Journal of Machine Learning Research 25 (2024) 1-27 Submitted 2/23; Revised 1/24; Published 1/24

Stochastic Modified Flows, Mean-Field Limits and Dynamics of Stochastic Gradient Descent

Benjamin Gess BENJAMIN.GESS@MATH.UNI-BIELEFELD.DE
Fakultät für Mathematik
Universität Bielefeld
33615 Bielefeld, Germany
and
Max Planck Institute for Mathematics in the Sciences
04103 Leipzig, Germany

Sebastian Kassing SEBASTIAN.KASSING@MATH.UNI-BIELEFELD.DE
Fakultät für Mathematik
Universität Bielefeld
33615 Bielefeld, Germany

Vitalii Konarovskyi VITALII.KONAROVSKYI@UNI-HAMBURG.DE
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg
20146 Hamburg, Germany
and
Institute of Mathematics of NAS of Ukraine
01024 Kyiv, Ukraine

Editor: Maxim Raginsky

Abstract

We propose new limiting dynamics for stochastic gradient descent in the small learning rate regime called stochastic modified flows. These SDEs are driven by a cylindrical Brownian motion and improve the so-called stochastic modified equations by having regular diffusion coefficients and by matching the multi-point statistics. As a second contribution, we introduce distribution dependent stochastic modified flows which we prove to describe the fluctuating limiting dynamics of stochastic gradient descent in the small learning rate - infinite width scaling regime.

Keywords: stochastic gradient descent, machine learning, overparametrization, stochastic modified equation, fluctuation mean field limit

1. Introduction

Stochastic gradient descent algorithms (SGD), going back to Robbins and Monro (1951), are the most common way to train neural networks. Due to the non-convexity and non-smoothness of the corresponding loss landscapes, the analysis of the optimization dynamics is highly challenging. The analysis of the implicit, algorithmic bias of SGD in overparameterized networks is one of the key open problems in the understanding of the empirically

©2024 Benjamin Gess, Sebastian Kassing and Vitalii Konarovskyi.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/23-0220.html.

observed good generalization properties of networks trained by SGD. Since the dynamics of SGD depend on many choices, like the choice of the loss function, the architecture of the network and the training data, their systematic understanding relies on the identification of universal structures that are invariant to these many degrees of freedoms, while retaining the essential properties of SGD. In recent years, several of such scaling limits and corresponding limiting dynamics have been identified. Among these, solutions to SDEs have been obtained as universal continuum objects in the small learning rate regime (Li et al. 2019; E et al. 2020), while (stochastic) Wasserstein gradient flows have been found in infinite width overparameterized limits (Nitanda and Suzuki 2017; Chizat and Bach 2018, 2020; Mei et al. 2018; Nguyen 2019; Rotskoff et al. 2019; Javanmard et al. 2020; Sirignano and Spiliopoulos 2020a,b; Gess et al. 2022; Rotskoff and Vanden-Eijnden 2022). In the present work, we introduce a new form of stochastic limiting dynamics which solves simultaneously three challenges met in previous works: (1) the irregularity of diffusion coefficients, (2) matching multi-point statistics, and (3) incorporating overparameterized limits.

Before we comment on each of these aspects in a few more details, let us recall the principle setup of SGD in supervised learning. For a given training data set $\Xi \subseteq \mathbb{R}^{n_0}$ sampled from a probability distribution ϑ, one aims to minimize the empirical risk

$$R(z) := \mathbb{E}_\vartheta \tilde{R}(z, \xi), \quad z \in \mathbb{R}^d,$$

where $\tilde{R} : \mathbb{R}^d \times \Xi \to \mathbb{R}$ is a loss function. Let ξ_n, $n \in \mathbb{N}_0 (:= \mathbb{N} \cup \{0\})$, be i.i.d. samples of training data drawn from ϑ. Then, the SGD dynamics is given by

$$Z^\eta_{n+1}(x) = Z^\eta_n(x) - \eta \nabla \tilde{R}(Z^\eta_n(x), \xi_n), \quad n \in \mathbb{N}_0, \tag{1}$$

where $Z_0(x) = x$, $x \in \mathbb{R}^d$ and $\eta > 0$. In particular, Z^η_n, $n \in \mathbb{N}_0$, allows to analyze the training dynamics of different initializations x subject to the same choice of training data.

We next address the above mentioned challenges in a few more details.

(1) The irregularity of diffusion coefficients: In the regime of small learning rate, the foundational works of Li et al. (2017, 2019) have suggested stochastic modified equations (SME) as universal continuum limits that capture both the average gradient descent performed by SGD and its fluctuations. More precisely, it is shown that the SGD dynamics Z^η_n, $n \in \mathbb{N}_0$, with learning rate η can be approximated to higher order in η by solutions to SMEs

$$dY^\eta_t(x) = -\nabla \left(R(Y^\eta_t(x)) + \frac{\eta}{4} |\nabla R(Y^\eta_t(x))|^2 \right) dt + \sqrt{\eta} \Sigma^{1/2}(Y^\eta_t(x)) dW_t, \tag{2}$$

where $Y^\eta_0(x) = x$ for $x \in \mathbb{R}^d$, W_t, $t \geq 0$, is a Brownian motion in $\mathbb{R}^d$ and $\Sigma : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ is the matrix defined by

$$\Sigma(y) = \mathbb{E}_\vartheta \left[(\nabla_y \tilde{R}(y, \xi) - \nabla R(y)) \otimes (\nabla_y \tilde{R}(y, \xi) - \nabla R(y)) \right], \quad y \in \mathbb{R}^d. \tag{3}$$

This convergence incorporates a certain degree of universality of (2), since the fluctuations in (2) are given in terms of Brownian motion, irrespective of the specific distribution ϑ. In machine learning, and, in particular, in overparameterized settings, the covariance matrix Σ is typically degenerate. As a result, the square root $\Sigma^{1/2}$ appearing in (2) has limited regularity properties,[1] which makes the analysis of (2) challenging, and leads to assumptions on $\Sigma^{1/2}$ that are in general not known to hold, see e.g. Ankirchner and Perko (2023); Perko (2023) and Example 2. The first contribution of this work is to resolve this issue by introducing a new model for the stochastic limiting dynamics, which we name stochastic modified flow (SMF),

$$\begin{aligned} dX_t^\eta(x) &= -\nabla\left(R(X_t^\eta(x)) + \frac{\eta}{4}|\nabla R(X_t^\eta(x))|^2\right)dt + \sqrt{\eta}\int_\Xi G(X_t^\eta(x),\xi)W(d\xi,dt), \\ X_0^\eta(x) &= x, \quad x\in\mathbb{R}^d, \end{aligned} \tag{4}$$

where $G(x,\xi) = \nabla\tilde{R}(x,\xi) - \nabla R(x)$ and W is a cylindrical Wiener process on the space $L_2((\Xi,\vartheta);\mathbb{R})$. It is important to notice that (4) satisfies the same martingale problem as (2), while avoiding the appearance of $\Sigma^{\frac{1}{2}}$, thereby bypassing the resulting irregularity of the diffusion coefficients. In contrast, only regularity assumptions on the individual losses $\tilde{R}$ are needed. More precisely, we get the following result.

Theorem 1 (see Theorem 12 and Corollary 14) *Let $\tilde{R}(\cdot,\xi)$ be regular enough for ϑ-a.e. $\xi\in\Xi$ and let $T>0$. Then for every $f\in \mathrm{C}_b^4(\mathbb{R}^d)$, one has*

$$\sup_{x\in\mathbb{R}^d}\sup_{n:n\eta\le T}\left|\mathbb{E}f(X_{n\eta}^\eta(x)) - \mathbb{E}f(Z_n^\eta(x))\right| \lesssim \eta^2.$$

We note that, under the assumptions of Corollary 14, there exists a unique solution to (4), see Theorem 5, such that the flow of solutions to (4) is differentiable with respect to the initial condition up to a certain order, see e.g. Section 4.6 in Kunita (1990).

(2) Matching multi-point statistics: In a variety of works, a dynamical systems approach to the dynamics of SGD has been introduced (Wu et al. 2018; Sato et al. 2022). This aims at using the concepts of attractors, Lyapunov exponents, stochastic synchronization etc. in the analysis of SGD dynamics, for example, in order to analyze asymptotic global stability, that is, if

$$|Z_n^\eta(x) - Z_n^\eta(y)| \to 0 \quad \text{for } n\to\infty \tag{5}$$

in probability. As before, the systematic analysis of such dynamical behavior of SGD relies on the identification of appropriate universal limiting models. It is thus tempting to analyze the dynamical features of SGD by means of those of (2). However, this is not correct, since (2) only captures the single-point motion of SGD, while dynamical features like stability (5) are properties of the multi-point motions. More precisely, (2) captures the limiting behavior of the law of single motions $\mathrm{Law}(Z_n^\eta(x))$, but not the joint multi-point laws $\mathrm{Law}(Z_n^\eta(x_1),\dots,Z_n^\eta(x_m))$ (see also Example 3).

As a second main contribution, in this work we prove that (SMF), in contrast to (2), captures the correct multi-point distributions of SGD, and therefore opens the way for an analysis of the dynamical properties of its (stochastic) flow. This can also be understood on the level of the corresponding Fokker-Planck equations. While the SME matches only the Fokker-Planck equation of the one-point motion of SGD, the infinite dimensional Fokker-Planck equation of SMF matches also the infinite dimensional Fokker-Planck equation of

1. The simple example $\Sigma(y) = y^2$, $\Sigma^{1/2}(y) = |y|$ shows that not more than Lipschitz continuity can be expected from $\Sigma^{1/2}$ in general.

the flow of SGD, and, therefore, in particular, the Fokker-Planck equations of all multi-point motions.

