scheme	mean	std
MPGD, $\gamma = 0.55$	0.2264 (**0.2238**)	0.0189 (0.0157)
MPGD, $\gamma = 0.6$	0.2267 (0.2247)	0.0127 (0.0152)
MPGD, $\gamma = 0.65$	0.2330 (0.2267)	0.0207 (0.0045)
MPGD, $\gamma = 0.7$	0.2393 (0.2270)	0.0139 (0.0058)

Table 6: Statistics for the results obtained in Table 2. We report the sample standard deviation (in %), denoted std, for the validation accuracy and the accuracy gap. The values in parenthesis refer to the sample standard deviation results for the MPGD variant (11).

Scheme	std(val. accuracy)	std(accuracy gap)
Baseline	1.75	1.97
Gaussian	1.13	0.53
MPGD, $\gamma = 0.55$	2.52 (2.55)	2.45 (2.21)
MPGD, $\gamma = 0.6$	1.52 (2.44)	1.65 (2.00)
MPGD, $\gamma = 0.65$	1.09 (2.86)	1.22 (3.00)
MPGD, $\gamma = 0.7$	4.40 (1.57)	3.55 (1.10)

sample standard deviation for the results obtained in Table 1. Table 5 shows the corresponding results when different values of β are used instead, illustrating the trade-offs induced by the selection of the stability parameter γ and the skewness parameter β for this particular setting. Here, we see that using the MPGD variant (11) with $\gamma = 0.6$ and $\beta = 0.5$ (see Table 1) leads to the lowest RMSE gap.

F.3 Details for the CIFAR-10 Classification Task

We use the implementation of ResNet18 and modify the implementation of the full-batch GD training provided at `https://github.com/JonasGeiping/fullbatchtraining` to set up MPGD and GD with Gaussian perturbations. Please refer to Section 5 and [GGP+21] for the relevant details. The experimental setup is in PyTorch, and all experiments are run on an NVIDIA A100-SXM4 GPU with 40 GB VRAM belonging to an internal SLURM cluster. Table 2 reports the mean validation accuracies over runs of 5 models trained with different seed values, whereas Table 6 reports the sample standard deviation for the results in Table 2.