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, $(V_n, W_n) \to (\mathrm{id}, \mathrm{L})$ 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 $n \to \infty$,

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

Therefore $((W_n, X_n), l_{r+d}) \overset{(\mathrm{d})}{\to} ((\mathrm{L}, X), \phi_{a,b})$ 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 $\mathrm{d}X_t = b \diamond \mathrm{d}\mathrm{L}_t^\alpha$ 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 $\mathrm{d}X_t = b(x^*) \diamond \mathrm{d}\mathrm{L}_t^\alpha$ 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),

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

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:

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

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

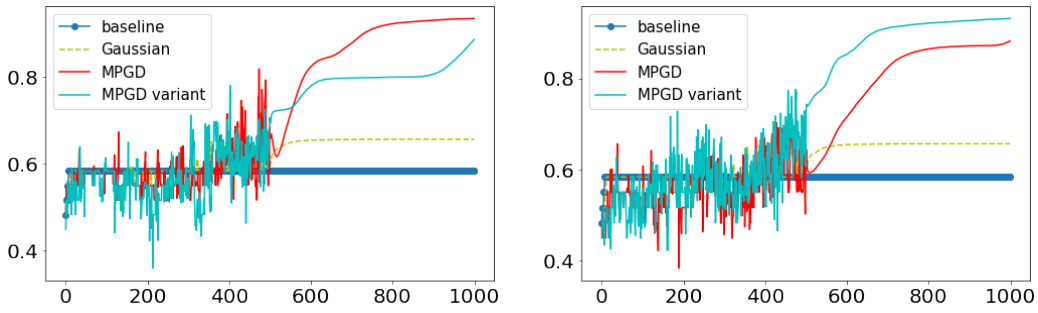

Figure 3: Mean test accuracy obtained under various GD schemes over 1000 epochs for the ECG classification task. Here we choose the learning rate $\eta = 0.05$. MPGD refers to the scheme (8), whereas MPGD variant refers to (11), both using $\mu = \sigma = 0.2$. The MPGDs considered in the left plot use $\gamma = 0.55, \beta = 0$, 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.