Theorem 2 (see Theorem 12 and Corollary 15) *Under the assumption of Theorem 1, for every* $\Phi \in \mathrm{C}_b^4(\mathcal{P}_2(\mathbb{R}^d))$ *one has*

$$\sup_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \sup_{n:n\eta\le T} \left|\mathbb{E}\Phi\left(\mu\circ (X_{n\eta}^{\eta})^{-1}\right) - \mathbb{E}\Phi\left(\mu\circ (Z_n^{\eta})^{-1}\right)\right| \lesssim \eta^2,$$

where $\mu\circ f^{-1}$ *denotes the push forward of the measure* μ *under a map* f*. Furthermore, for every* $m\in\mathbb{N}$ *and* $f\in \mathrm{C}_b^4(\mathbb{R}^{dm})$*,*

$$\sup_{x_1,\dots,x_m\in\mathbb{R}^d} \sup_{n:n\eta\le T} \left|\mathbb{E}f(X_{n\eta}^{\eta}(x_1),\dots,X_{n\eta}^{\eta}(x_m)) - \mathbb{E}f(Z_n^{\eta}(x_1),\dots,Z_n^{\eta}(x_m))\right| \lesssim \eta^2.$$

(3) Overparameterized limits: As a third main contribution, we extend the small learning rate limit to also incorporate the infinite width limit. We here consider networks with quadratic loss function. Let $\mathcal{D}\subseteq\mathbb{R}^{n_0}\times\mathbb{R}^{k_0}$ be a given training data set with inputs $\Xi=\{\xi:\ (\xi,f(\xi))\in\mathcal{D}\}$ and labels $\{f(\xi):\ (\xi,f(\xi))\in\mathcal{D}\}$.[2] For the approximation of f we choose a parameterized hypotheses space $\mathcal{M}:=\{f^M(z,\cdot): z\in\mathbb{R}^{Md}\}$, $M,d\in\mathbb{N}$, where

$$f^M(z,\xi)=\frac{1}{M}\sum_{i=1}^M \Psi(z^i,\xi),\quad \xi\in\Xi, \tag{6}$$

with $\Psi:\mathbb{R}^d\times\Xi\to\mathbb{R}^{k_0}$, $z=(z^i)_{i\in[M]}$ and $[M]:=\{1,\dots,M\}$. For example, one can choose $\mathcal{M}$ to be the space of response functions of fully connected feed-forward neural networks with one hidden layer containing M hidden neurons. In that case, we choose a function $\phi:\mathbb{R}\to\mathbb{R}$, the activation function, and we write $z=(z^i)_{i\in[M]}$ with $z^i=(c^i,U^i,b^i)\in\mathbb{R}^{k_0}\times\mathbb{R}^{n_0}\times\mathbb{R}$ and $\Psi(z^i,\xi)=c^i\phi(U^i\cdot\xi+b^i)$. Then,

$$f^M(z,\xi)=\frac{1}{M}\sum_{i=1}^M c^i\phi(U^i\cdot\xi+b^i),\quad \xi\in\Xi.$$

The aim of risk minimization (with respect to the square loss) is to select a suitable model $f^M(z,\cdot)$ minimizing the risk $R(z)=\mathbb{E}_\vartheta\tilde{R}(z,\xi)$, $z\in\mathbb{R}^{Md}$, for

$$\tilde{R}(z,\xi)=\frac{1}{2}|f(\xi)-f^M(z,\xi)|^2,\quad z\in\mathbb{R}^{Md},\xi\in\Xi.$$

As before, this optimization task is executed by the stochastic gradient descent algorithm (1) with the starting value $Z_0^\eta=(Z_0^{i,\eta})_{i\in[M]}$ being a tuple of i.i.d. random variables with distribution $\mu\in\mathcal{P}_2(\mathbb{R}^d)$ that are independent of ξ_n, $n\in\mathbb{N}_0$.

A simple computation gives that

$$R(z)=C_f-\frac{1}{M}\sum_{i=1}^M F(z^i)+\frac{1}{2M^2}\sum_{i,j=1}^M K(z^i,z^j),$$

2. For simplicity we assume that the ground-truth is given by a function $f:\mathbb{R}^{n_0}\to\mathbb{R}^{k_0}$.

where $C_f = \frac{1}{2}\mathbb{E}_\vartheta |f(\xi)|^2$ and

$$F(z^i) = \mathbb{E}_\vartheta \left[f(\xi) \cdot \Psi(z^i, \xi) \right], \quad K(z^i, z^j) = \mathbb{E}_\vartheta \left[\Psi(z^i, \xi) \cdot \Psi(z^j, \xi) \right]. \tag{7}$$

Taking

$$\begin{aligned} V(\nu, z^i) &= \nabla F(z^i) - \int_{\mathbb{R}^d} \nabla_{z^i} K(z^i, y)\nu(dy), \\ G(\nu, z^i, \xi) &= \left(f(\xi) - \int_{\mathbb{R}^d} \Psi(y, \xi)\nu(dy) \right) \nabla_{z^i} \Psi(z^i, \xi) \\ &\quad - \mathbb{E}_\vartheta \left[\left(f(\xi) - \int_{\mathbb{R}^d} \Psi(y, \xi)\nu(dy) \right) \nabla_{z^i} \Psi(z^i, \xi) \right] \end{aligned} \tag{8}$$

and replacing η in (1) by $M\eta$, we can rewrite the expression for the dynamics of $Z_n^\eta = (Z_n^{i,\eta})_{i\in[M]}$, $n \in \mathbb{N}_0$, as follows

$$\begin{aligned} Z_{n+1}^{i,\eta} &= Z_n^{i,\eta} + \eta V(\Gamma_n^{M,\eta}, Z_n^{i,\eta}) + \eta G(\Gamma_n^{M,\eta}, Z_n^{i,\eta}, \xi_n), \\ \Gamma_n^{M,\eta} &= \frac{1}{M} \sum_{j=1}^M \delta_{Z_n^{j,\eta}}, \quad i \in [M], \quad n \in \mathbb{N}_0, \end{aligned} \tag{9}$$

where δ_z denotes the δ-measure in z.

We obtain quantified estimates on the approximation of the dynamics of the empirical measure $\Gamma_n^{M,\eta}$, $n \in \mathbb{N}_0$, of SGD by the solution to a distribution dependent stochastic modified flow (DDSMF)

$$\begin{aligned} dX_t^\eta(x) &= \left[V(\Lambda_t^\eta, X_t^\eta(x)) - \frac{\eta}{4}\nabla |V(\Lambda_t^\eta, X_t^\eta(x))|^2 - \frac{\eta}{4}\left\langle \mathrm{D}|V(\Lambda_t^\eta, X_t^\eta(x))|^2, \Lambda_t^\eta \right\rangle \right] dt \\ &\quad + \sqrt{\eta} \int_\Xi G(\Lambda_t^\eta, X_t^\eta(x), \xi) W(d\xi, dt), \\ X_0^\eta(x) &= x, \quad \Lambda_t^\eta = \mu \circ (X_t^\eta)^{-1}, \quad x \in \mathbb{R}^d, \quad t \geq 0, \end{aligned} \tag{10}$$

where D denotes the differentiation with respect to the measure dependent argument in the sense of Lions,[3] $\langle \varphi, \nu \rangle$ denotes the integration of a function $\varphi : \mathbb{R}^d \to \mathbb{R}$ with respect to a measure ν and W is a cylindrical Wiener process on $L_2((\Xi, \vartheta); \mathbb{R})$. We remark that

$$\frac{1}{2}\left\langle \mathrm{D}|V(\nu, z)|^2, \nu \right\rangle = V(\nu, z)\left\langle \nabla_x \nabla_z K(z, x), \nu(dx) \right\rangle,$$

according to the form of V in (8) and properties of Lions derivative.

Theorem 3 (see Theorem 12, Corollary 16 and Remark 17) *Let Ψ be regular enough and $T > 0$. Then for every $\Phi \in \mathrm{C}_b^4(\mathcal{P}_2(\mathbb{R}^d))$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ with a finite p moment with $p > 2$, one has*

$$\sup_{n: n\eta \leq T} \left| \mathbb{E}\Phi(\Lambda_{n\eta}^\eta) - \mathbb{E}\Phi(\Gamma_n^{M,\eta}) \right| \lesssim \eta^2$$

for every $\eta > 0$ and M large enough.

3. For more details see Section 2.2.

Again, under the assumption of Corollary 16, there exists a unique solution to the DDSMF (10), see Theorem 5.

This extends the framework of SMEs and SMFs to (10) which can thus serve as the starting point to analyze the stochastic dynamics of SGD in large, shallow networks.

Overview of the literature. Stochastic modified equations as limiting objects of SGD in the regime of small learning rates have been introduced in Li et al. (2017, 2019). Following these original papers several results were derived for diffusion approximations with SMEs, e.g., generator based proofs (Feng et al. 2018; Hu et al. 2019), approximations for SGD without reshuffling (Ankirchner and Perko 2022) and uniform-in-time estimates for strongly convex objective functions (Feng et al. 2020; Li and Wang 2022). The approximating SME can be used to derive optimal hyperparameter schedules, e.g. for the learning rate (Li et al. 2017) or the batch-size (Zhao et al. 2022; Perko 2023). For a discussion on the validity of the diffusion approximation for finite (non-infinitesimal) learning rate see Li et al. (2021a).

The derivation of stochastic continuum limits of SGD has proven instrumental in the analysis of optimization dynamics in several regards. For example, in Wojtowytsch (2024) an analysis of the corresponding Fokker-Planck equation has been performed, proving that the limiting distribution carries more mass on flatter minima. This extends earlier work in Zhu et al. (2019); Xie et al. (2020), where the specific structure of the noise in supervised learning is shown to help escaping from sharp and poor minima. In Li et al. (2021b), an SDE approximation suggests that, along a manifold of minimizers, SGD has an implicit bias towards minimizing the trace of the Hessian. In Gess and Kassing (2023), a continuous time model was used in order to derive a Lyapunov function for the convergence rate of momentum SGD, see also Moucer et al. (2023) for a general approach for finding a Lyapunov function for continuous time optimization methods.

In Chizat and Bach (2018); Mei et al. (2018); Rotskoff and Vanden-Eijnden (2018a); Javanmard et al. (2020); Sirignano and Spiliopoulos (2020b), the convergence of gradient descent dynamics for overparameterized neural networks to a Wasserstein gradient flow has been analyzed. The conservative SPDE describing the mean-field limit that incorporates the fluctuations of the stochastic gradient descent was suggested in Rotskoff and Vanden-Eijnden (2018b, 2022). The rigorous study of the well-posedness of this conservative SPDE and proof of quantified central limit theorem has been done in Gess et al. (2022), using the observation that its solutions can be described by the SDE with interaction (11) below, which was investigated e.g., in Pilipenko (2006); Dorogovtsev (2024); Dorogovtsev and Ostapenko (2010); Belozerova (2020); Wang (2021) (see also Kurtz and Xiong 1999; Dorogovtsev 2004; Carmona et al. 2016; Wang 2021 for its connection with McKean–Vlasov SDEs with common noise). It should be noted that the stochastic modified flows proposed in this work are of a particular form of the SDE with interaction (11). In Rotskoff and Vanden-Eijnden (2018a); Sirignano and Spiliopoulos (2020a), a linear SPDE has been rigorously identified in the context of central limit fluctuations of stochastic gradient descent in the overparameterized regime.

The paper is organized as follows: In Section 2, we introduce a stochastic differential equation with interaction (see (11)) that covers both the SMF (4) and the DDSMF (10) and recall existence and uniqueness results assuming Lipschitz-continuity of its coefficients. Moreover, we state a result for the continuous dependence of solutions to the SDE with interaction with respect to its initial distribution, as well as an analog of Kolmogorov's equation in the setting of SDEs with interaction. Section 3 is devoted to the main result of this article, Theorem 12, which compares the dynamics of a discrete time Markov chain with those of a solution to a corresponding SDE with interaction. Theorem 1, Theorem 2 and Theorem 3 then follow as consequences of Theorem 12, see Corollary 14, Corollary 15 and Corollary 16, respectively.

2. Measure-valued Diffusion and Stochastic Modified Flows

The goal of this section is to prove the well-posedness for stochastic modified flows and investigate some properties of the associated semigroup. We recall that $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of probability measures μ on $\mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} |x|^2 \mu(dx) < \infty$$

with the Wasserstein distance defined by

$$\mathcal{W}_2(\mu,\nu) = \inf_{\chi \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |x-y|^2 \chi(dx,dy) \right)^{\frac{1}{2}},$$

where $\Pi(\mu,\nu)$ is the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν. It is well-know that $\mathcal{P}_2(\mathbb{R}^d)$ equipped with the Wasserstein distance $\mathcal{W}_2$ is a Polish space.

Let $L_2((E,\nu);\mathbb{R}^k)$ be the space of all 2-integrable functions from a measure space $(E,\mathcal{E},\nu)$ to $\mathbb{R}^k$ with the usual inner product $\langle\cdot,\cdot\rangle_\nu$ and the associated norm $\|\cdot\|_\nu$. We will further fix a measure space $(\Xi,\mathcal{G},\vartheta)$ such that ϑ is a finite measure and the space $L_2((\Xi,\vartheta);\mathbb{R})$ is separable. We will also consider a cylindrical Wiener process W_t, $t \geq 0$, on $L_2((\Xi,\vartheta);\mathbb{R})$ defined on a filtered complete probability space $(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\geq 0},\mathbb{P})$,[4] that is,

(i) for every $t \geq 0$, the map $W_t : L_2((\Xi,\vartheta);\mathbb{R}) \to L_2((\Omega,\mathbb{P});\mathbb{R})$ is linear;

(ii) for every $h \in L_2((\Xi,\vartheta);\mathbb{R})$, $W_t(h)$, $t \geq 0$, is an $(\mathcal{F}_t)_{t\geq 0}$-Brownian motion with $\operatorname{Var} W_t(h) = \|h\|_\vartheta^2 t$.

We will assume that $(\mathcal{F}_t)_{t\geq 0}$ is the complete right-continuous filtration generated by W_t, $t \geq 0$. For an $(\mathcal{F}_t)_{t\geq 0}$-progressively measurable $L_2((\Xi,\vartheta);\mathbb{R}^k)$-valued process $g(t,\cdot) = \{g(t,\xi),\ \xi \in \Xi\}$, $t \geq 0$, with

$$\int_0^t \|g(s,\cdot)\|_\vartheta^2 ds < \infty$$

a.s. for every $t \geq 0$, we will write

$$\int_0^t \int_\Xi g(s,\xi) W(d\xi,ds) := \int_0^t \Upsilon(s) dW_s$$

for $\Upsilon(s)h = \langle g(s,\cdot),h\rangle_\vartheta = (\langle g_i(s,\cdot),h\rangle_\vartheta)_{i\in[k]}$, $h \in L_2((\Xi,\vartheta);\mathbb{R})$.[5] We note that the stochastic integral with respect to a cylindrical Wiener process can be rewritten as an infinite sum of

4. See Section 2.1.2 in Gawarecki and Mandrekar (2011) for the definition and further properties.

5. For the definition of the integral with respect to a cylindrical Wiener process see, e.g., (Gawarecki and Mandrekar, 2011, Section 2.2.4).

stochastic integrals with respect to independent standard Brownian motions. Indeed, since $L_2((\Xi,\vartheta);\mathbb{R}^k)$ is a separable Hilbert space, there exists an orthonormal basis $\{e_n,\ n\in\mathbb{N}\}$ in $L_2((\Xi,\vartheta);\mathbb{R}^k)$. Then, according to Lemma 2.8 in Gawarecki and Mandrekar (2011),

$$\int_0^t \int_\Xi g(s,\xi)W(d\xi,ds) = \sum_{n=1}^\infty \int_0^t \langle g(s,\cdot),e_n\rangle_\vartheta dW_s(e_n),$$

where $W_t(e_n)$, $t\geq 0$, $n\in\mathbb{N}$, are independent standard $(\mathcal{F}_t)_{t\geq 0}$-Brownian motions and the series on the right-hand side of the equation converges in $L_2((\Omega,\mathbb{P});\mathbb{R})$ for each $t\geq 0$.

2.1 Stochastic Modified Flows

For measurable functions $B:[0,\infty)\times\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d\to\mathbb{R}^d$, $G:[0,\infty)\times\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d\to L_2((\Xi,\vartheta);\mathbb{R}^d)$ and a probability measure $\mu\in\mathcal{P}_2(\mathbb{R}^d)$, we consider the following stochastic differential equation

$$\begin{aligned} dX_t(x) &= B(t,\Lambda_t,X_t(x))dt + \int_\Xi G(t,\Lambda_t,X_t(x),\xi)W(d\xi,dt), \\ X_0(x) &= x, \quad \Lambda_t=\mu\circ X_t^{-1}, \quad x\in\mathbb{R}^d, \quad t\geq 0. \end{aligned} \tag{11}$$

It is clear that the equations (4) and (10) can be written in the form of (11). Therefore, in this section we will only focus on (11) which is called the stochastic differential equation with interaction and was studied, e.g., in Dorogovtsev and Kotelenez (1997); Pilipenko (2006); Wang (2021). Let $\mathcal{B}(E)$ denote the Borel σ-algebra on a topological space E. Following the definition from (Dorogovtsev, 2024, Definition 2.3.1) or (Gess et al., 2022, Definition 2.5), we introduce the notion of a solution to (11).

Definition 4 *A family of continuous processes $\{X_t(x),\ t\geq 0\}$, $x\in\mathbb{R}^d$, is called a (strong) solution to the SDE with interaction* (11) *with initial mass distribution $\mu\in\mathcal{P}_2(\mathbb{R}^d)$ if, for each $t\geq 0$ the restriction of X to the time interval $[0,t]$ is $\mathcal{B}([0,t])\otimes\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{F}_t$-measurable, $\Lambda_t=\mu\circ X_t^{-1}$, $t\geq 0$, is a continuous process in $\mathcal{P}_2(\mathbb{R}^d)$ and for every $x\in\mathbb{R}^d$, a.s.,*

$$X_t(x) = x + \int_0^t B(s,\Lambda_s,X_s(x))ds + \int_0^t\int_\Xi G(s,\Lambda_s,X_s(x),\xi)W(d\xi,ds),$$

for all $t\geq 0$. For convenience, we will also call the measure-valued process Λ_t, $t\geq 0$, a solution to (11)*.*

Let $\phi_p(x)=|x|^p$, $x\in\mathbb{R}^d$. The following theorem was proved in Gess et al. (2022). See Theorem 2.9 and Corollary 2.10 for the well-posedness and the estimates; the existence of a continuous modification of X was observed in the proof of Theorem 2.9 ibid.

Theorem 5 *Assume that the coefficients B, G of* (11) *are Lipschitz continuous with respect to $(\mu,x)\in\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d$, that is, for every $T>0$ there exists $L>0$ such that for each $t\in[0,T]$, $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$ and $x,y\in\mathbb{R}^d$*

$$\begin{aligned} &|B(t,\mu,x)-B(t,\nu,y)| + \|G(t,\mu,x,\cdot)-G(t,\nu,y,\cdot)\|_\vartheta \\ &\qquad \leq L\left(\mathcal{W}_2(\mu,\nu)+|x-y|\right), \end{aligned} \tag{12}$$

and

$$|B(t,\delta_0,0)| + \|G(t,\delta_0,0,\cdot)\|_\vartheta \leq L, \tag{13}$$

where δ_0 denotes the δ-measure at 0 on $\mathbb{R}^d$. Then, for every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists a unique strong solution $X_t(x)$, $t \geq 0$, $x \in \mathbb{R}^d$, to the SDE with interaction (11). Moreover, there exists a version of $X.(x)$, $x \in \mathbb{R}^d$, that is a continuous in (t,x), and for every $T > 0$ and $p \geq 2$ there exists a constant $C > 0$ such that

$$\mathbb{E} \sup_{t\in[0,T]} |X_t(x)|^p \leq C(1 + \langle \phi_p, \mu\rangle + |x|^p),$$

for all $x \in \mathbb{R}^d$. In particular,

$$\mathbb{E} \sup_{t\in[0,T]} \langle \phi_p, \Lambda_t\rangle \leq C(1 + \langle \phi_p, \mu\rangle),$$

where $\Lambda_t = \mu \circ X_t^{-1}$.

From now on, we will only consider the version $X_t(x)$, $t \geq 0$, $x \in \mathbb{R}^d$, of a solution to the SDE with interaction (11) which is continuous in (x,t). In order to reflect the dependency on the initial mass distribution we will write $X_t(\mu, x)$ and $\Lambda_t(\mu)$ instead of $X_t(x)$ and Λ_t. We next recall the result on the continuous dependence of $\Lambda_t(\mu)$, $t \geq 0$, with respect to the initial condition μ, that was obtained in (Gess et al., 2022, Theorem 2.14).

Proposition 6 *Under the assumption of Theorem 5, for every $T > 0$ there exists a constant $C > 0$ depending only on T and the Lipschitz constant L such that*

$$\mathbb{E} \sup_{t\in[0,T]} |X_t(\mu,x) - X_t(\nu,y)|^2 \leq C\left(\mathcal{W}_2^2(\mu,\nu) + |x-y|^2\right)$$

and

$$\mathbb{E} \sup_{t\in[0,T]} \mathcal{W}_2^2\left(\Lambda_t(\mu), \Lambda_t(\nu)\right) \leq C\mathcal{W}_2^2(\mu,\nu)$$

for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $x, y \in \mathbb{R}^d$.

2.2 Measure-valued Diffusion

The goal of this section is to obtain an analog of Kolmogorov's equation for the process $\Lambda_t(\mu)$, $t \geq 0$, given in (11). For this purpose, we need to recall the notion of Lions derivative according to Cardaliaguet (2013). We say that a function $f : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^k$ is L-differentiable at μ, if there exists an element $\mathrm{D}f(\mu)$ in $L_2((\mathbb{R}^d,\mu);\mathbb{R}^k \times \mathbb{R}^d)$ such that

$$\lim_{\|h\|_\mu \to 0} \frac{f(\mu \circ (\mathrm{id}+h)^{-1}) - f(\mu) - \langle \mathrm{D}f(\mu), h\rangle_\mu}{\|h\|_\mu} = 0,$$

where id denotes the identity map on $\mathbb{R}^d$ and the limit is taken over $h \in L_2((\mathbb{R}^d,\mu);\mathbb{R}^d)$. In this case, $\mathrm{D}f(\mu)$ is called the L-derivative of f at μ.

Remark 7 *For the interested reader, we would like to note that the L-derivative is the Lions derivative introduced via the lifting of function on $\mathcal{P}_2(\mathbb{R}^d)$ to functions defined on a Hilbert space $L_2((\tilde{\Omega}, \tilde{\mathbb{P}}); \mathbb{R}^d)$ for a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$,[6] according to the discussion in Section 2 in Ren and Wang (2020). Moreover, under some regularity assumptions on the L-derivative $\mathrm{D}f$, it coincides with the Wasserstein gradient and the gradient of the linear functional derivative,[7] by Theorem 5.64 and Proposition 5.48 in Carmona and Delarue (2018), respectively.*

Let $f : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^k$ be continuous. If for every $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ the function f is L-differentiable at μ and its derivative has a μ-version $\mathrm{D}f(\mu, x)$ such that $\mathrm{D}f(\mu, x)$ is jointly continuous in $(\mu, x) \in \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d$, we will say that f is continuously differentiable on $\mathcal{P}_2(\mathbb{R}^d)$. The set of all continuously differentiable functions on $\mathcal{P}_2(\mathbb{R}^d)$ will be denoted by $\mathrm{C}^1(\mathcal{P}_2(\mathbb{R}^d))$. We will also consider functions defined on the product space $[0, T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m$. Therefore, we define $\mathrm{C}_b^{0,1,1}(\mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m)$ as a class of all continuous bounded functions $f : [0, T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m \to \mathbb{R}^k$ that are continuously differentiable in the second and third variables and their derivatives are jointly continuous and bounded in all variables. Similarly, for $l \in \mathbb{N}$ we can introduce the space $\mathrm{C}_b^{0,l,l}(\mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m)$ by assuming that all mixed derivatives in the second and third variables to the l-th order exist and are jointly continuous and bounded in all variables.

Similarly to $\mathrm{C}_b^{0,l,l}([0, T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m)$, we define the class $\tilde{\mathrm{C}}_b^{0,l,l}([0, T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m)$ as the set of continuous and bounded functions $f : [0, T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m \to L_2((\Xi, \vartheta); \mathbb{R}^k)$ such that for ϑ-a.e. $\xi \in \Xi$ we have $f(\cdot, \cdot, \cdot, \xi) \in \mathrm{C}_b^{0,l,l}([0, T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m)$ and all its mixed derivatives up to the l-th order are continuous and bounded as $L_2((\Xi, \vartheta); \mathbb{R}^k)$-valued functions.

If $f \in \mathrm{C}_b^{0,l,l}([0, T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m)$ is independent of the first variable (resp. first and third variables), we will simply write $f \in \mathrm{C}_b^{l,l}(\mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m)$ (resp. $f \in \mathrm{C}_b^l(\mathcal{P}_2(\mathbb{R}^d))$). The set $\tilde{\mathrm{C}}_b^{l,l}([0, T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m)$ is defined analogously.

Example 1 *If $\varphi_i \in \mathrm{C}_b^1(\mathbb{R}^d)$, $i \in [n]$, and $h \in \mathrm{C}^1(\mathbb{R}^n)$ then $f(\mu) = h(\langle \varphi_1, \mu \rangle, \dots, \langle \varphi_n, \mu \rangle)$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, belongs to $\mathrm{C}_b^1(\mathcal{P}_2(\mathbb{R}^d))$ and*

$$\mathrm{D}f(\mu, x) = \sum_{i=1}^{n} \partial_i h(\langle \varphi_1, \mu \rangle, \dots, \langle \varphi_n, \mu \rangle) \nabla \varphi_i(x), \quad x \in \mathbb{R}^d, \quad \mu \in \mathcal{P}_2(\mathbb{R}^d).$$

For the coefficients B and G of the equation (11), we define the following second-order differential operator

$$\begin{aligned}\mathcal{L}_t f(\mu) := &\frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{A}(t, \mu, x, y) : \mathrm{D}^2 f(\mu, x, y) \mu(dx) \mu(dy) \\ &+ \frac{1}{2} \int_{\mathbb{R}^d} A(t, \mu, x) : \nabla \mathrm{D}f(\mu, x) \mu(dx) \\ &+ \int_{\mathbb{R}^d} B(t, \mu, x) \cdot \mathrm{D}f(\mu, x) \mu(dx),\end{aligned} \tag{14}$$

6. See Definition 6.1 in Cardaliaguet (2013) or Definition 5.22 in Carmona and Delarue (2018).
7. For the definitions of the Wasserstein gradient and the linear functional derivative see e.g. Definitions 5.62 and 5.43 in Carmona and Delarue (2018), respectively.

for $f \in \mathrm{C}_b^2(\mathcal{P}_2(\mathbb{R}^d))$, where

$$\begin{aligned}\tilde{A}(t,\mu,x,y) &= \mathbb{E}_\vartheta \left[G(t,\mu,x,\xi) \otimes G(t,\mu,y,\xi)\right] \\ &= \left(\langle G_i(t,\mu,x,\cdot), G_j(t,\mu,y,\cdot)\rangle_\vartheta\right)_{i,j\in[d]}, \\ A(t,\mu,x) &= \tilde{A}(t,\mu,x,x)\end{aligned}$$

and we use the notation $C : D = \sum_{i,j=1}^d c_{i,j} d_{i,j}$ for $C = (c_{i,j})_{i,j\in[d]}, D = (d_{i,j})_{i,j\in[d]}$ and $a \cdot b = \sum_{i=1}^d a_i b_i$ for $a = (a_i)_{i\in[d]}$, $b = (b_i)_{i\in[d]}$.

We next provide the well posedness of the Kolmogorov equation associated to (11). This result can be obtained as in the proof of Theorem 3.1 in Wang (2021) with slight changes, where a similar equation driven by a finite dimensional noise was considered.

Proposition 8 (Kolmogorov equation) *Let $T > 0$ and the coefficients B, G of (11) belong to $\mathrm{C}_b^{0,2,2}([0,T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d)$ and $\tilde{\mathrm{C}}_b^{0,2,2}([0,T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d)$, respectively. For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, let $\Lambda_t(\mu)$, $t \in [0,T]$, be a solution to (11) with initial mass distribution $\Lambda_0(\mu) = \mu$. Then, for every $\Phi \in \mathrm{C}_b^2(\mathcal{P}_2(\mathbb{R}^d))$, the function*

$$U(t,\mu) = \mathbb{E}\Phi(\Lambda_t(\mu)), \quad (t,\mu) \in [0,T] \times \mathcal{P}_2(\mathbb{R}^d),$$

is a unique solution to the equation

$$\begin{aligned}&\partial_t U(t,\mu) = \mathcal{L}_t U(t,\mu), \\ &U(0,\mu) = \Phi(\mu), \quad (t,\mu) \in [0,T] \times \mathcal{P}_2(\mathbb{R}^d),\end{aligned} \tag{15}$$

in the class $\mathrm{C}_b^{0,2}([0,T] \times \mathcal{P}_2(\mathbb{R}^d))$ with $\partial_t U \in \mathrm{C}([0,T] \times \mathcal{P}_2(\mathbb{R}^d))$.

If, additionally, $B \in \mathrm{C}_b^{0,l,l}([0,T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d)$, $G \in \tilde{\mathrm{C}}_b^{0,l,l}([0,T] \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d)$ and $\Phi \in \mathrm{C}_b^l(\mathcal{P}_2(\mathbb{R}^d))$, for some $l > 2$, then $U \in \mathrm{C}_b^{0,l}([0,T] \times \mathcal{P}_2(\mathbb{R}^d))$.

Remark 9 *In general, the constants that estimate the uniform norm of derivatives of a solution U to the equation (15) will depend on T and increase exponentially fast as $T \to \infty$.*[8] *However, we believe that additional convexity assumptions (e.g. choosing $B = \nabla R$ for a strongly convex function R) can prevent a blow-up of the corresponding constants at infinity and open the way for a uniform-in-time analysis of the stochastic modified equation, see Feng et al. (2020); Li and Wang (2022), and the mean-field dynamics, see Chen et al. (2022); Suzuki et al. (2023).*

3. Diffusion Approximation via Stochastic Modified Flows

The goal of this section is to prove the theorems stated in the introduction. For this, we first show a general result comparing the dynamics of a Markov chain defined below with a corresponding SDE with interaction and cylindrical noise. Then, we show that the results given in the introduction immediately follow from the general comparison statement. We fix measurable functions $V : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ and $G : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to L_2((\Xi,\vartheta);\mathbb{R}^d)$ such

8. We refer the reader to estimates in Section 3 in Wang (2021) for more details.

that $\mathbb{E}_\vartheta G(\mu, x, \xi) = 0$ for all $(\mu, x) \in \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d$. For $\eta > 0$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we consider a Markov chain defined by

$$\begin{aligned} Z^\eta_{n+1}(z) &= Z^\eta_n(z) + \eta V(\Gamma^\eta_n, Z^\eta_n(z)) + \eta G(\Gamma^\eta_n, Z^\eta_n(z), \xi_n), \\ Z^\eta_0(z) &= z, \quad \Gamma^\eta_n = \mu \circ (Z^\eta_n)^{-1}, \quad z \in \mathbb{R}^d, \ \ n \in \mathbb{N}_0, \end{aligned} \tag{16}$$

where ξ_n, $n \in \mathbb{N}_0$, are i.i.d. sampled from the distribution ϑ. We remark that, e.g., the SGD dynamics in the overparameterized shallow neural network in (9) can be written in form of (16) by taking $\mu = \frac{1}{M}\sum_{i=1}^M \delta_{Z_0^i}$ and $Z_n^{i,\eta} = Z_n^\eta(Z_0^{i,\eta})$, $i \in [M]$. We will approximate Γ^η_n, $n \in \mathbb{N}_0$, by solutions to the DDSMF

$$\begin{aligned} dX^\eta_t(x) &= \left[V(\Lambda^\eta_t, X^\eta_t(x)) - \frac{\eta}{4}\nabla |V(\Lambda^\eta_t, X^\eta_t(x))|^2 - \frac{\eta}{4}\left\langle \mathrm{D}|V(\Lambda^\eta_t, X^\eta_t(x))|^2, \Lambda^\eta_t \right\rangle \right] dt \\ &\quad + \sqrt{\eta}\int_\Xi G(\Lambda^\eta_t, X^\eta_t(x)) W(d\xi, dt), \\ X^\eta_0(x) &= x, \quad \Lambda^\eta_t = \mu \circ (X^\eta_t)^{-1}, \quad x \in \mathbb{R}^d, \ \ t \geq 0, \end{aligned} \tag{17}$$

where W is a cylindrical Wiener process on $L_2((\Xi, \vartheta); \mathbb{R})$. We first prove some auxiliary statements that will imply the well-posedness of the DDSMF (17).

Lemma 10 *Let $\varepsilon > 0$ and γ_s, $s \in [0, \varepsilon]$, be a family of square integrable random variables on $\mathbb{R}^k$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If*

$$\gamma_0' := \lim_{s \to 0+} \frac{\gamma_s - \gamma_0}{s}$$

exists in $L_2((\Omega, \mathbb{P}); \mathbb{R}^k)$, then for every $f \in \mathrm{C}^1(\mathcal{P}_2(\mathbb{R}^k))$ one has

$$\lim_{s \to 0+} \frac{f(\mathrm{Law}(\gamma_s)) - f(\mathrm{Law}(\gamma_0))}{s} = \mathbb{E}\left[\mathrm{D}f(\mathrm{Law}(\gamma_0), \gamma_0) \cdot \gamma_0' \right].$$

Proof This statement was obtained in (Wang, 2021, Lemma 2.4). ■

Lemma 11 *Let the functions V and G belong belong to $\mathrm{C}_b^{1,1}(\mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d)$ and $\tilde{\mathrm{C}}_b^{1,1}(\mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d)$, respectively. Then, for every $x, y \in \mathbb{R}^d$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we have*

$$|V(\mu, x) - V(\nu, y)| + \|G(\mu, x, \cdot) - G(\nu, y, \cdot)\|_\vartheta \leq L\left(\mathcal{W}_2(\mu, \nu) + |x - y|\right),$$

with

$$\begin{aligned} L = &\sup_{x, y \in \mathbb{R}^d, \mu \in \mathcal{P}_2(\mathbb{R}^d)} \left(|\nabla V(\mu, x)| + |\mathrm{D}V(\mu, x, y)| \right) \\ &+ \sup_{x, y \in \mathbb{R}^d, \mu \in \mathcal{P}_2(\mathbb{R}^d)} \left(\|\nabla G(\mu, x, \cdot)\|_\vartheta + \|\mathrm{D}G(\mu, x, y, \cdot)\|_\vartheta \right). \end{aligned}$$

Proof Let $x, y \in \mathbb{R}^d$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be fixed. We take an arbitrary probability measure χ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ, ν and consider random variables ζ_0, ζ_1 on the probability space $(\mathbb{R}^d\times\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d\times\mathbb{R}^d), \chi)$ defined by $\zeta_0(x,y) = y$ and $\zeta_1(x,y) = x$ for all $(x,y) \in \mathbb{R}^d\times\mathbb{R}^d$. Then $\mathrm{Law}(\zeta_0) = \nu$ and $\mathrm{Law}(\zeta_1) = \mu$. Set $\gamma_s = (1-s)\zeta_0 + s\zeta_1$, $s \in [0,1]$, and note that $\gamma_i = \zeta_i$, for $i \in \{0,1\}$, and $\gamma_s' = (\zeta_1 - \zeta_0)$, for all $s \in [0,1]$. We have

$$|V(\mu,x) - V(\nu,y)| \le |V(\mu,x) - V(\mu,y)| + |V(\mu,y) - V(\nu,y)|$$

and we can bound the terms on the right hand side of the inequality as follows. With the mean-value theorem, the first term can be bounded by $\sup_{z\in\mathbb{R}^d,\rho\in\mathcal{P}_2(\mathbb{R}^d)} |\nabla V(\rho,z)|\,|x-y|$. To bound the second term, we will use Lemma 10 and the mean-value theorem:

$$\begin{aligned}
|V(\mu,y) - V(\nu,y)| &= |V(\mathrm{Law}(\zeta_1),y) - V(\mathrm{Law}(\zeta_0),y)| \le \sup_{s\in[0,1]} \left|\frac{d}{ds}V(\mathrm{Law}(\gamma_s),y)\right| \\
&\le \sup_{s\in[0,1]} \left| \int_{\mathbb{R}^d}\int_{R^d} \mathrm{D}V(\mathrm{Law}(\gamma_s), y, \gamma_s(z_1,z_2)) \cdot \gamma_s'(z_1,z_2)\chi(dz_1,dz_2) \right| \\
&\le \sup_{z_1,z_2\in\mathbb{R}^d,\rho\in\mathcal{P}_2(\mathbb{R}^d)} |\mathrm{D}V(\rho,z_1,z_2)| \int_{\mathbb{R}^d}\int_{R^d} |\zeta_1(z_1,z_2) - \zeta_0(z_1,z_2)|\chi(dz_1,dz_2) \\
&\le \sup_{z_1,z_2\in\mathbb{R}^d,\rho\in\mathcal{P}_2(\mathbb{R}^d)} |\mathrm{D}V(\rho,z_1,z_2)| \left(\int_{\mathbb{R}^d}\int_{R^d} |z_2 - z_1|^2\chi(dz_1,dz_2)\right)^{\frac{1}{2}}.
\end{aligned}$$

Taking the infimum over all probability measures χ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ, ν, we obtain

$$|V(\mu,y) - V(\nu,y)| \le \sup_{z_1,z_2\in\mathbb{R}^d,\rho\in\mathcal{P}_2(\mathbb{R}^d)} |\mathrm{D}V(\rho,z_1,z_2)|\mathcal{W}_2(\mu,\nu).$$

The estimate for $\|G(x,\mu) - G(y,\nu)\|_\vartheta$ can be obtained similarly. ■

Now we are ready to proof the main result of this work.

Theorem 12 *Let $V \in \mathrm{C}_b^{5,5}(\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d)$, $G \in \tilde{\mathrm{C}}_b^{4,4}(\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d)$. For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\eta > 0$, let $\Gamma_n^\eta(\mu)$, $n \in \mathbb{N}_0$, and $\Lambda_t^\eta(\mu)$, $t \ge 0$, be defined by (16), and (17), respectively. Assume that $\mathbb{E}_\vartheta G(\mu,x,\xi) = 0$ for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $x \in \mathbb{R}^d$ as well as*

$$\sup_{x\in\mathbb{R}^d}\sup_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathbb{E}_\vartheta|G(\mu,x,\xi)|^3 < \infty. \tag{18}$$

Then, for every $\Phi \in \mathrm{C}_b^4(\mathcal{P}_2(\mathbb{R}^d))$ and $T > 0$ there exists a constant C independent of η such that

$$\sup_{\mu\in\mathcal{P}_2(\mathbb{R}^d)}\sup_{n:n\eta\le T} \left|\mathbb{E}\Phi(\Lambda_{n\eta}^\eta(\mu)) - \mathbb{E}\Phi(\Gamma_n^\eta(\mu))\right| \le C\eta^2, \tag{19}$$

for all $\eta > 0$.

Proof We first remark that the measure-valued process $\Lambda_t^\eta(\mu)$, $t \ge 0$, is uniquely defined due to Theorem 5 and Lemma 11. Without loss of generality, we consider $\eta \le T$. The proof of this theorem relies on the comparison of the generators associated with the processes $\Gamma_n^\eta(\mu)$, $n \in \mathbb{N}_0$, and $\Lambda_t^\eta(\mu)$, $t \ge 0$, up to a certain order of η. We first demonstrate how such a bound on their difference can be used to conclude the proof.

We start from the definition of the transition semigroup for the process $\Gamma_n^\eta(\mu)$, $n \in \mathbb{N}_0$. For convenience of notation, we will drop the superscript η in Γ_n^η and Λ_t^η and simply write Γ_n and Λ_t, respectively. Note that $\Gamma_{n+1} = \Gamma_n \circ Y_n^{-1}(\Gamma_n, \cdot)$, where

$$Y_n(\mu, y) = y + \eta V(\mu, y) + \eta G(\mu, y, \xi_n), \quad \mu \in \mathcal{P}_2(\mathbb{R}^d), \quad y \in \mathbb{R}^d.$$

Indeed, by (16), $Z_{n+1}(z) = Y_n(\Gamma_n, Z_n(z))$, $z \in \mathbb{R}^d$, and, hence,

$$\begin{aligned} \Gamma_n \circ Y_n^{-1}(\Gamma_n, \cdot) &= \mu \circ Z_n^{-1} \circ Y_n^{-1}(\Gamma_n, \cdot) \\ &= \mu \circ Y_n(\Gamma_n, Z_n(\cdot))^{-1} = \mu \circ Z_{n+1}^{-1} = \Gamma_{n+1}, \end{aligned}$$

for all $n \in \mathbb{N}_0$. Therefore, defining the linear operator $\mathcal{S}$ on the set of all bounded measurable functions $\Psi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ by

$$\mathcal{S}\Psi(\mu) = \mathbb{E}_\vartheta \Psi(\mu \circ Y_1^{-1}(\mu, \cdot)), \quad \mu \in \mathcal{P}_2(\mu),$$

we conclude that

$$\begin{aligned} \mathbb{E}_\vartheta \Psi(\Gamma_n(\mu)) &= \mathbb{E}_\vartheta \Psi(\Gamma_{n-1}(\mu) \circ Y_{n-1}^{-1}(\Gamma_{n-1}(\mu), \cdot)) \\ &= \mathbb{E}_\vartheta \left[\mathbb{E}_\vartheta \left[\Psi(\Gamma_{n-1}(\mu) \circ Y_{n-1}^{-1}(\Gamma_{n-1}(\mu), \cdot)) \middle| \Gamma_{n-1}(\mu) \right] \right] \\ &= \mathbb{E}_\vartheta \mathcal{S}\Psi(\Gamma_{n-1}(\mu)) = \cdots = \mathcal{S}^n \Psi(\mu), \end{aligned} \tag{20}$$

for all $n \in \mathbb{N}$. Hence, defining $U(t, \mu) = \mathbb{E}\Phi(\Lambda_t(\mu))$, $t \geq 0$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, and using (20), we get for each $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $n \in \mathbb{N}$

$$\begin{aligned} \mathbb{E}\Phi(\Gamma_n(\mu))) - \mathbb{E}\Phi(\Lambda_{n\eta}(\mu)) &= \mathcal{S}^n \Phi(\mu) - U(t_n, \mu) \\ &= \sum_{i=0}^{n-1} \mathcal{S}^{n-i-1} \left(\mathcal{S}U(t_i, \mu) - U(t_{i+1}, \mu) \right), \end{aligned} \tag{21}$$

where $t_i := i\eta$.

Thus, by (21), and by the inequality

$$\sup_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} |\mathcal{S}\Psi(\mu)| \leq \sup_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} |\Psi(\mu)|,$$

we deduce that for all $n \in \mathbb{N}$ with $n\eta \leq T$

$$\sup_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} |\mathbb{E}\Phi(\Lambda_{n\eta}(\mu)) - \mathbb{E}\Phi(\Gamma_n(\mu))| \leq \sup_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=0}^{n-1} |\mathcal{S}U(t_i, \mu) - U(t_{i+1}, \mu)|. \tag{22}$$

In conclusion, to prove (19), it remains to compare $\mathcal{S}U(t_i, \mu)$ with $U(t_{i+1}, \mu)$. For this, we will expand the generators associated with the processes $\Gamma_n^\eta(\mu)$, $n \in \mathbb{N}_0$, and $\Lambda_t^\eta(\mu)$, $t \geq 0$, with respect to η up to the second order.

To obtain the expansion of $\mathcal{S}\Psi(\mu)$ for $\Psi \in \mathrm{C}_b^3(\mathcal{P}_2(\mathbb{R}^d))$, we fix $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\xi \in \Xi$ and consider $Y(\mu, y) = y + \eta V(\mu, y) + \eta G(\mu, y, \xi)$, $y \in \mathbb{R}^d$, as a random variable on the probability space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu)$. Define

$$\gamma_s(y) = (1 - s)y + sY(\mu, y), \quad y \in \mathbb{R}^d, \quad s \in [0, 1].$$

Then $\gamma_0(y) = y$, $\gamma_1(y) = Y(\mu, y)$, $\gamma_s'(y) = \eta(V(\mu, y) + G(\mu, y, \xi))$ and $\mathrm{Law}(\gamma_s) := \mu \circ \gamma_s^{-1}$ for all $y \in \mathbb{R}^d$, $s \in [0, 1]$. Using Taylor's formula, we obtain

$$\begin{aligned}\Psi(\mu \circ Y^{-1}(\mu, \cdot)) = \Psi(\mathrm{Law}(\gamma_1)) = \Psi(\mathrm{Law}(\gamma_0)) + \frac{d}{ds}\Psi(\mathrm{Law}(\gamma_s))\big|_{s=0} \\ + \frac{1}{2}\frac{d^2}{ds^2}\Psi(\mathrm{Law}(\gamma_s))\big|_{s=0} + \frac{1}{2}\int_0^1 \frac{d^3}{ds^3}\Psi(\mathrm{Law}(\gamma_s))(1-s)^3 ds.\end{aligned} \tag{23}$$

We next compute the derivatives appearing in the expression above. By Lemma 10, we get

$$\frac{d}{ds}\Psi(\mathrm{Law}(\gamma_s)) = \eta \int_{\mathbb{R}^d} \mathrm{D}\Psi(\mathrm{Law}(\gamma_s), \gamma_s(x)) \cdot (V(\mu, x) + G(\mu, x, \xi))\mu(dx)$$

and

$$\begin{aligned}\frac{d^2}{ds^2}\Psi(\mathrm{Law}(\gamma_s)) &= \eta \frac{d}{ds}\int_{\mathbb{R}^d} \mathrm{D}\Psi(\mathrm{Law}(\gamma_s), \gamma_s(x)) \cdot (V(\mu, x) + G(\mu, x, \xi))\mu(dx) = \\ &= \eta^2 \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathrm{D}^2\Psi(\mathrm{Law}(\gamma_s), \gamma_s(x), \gamma_s(y)) \\ &\qquad : (V(\mu, x) + G(\mu, x, \xi)) \otimes (V(\mu, y) + G(\mu, y, \xi))\mu(dx)\mu(dy) \\ &+ \eta^2 \int_{\mathbb{R}^d} \nabla\mathrm{D}\Psi(\mathrm{Law}(\gamma_s), \gamma_s(x)) \\ &\qquad : (V(\mu, x) + G(\mu, x, \xi)) \otimes (V(\mu, x) + G(\mu, x, \xi))\mu(dx).\end{aligned}$$

The third derivative $\frac{d^3}{ds^3}\Psi(\mathrm{Law}(\gamma_s))$ can be computed analogously. Since its precise form is not needed, we omit its computation and observe only that $\frac{d^3}{ds^3}\Psi(\mathrm{Law}(\gamma_s))$, $s \in [0, 1]$, is uniformly bounded by $C\|\Psi\|_{\mathrm{C}_b^3}$ for some constant $C > 0$ depending in particular on $\int |G(\mu, x, \xi)|^3\mu(dx)$.

Taking the expectation of (23) with respect to ϑ and using Fubini's theorem, the equalities $\gamma_0(x) = x$, $\mathrm{Law}(\gamma_0) = \mu$, $\mathbb{E}_\vartheta G(\mu, x, \xi) = 0$, the assumption in (18) and the fact that $\Psi \in \mathrm{C}_b^3(\mathcal{P}_2(\mathbb{R}^d))$, we obtain

$$\begin{aligned}\mathcal{S}\Psi(\mu) = \mathbb{E}_\vartheta\Psi(\mathrm{Law}(\gamma_1)) &= \Psi(\mu) + \eta\int_{\mathbb{R}^d} \mathrm{D}\Psi(\mu, x) \cdot V(\mu, x)\mu(dx) \\ &+ \frac{\eta^2}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathrm{D}^2\Psi(\mu, x, y) : V(\mu, x) \otimes V(\mu, y)\mu(dx)\mu(dy) \\ &+ \frac{\eta^2}{2}\int_{\mathbb{R}^d} \nabla\mathrm{D}\Psi(\mu, x) : V(\mu, x) \otimes V(\mu, x)\mu(dx) \\ &+ \frac{\eta^2}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathrm{D}^2\Psi(\mu, x, y) : \tilde{A}(\mu, x, y)\mu(dx)\mu(dy) \\ &+ \frac{\eta^2}{2}\int_{\mathbb{R}^d} \nabla\mathrm{D}\Psi(\mu, x) : A(\mu, x)\mu(dx) + \eta^3 R_1(\Psi, \mu),\end{aligned} \tag{24}$$

where $\sup_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} |R_1(\Psi, \mu)| \le C\|\Psi\|_{\mathrm{C}_b^3}$, for a constant $C > 0$ and

$$\tilde{A}(\mu, x, y) = \mathbb{E}_\vartheta\left[G(\mu, x, \xi) \otimes G(\mu, y, \xi)\right], \quad A(\mu, x) = \tilde{A}(\mu, x, x).$$

We next expand the generator of the process $\Lambda_t^\eta(\mu)$, $t \geq 0$. Recall that $U(t,\mu) = \mathbb{E}\Phi(\Lambda_t(\mu))$, $t \geq 0$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. According to Proposition 8, we can conclude that for every $t \geq t_i$

$$U(t,\mu) = U(t_i,\mu) + \int_{t_i}^t \mathcal{L}U(r,\mu)dr, \tag{25}$$

where $\mathcal{L} = \mathcal{L}_1 + \eta\mathcal{L}_2$ and

$$\begin{aligned}
\mathcal{L}_1 U(r,\mu) &:= \int_{\mathbb{R}^d} V(\mu,x)\cdot \mathrm{D}U(r,\mu,x)\mu(dx),\\
\mathcal{L}_2 U(r,\mu) &:= \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \tilde{A}(\mu,x,y) : \mathrm{D}^2U(r,\mu,x,y)\mu(dx)\mu(dy)\\
&\quad + \frac{1}{2}\int_{\mathbb{R}^d} A(\mu,x) : \nabla\mathrm{D}U(r,\mu,x)\mu(dx)\\
&\quad - \frac{1}{4}\int_{\mathbb{R}^d} \nabla|V(\mu,x)|^2 \cdot \mathrm{D}U(r,\mu,x)\mu(dx)\\
&\quad - \frac{1}{4}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathrm{D}|V(\mu,x)|^2(y)\cdot \mathrm{D}U(r,\mu,x)\mu(dx)\mu(dy).
\end{aligned}$$

Iterating the equality (25) as in the proof of Lemma 3 in Li and Wang (2022), we obtain

$$U(t_{i+1},\mu) = U(t_i,\mu) + \eta\mathcal{L}_1U(t_i,\mu) + \eta^2\left(\mathcal{L}_2 + \frac{1}{2}\mathcal{L}_1^2\right)U(t_i,\mu) + \eta^3 R_2(\mu), \tag{26}$$

where $\sup_{\mu\in\mathcal{P}_2(\mathbb{R}^d)}|R_2(\mu)| \leq C\|U\|_{\mathrm{C}_b^{0,4}([0,T]\times\mathcal{P}_2(\mathbb{R}^d))}$ for a constant $C > 0$.

In order to compare $\mathcal{S}U(t_i,\mu)$ and $U(t_{i+1},\mu)$, we next express $\mathcal{L}_2 + \frac{1}{2}\mathcal{L}_1^2$ in terms of the coefficients of the equation (17). Note that, according to Example 1, we have

$$\begin{aligned}
\mathrm{D}\mathcal{L}_1U(r,\mu,x) &= \nabla\left[V(\mu,x)\cdot\mathrm{D}U(r,\mu,x)\right] + \int_{\mathbb{R}^d}\mathrm{D}\left[V(\mu,y)\cdot\mathrm{D}U(r,\mu,y)\right](x)\mu(dy)\\
&= \mathrm{D}U(r,\mu,x)\nabla V(\mu,x) + V(\mu,x)\nabla\mathrm{D}U(r,\mu,x)\\
&\quad + \int_{\mathbb{R}^d}\mathrm{D}U(r,\mu,y)\mathrm{D}V(\mu,y,x)\mu(dy) + \int_{\mathbb{R}^d}V(\mu,y)\mathrm{D}^2U(r,\mu,y,x)\mu(dy).
\end{aligned}$$

Thus, using the equalities

$$\frac{1}{2}\nabla|V(\mu,x)|^2 = V(\mu,x)\nabla V(\mu,x) \quad \text{and} \quad \frac{1}{2}\mathrm{D}|V(\mu,x)|^2(y) = V(\mu,x)\mathrm{D}V(\mu,x,y),$$

we get

$$\begin{aligned}
\mathcal{L}_1^2U(r,\mu,x) &= \frac{1}{2}\int_{\mathbb{R}^d}\nabla|V(\mu,x)|^2\cdot\mathrm{D}U(r,\mu,x)\mu(dx)\\
&\quad + \int_{\mathbb{R}^d}\nabla\mathrm{D}U(r,\mu,x) : V(\mu,x)\otimes V(\mu,x)\mu(dx)\\
&\quad + \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\mathrm{D}|V(\mu,x)|^2(y)\cdot\mathrm{D}U(r,\mu,x)\mu(dx)\mu(dy)\\
&\quad + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\mathrm{D}^2U(r,\mu,x,y) : V(\mu,x)\otimes V(\mu,y)\mu(dx)\mu(dy).
\end{aligned}$$

Consequently,

$$
\begin{aligned}
\left(\mathcal{L}_2 + \frac{1}{2}\mathcal{L}_1^2\right) U(r,\mu) &= \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \tilde{A}(\mu,x,y) : \mathrm{D}^2 U(r,\mu,x,y)\mu(dx)\mu(dy) \\
&+ \frac{1}{2}\int_{\mathbb{R}^d} A(\mu,x) : \nabla \mathrm{D} U(r,\mu,x)\mu(dx) \\
&- \frac{1}{4}\int_{\mathbb{R}^d} \nabla |V(\mu,x)|^2 \cdot \mathrm{D} U(r,\mu,x)\mu(dx) \\
&- \frac{1}{4}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathrm{D}|V(\mu,x)|^2(y) \cdot \mathrm{D} U(r,\mu,x)\mu(dx)\mu(dy) \\
&+ \frac{1}{4}\int_{\mathbb{R}^d} \nabla |V(\mu,x)|^2 \cdot \mathrm{D} U(r,\mu,x)\mu(dx) \\
&+ \frac{1}{2}\int_{\mathbb{R}^d} \nabla \mathrm{D} U(r,\mu,x) : V(\mu,x)\otimes V(\mu,x)\mu(dx) \\
&+ \frac{1}{4}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathrm{D}|V(\mu,x)|^2(y) \cdot \mathrm{D} U(r,\mu,x)\mu(dx)\mu(dy) \\
&+ \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathrm{D}^2 U(r,\mu,x,y) : V(\mu,x)\otimes V(\mu,y)\mu(dx)\mu(dy) \\
&= \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \tilde{A}(\mu,x,y) : \mathrm{D}^2 U(r,\mu,x,y)\mu(dx)\mu(dy) \\
&+ \frac{1}{2}\int_{\mathbb{R}^d} A(\mu,x) : \nabla \mathrm{D} U(r,\mu,x)\mu(dx) \\
&+ \frac{1}{2}\int_{\mathbb{R}^d} \nabla \mathrm{D} U(r,\mu,x) : V(\mu,x)\otimes V(\mu,x)\mu(dx) \\
&+ \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathrm{D}^2 U(r,\mu,x,\mu) : V(\mu,x)\otimes V(\mu,y)\mu(dx)\mu(dy).
\end{aligned}
$$

Comparing (26) with (24) for $\Psi = U(t_i,\cdot)$, we conclude that

$$
\begin{aligned}
\mathcal{S}U(t_i,\mu) &= U(t_i,\mu) + \eta\mathcal{L}_1 U(t_i,\mu) + \eta^2\left(\mathcal{L}_2 + \frac{1}{2}\mathcal{L}_1^2\right) U(t_i,\mu) + \eta^3 R_1(U(t_i,\cdot),\mu) \\
&= U(t_{i+1},\mu) + \eta^3 R_1(U(t_i,\cdot),\mu) - \eta^3 R_2(\mu).
\end{aligned}
$$

Inserting into (21), and using the fact that $U \in \mathrm{C}_b^{0,4}([0,T]\times\mathcal{P}_2(\mathbb{R}^d))$ (see Proposition 8), yields, for all $n\in\mathbb{N}$ with $n\eta \le T$

$$
\begin{aligned}
\sup_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} |\mathbb{E}\Phi(\Lambda_{n\eta}(\mu)) - \mathbb{E}\Phi(\Gamma_n(\mu))| &\le \sup_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \sum_{i=0}^{n-1} \eta^3 |R_1(U(t_i,\cdot),\mu) - R_2(\mu)| \\
&\le Cn\eta^3 \le CT\eta^2,
\end{aligned}
$$

where C is a constant that depends on T, see Remark 9. This completes the proof of the theorem. ■

Remark 13 *From the proof of Theorem 12 one can see that for every* $\Phi \in \mathrm{C}_b^4(\mathcal{P}_2(\mathbb{R}^d))$ *and* $T > 0$ *there exists a constant* $C > 0$ *such that*

$$\sup_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sup_{n: n\eta \leq T} |\mathbb{E}\Phi(\Lambda_{n\eta}(\mu)) - \mathbb{E}\Phi(\Gamma_n(\mu))| \leq C\eta,$$

for all $\eta > 0$, *if* $\Lambda_t = \Lambda_t(\mu)$, $t \geq 0$, *is defined by the SDE with interaction*

$$\begin{aligned} dX_t(x) &= V(\Lambda_t, X_t(x))dt + \sqrt{\eta} \int_\Xi G(\Lambda_t, X_t(x)) W(d\xi, dt), \\ X_0(x) &= x, \quad \Lambda_t = \mu \circ X_t^{-1}, \quad x \in \mathbb{R}^d, \quad t \geq 0. \end{aligned}$$

We now apply Theorem 12 to the comparison of the SGD dynamics and stochastic modified flows considered in the introduction. First, we recover a variant of the statement for stochastic modified equations.

Corollary 14 *Let* $Z_n^\eta(x)$, $n \in \mathbb{N}_0$, *be defined by* (1) *for a loss function* $\tilde{R}$ *and* $X_t^\eta(x)$, $t \geq 0$, *be a solution to* (4). *Let also* $\tilde{R}(\cdot, \xi) \in \mathrm{C}_b^6(\mathbb{R}^d)$ *for* ϑ*-a.e.* $\xi \in \Xi$ *and assume that*

$$\int_\Xi \|\tilde{R}(\cdot, \xi)\|_{\mathrm{C}_b^6}^2 \vartheta(d\xi) < \infty \quad \text{ and } \quad \sup_{x \in \mathbb{R}^d} \mathbb{E}_\vartheta |\nabla \tilde{R}(x, \xi)|^3 < \infty.$$

Then, for every $f \in \mathrm{C}_b^4(\mathbb{R}^d)$ *and* $T > 0$, *there exists a constant* $C > 0$ *independent of* η *such that*

$$\sup_{x \in \mathbb{R}^d} \sup_{n: n\eta \leq T} \left|\mathbb{E}f(X_{n\eta}^\eta(x)) - \mathbb{E}f(Z_n^\eta(x))\right| \leq C\eta^2$$

for all $\eta > 0$.

Proof Using the dominated convergence theorem it is easily seen that the functions $V := -\nabla R$ and $G := \nabla \tilde{R} - \nabla R$ belong to $\mathrm{C}_b^{5,5}(\mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d)$ and $\tilde{\mathrm{C}}_b^{4,4}(\mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d)$, respectively, where $R = \mathbb{E}_\vartheta \tilde{R}$. Moreover, $\mathbb{E}_\vartheta G(x, \xi) = 0$ for all $x \in \mathbb{R}^d$ and $\sup_{x \in \mathbb{R}^d} \mathbb{E}_\vartheta |G(x, \xi)|^3 < \infty$. Hence, applying Theorem 12 to the function $\Phi(\mu) = \langle f, \mu \rangle$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, that trivially belongs to $\mathrm{C}_b^4(\mathcal{P}_2(\mathbb{R}^d))$, we obtain

$$\begin{aligned} &\sup_{x \in \mathbb{R}^d} \sup_{n: n\eta \leq T} \left|\mathbb{E}f(X_{n\eta}^\eta(x)) - \mathbb{E}f(Z_n^\eta(x))\right| \\ &\qquad = \sup_{\mu = \delta_x, x \in \mathbb{R}^d} \sup_{n: n\eta \leq T} \left|\mathbb{E}\langle f(X_{n\eta}^\eta), \mu \rangle - \mathbb{E}\langle f(Z_n^\eta), \mu \rangle\right| \\ &\qquad \leq \sup_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sup_{n: n\eta \leq T} \left|\mathbb{E}\Phi(\Lambda_{n\eta}^\eta(\mu)) - \mathbb{E}\Phi(\Gamma_n^\eta(\mu))\right| \leq C\eta^2, \end{aligned}$$

for all $\eta > 0$ and some constant $C > 0$ independent of η, where $\Lambda_t^\eta(\mu) = \mu \circ (X_t^\eta)^{-1}$ and $\Gamma_n^\eta(\mu) = \mu \circ (Z_n^\eta)^{-1}$. This completes the proof of the statement. ■

The next example shows the limited regularity properties of the solution to the SME (2).

Example 2 *Let $(\Xi, \mathcal{G}, \vartheta)$ be a probability space with $\vartheta(\{1\}) = \vartheta(\{-1\}) = \frac{1}{2}$ and let $\tilde{R}(x,\xi) = \frac{1}{2}\xi x^2$ for all $x \in \mathbb{R}$ and $\xi \in \Xi$. Then $R \equiv 0$ and $\Sigma^{1/2}(x) = |x|$ for all $x \in \mathbb{R}$. Let $x, \eta > 0$. Then $Y_t^\eta(x) := x \exp(\sqrt{\eta} W_t - \frac{\eta}{2} t))$, $t \geq 0$, solves the SDE (2) with $Y_0^\eta(x) = x$. Contrariwise, for $x \leq 0$ the SDE (2) is solved by $Y_t^\eta(x) := x \exp(-\sqrt{\eta} W_t - \frac{\eta}{2} t))$, $t \geq 0$. Thus, for fixed $t > 0$ the family of random variables $Y_t^\eta(x)$, $x \in \mathbb{R}$, is not differentiable w.r.t. x at the origin. However, since $\tilde{R}$ is a smooth function, the solution to the SMF (4) is smooth w.r.t. the initial condition.*

Corollary 15 *Under the assumptions of Corollary 14, for every $m \in \mathbb{N}$, $f \in \mathrm{C}_b^4(\mathbb{R}^{dm})$, $\Phi \in \mathrm{C}_b^4(\mathcal{P}_2(\mathbb{R}^d))$ and $T > 0$ there exists a constant $C > 0$ independent of η such that*

$$\sup_{x_1,\ldots,x_m \in \mathbb{R}^d} \sup_{n: n\eta \leq T} \left| \mathbb{E} f(X_{n\eta}^\eta(x_1), \ldots, X_{n\eta}^\eta(x_m)) - \mathbb{E} f(Z_n^\eta(x_1), \ldots, Z_n^\eta(x_m)) \right| \leq C\eta^2 \tag{27}$$

and

$$\sup_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sup_{n: n\eta \leq T} \left| \mathbb{E}\Phi(\mu \circ (X_{n\eta}^\eta)^{-1}) - \mathbb{E}\Phi(\mu \circ (Z_n^\eta)^{-1}) \right| \leq C\eta^2 \tag{28}$$

for all $\eta > 0$.

Proof The estimate (28) can be obtained by the same argument as in the proof of Corollary 14. To prove (27), we will apply Corollary 14 to the function

$$\tilde{R}^{\mathrm{ext}}(z,\xi) = \tilde{R}(z_1,\xi) + \ldots + \tilde{R}(z_m,\xi), \quad z = (z_i)_{i\in[m]} \in \mathbb{R}^{dm}, \quad \xi \in \Xi.$$

Note that

$$\nabla \tilde{R}^{\mathrm{ext}}(z,\xi) = \left(\nabla_{z_i} \tilde{R}(z_i,\xi) \right)_{i\in[m]}$$

for all $z = (z_i)_{i\in[m]} \in \mathbb{R}^{dm}$ and $\xi \in \Xi$. Defining $Z_n^{\mathrm{ext},\eta}(x)$, $n \in \mathbb{N}_0$, by (1) with $\tilde{R}$ and $\mathbb{R}^d$ replaced by $\tilde{R}_{\mathrm{ext}}$ and $\mathbb{R}^{dm}$, respectively, it is easily seen that

$$Z_n^{\mathrm{ext},\eta}(x) = \left(Z_n^\eta(x_i) \right)_{i\in[m]}, \quad n \in \mathbb{N}_0,$$

for all $x = (x_i)_{i\in[m]} \in \mathbb{R}^{dm}$.

We next set $R^{\mathrm{ext}}(z) = \mathbb{E}_\vartheta \tilde{R}^{\mathrm{ext}}(z,\xi)$, $z = (z_i)_{i\in[m]}$. Then

$$\nabla R^{\mathrm{ext}}(z) = \left(\nabla_{z_i} R(z_i) \right)_{i\in[m]}$$

and

$$\nabla |\nabla R^{\mathrm{ext}}(z)|^2 = \left(\nabla_{z_i} |\nabla_{z_i} R(z_i)|^2 \right)_{i\in[m]}$$

for all $z = (z_i)_{i\in[m]} \in \mathbb{R}^{dm}$. Moreover,

$$G^{\mathrm{ext}}(z,\xi) := \nabla \tilde{R}^{\mathrm{ext}}(z,\xi) - \nabla R^{\mathrm{ext}}(z,\xi) = \left(G(z_i,\xi) \right)_{i\in[m]},$$

where G is the coefficient of (4) that equals $\nabla \tilde{R} - \nabla R$. Under the assumptions of the corollary, equation (4) with R and G replaced by R^{ext} and G^{ext}, respectively, has a unique solution $X_t^{\mathrm{ext},\eta}(x)$, $x \in \mathbb{R}^{dm}$, $t \geq 0$. Moreover,

$$X_t^{\mathrm{ext},\eta}(x) = \left(X_t^\eta(x_i) \right)_{i\in[m]}, \quad t \geq 0,$$

a.s. for all $x = (x_i)_{i\in[m]}$. Since $\tilde{R}^{\mathrm{ext}}$ satisfies the assumptions of Corollary 14, one gets for every $f \in \mathrm{C}_b^4(\mathbb{R}^{dm})$

$$\begin{aligned}&\sup_{x\in\mathbb{R}^{dm}} \left|\mathbb{E}f(X_{n\eta}^{\mathrm{ext},\eta}(x)) - f(Z_n^{\mathrm{ext},\eta}(x))\right|\\ &= \sup_{x_1,\dots,x_m\in\mathbb{R}^d} \sup_{n:n\eta\le T} \left|\mathbb{E}f(X_{n\eta}^{\eta}(x_1),\dots,X_{n\eta}^{\eta}(x_m)) - \mathbb{E}f(Z_n^{\eta}(x_1),\dots,Z_n^{\eta}(x_m))\right| \le C\eta^2\end{aligned}$$

for a constant $C > 0$ independent of η. This completes the proof of the statement. ■

In the next example, we show that Corollary 15 cannot hold for the solution to the classical stochastic modified equation (2), since the distribution of the two-point motion is different from the distribution of the two-point motion of (4).

Example 3 *The covariation of the two-point motion $(X_t^\eta(x), X_t^\eta(\bar{x}))$, $t \ge 0$, from the SMF (4) equals*

$$[X^\eta(x), X^\eta(\bar{x})]_t = \eta \int_0^t \tilde{A}(X_s^\eta(x), X_s^\eta(\bar{x}))ds, \quad t \ge 0, \tag{29}$$

where $\tilde{A}(x,y) = \langle G(x,\cdot) \otimes G(y,\cdot)\rangle_\vartheta$. However, the covariation of the two-point motion $(Y_t^\eta(x), Y_t^\eta(\bar{x}))$, $t \ge 0$, obtained from the SDE (2), is given by

$$[Y^\eta(x), Y^\eta(\bar{x})]_t = \eta \int_0^t \Sigma(Y_s(x))^{1/2}\Sigma(Y_s(\bar{x}))^{1/2}ds, \quad t \ge 0, \tag{30}$$

for $\Sigma(x) = \tilde{A}(x,x)$. This implies that the processes $(X^\eta(x), X^\eta(\bar{x}))$ and $(Y^\eta(x), Y^\eta(\bar{x}))$ have different distributions in general. We further notice that the covariance of the one step SGD dynamics defined by (1) satisfies

$$\mathrm{cov}(Z_1^\eta(x), Z_1^\eta(y)) = \eta^2 \tilde{A}(x,y),$$

which is comparable with (29), but not with (30).

Next, we consider the SGD scheme $Z_n^\eta = (Z_n^{i,\eta})_{i\in[M]}$, $n \in \mathbb{N}_0$, incorporating the infinite width limit that is defined by (9), where $Z_0^{i,\eta}$, $i \in [M]$, are i.i.d. random variables sampled from a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. We prove the convergence of the empirical distribution process $\Gamma_n^{M,\eta} = \frac{1}{M}\sum_{i=1}^M \delta_{Z_n^{i,\eta}}$, $n \in \mathbb{N}_0$, to a mean-field solution $\Lambda_t^\eta = \mu \circ (X_t^\eta)^{-1}$, $t \ge 0$, of the DDSMF defined by (10).

Corollary 16 *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mu^M = \frac{1}{M}\sum_{j=1}^m \delta_{Z_0^{j,\eta}}$, where $Z_0^{j,\eta}$, $j \in [M]$, are i.i.d. random variables with distribution μ. Let $\Gamma_n^{M,\eta}$, $n \in \mathbb{N}_0$, and Λ_t^η, $t \ge 0$, be as in (9) and (10), respectively, with $\Gamma_0^{M,\eta} = \mu^M$ and $\Lambda_0^\eta = \mu$. Assume that the function Ψ in (6) satisfies: $\Psi(\cdot,\xi) \in \mathrm{C}_b^6(\mathbb{R}^d)$ for ϑ-a.e. $\xi \in \Xi$,*

$$\int_\Xi \left(\|\Psi(\cdot,\xi)\|_{\mathrm{C}_b^6}^2 + |f(\xi)|^2\right) \|\Psi(\cdot,\xi)\|_{\mathrm{C}_b^6}^2 \vartheta(d\xi) < \infty$$

and

$$\sup_{x\in\mathbb{R}^d}\int_\Xi\Big(\|\Psi(\cdot,\xi)\|^3_{C_b^0}+|f(\xi)|^3\Big)|\nabla_x\Psi(x,\xi)|^3\vartheta(d\xi)<\infty.$$

Then, for every $\Phi\in \mathrm{C}_b^4(\mathcal{P}_2(\mathbb{R}^d))$ *there exists a constant* $C>0$ *independent of* η *and* M *such that*

$$\sup_{n:n\eta\le T}\left|\mathbb{E}\Phi(\Lambda^\eta_{n\eta})-\mathbb{E}\Phi(\Gamma^{M,\eta}_n)\right|\le C\eta^2+C\sqrt{\mathbb{E}\mathcal{W}_2^2(\mu,\mu^M)} \tag{31}$$

for all $\eta>0$ *and* $M\in\mathbb{N}$. *In particular, if* μ *has finite pth moment for some* $p>2$, *with* $p\ne 4$ *for* $d\le 4$ *and* $p\ne\frac{d}{d-2}$ *for* $d\ge 5$, *then for every* $a>0$ *there exists a constant* $C>0$ *independent of* η *and* M *such that*

$$\sup_{n:n\eta\le T}\left|\mathbb{E}\Phi(\Lambda^\eta_{n\eta})-\mathbb{E}\Phi(\Gamma^{M,\eta}_n)\right|\le C\eta^2 \tag{32}$$

for all $\eta>0$ *and* $M\in\mathbb{N}$ *satisfying* $\frac{K(M)}{\eta^4}\le a$, *where*

$$K(M)=\begin{cases}M^{-\frac12}+M^{-\frac{p-2}{2}} & \text{if } d\le 3,\\ M^{-\frac12}\ln(1+M)+M^{-\frac{p-2}{2}} & \text{if } d=4,\\ M^{-\frac2d}+M^{-\frac{p-2}{2}} & \text{if } d\ge 5.\end{cases}$$

Proof First, we show that $V\in \mathrm{C}_b^{5,5}(\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d)$ and $G\in\tilde{\mathrm{C}}_b^{4,4}(\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d)$, where V and G are given by (8). Analogously to the proof of Corollary 14, we get that $F\in \mathrm{C}_b^6(\mathbb{R}^d)$, where $F(z)=\mathbb{E}_\vartheta\left[f(\xi)\cdot\Psi(z,\xi)\right]$, $z\in\mathbb{R}^d$, and, thus, $\nabla F\in \mathrm{C}_b^5(\mathbb{R}^d)$. For $K(z^1,z^2)=\mathbb{E}_\vartheta\left[\Psi(z^1,\xi)\cdot\Psi(z^2,\xi)\right]$, $z_1,z_2\in\mathbb{R}^d$, we use the dominated convergence theorem to get that $K\in \mathrm{C}_b^6(\mathbb{R}^{2d})$. Using Example 1, we get, for $\tilde K(\mu,z)=\langle\nabla_z K(z,\cdot),\mu\rangle$, $\mu\in\mathcal{P}_2(\mathbb{R}^d), z\in\mathbb{R}^d$, that

$$D\tilde K(\mu,z^1,z^2)=\nabla_{z^2}\nabla_{z^1}K(z^1,z^2),$$

with analogous expressions for higher derivatives. Thus, $\tilde K\in \mathrm{C}_b^{5,5}(\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d)$ and, therefore, $V\in \mathrm{C}_b^{5,5}(\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d)$. To see that $G\in\tilde{\mathrm{C}}_b^{4,4}(\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d)$ note that

$$\tilde G(\mu,z,\xi)=\left(f(\xi)-\langle\Psi(\cdot,\xi),\mu\rangle\right)\nabla_z\Psi(z,\xi),\quad \mu\in\mathcal{P}_2(\mathbb{R}^d),\ z\in\mathbb{R}^d,\ \xi\in\Xi,$$

satisfies $\tilde G\in\tilde{\mathrm{C}}_b^{4,4}(\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d)$ and $\mathbb{E}_\vartheta[\tilde G(\cdot,\cdot,\xi)]\in \mathrm{C}_b^{4,4}(\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d)$. Moreover, we clearly have $\mathbb{E}_\vartheta G(\mu,x,\xi)=0$ for all $\mu\in\mathcal{P}_2(\mathbb{R}^d)$, $x\in\mathbb{R}^d$ and $\sup_{x\in\mathbb{R}^d}\sup_{\mu\in\mathcal{P}_2(\mathbb{R}^d)}\mathbb{E}_\vartheta|G(\mu,x,\xi)|^3<\infty$.

Note that one needs to check the estimate (31) only for $\eta\in(0,T]$. Let $\Lambda^\eta_t(\mu)$, $t\ge 0$, be defined by (10) for every $\mu\in\mathcal{P}_2(\mathbb{R}^d)$. We next fix $\mu\in\mathcal{P}_2(\mathbb{R}^d)$ and consider the empirical distribution $\mu^M=\frac1M\sum_{i=1}^M\delta_{Z_0^{i,\eta}}$ associated with the family of i.i.d. random variables $Z_0^{i,\eta}$, $i\in[M]$, sampled from the distribution μ. By Theorem 12, there exists a constant $C>0$ independent of η such that

$$\sup_{\mu\in\mathcal{P}_2(\mathbb{R}^d)}\sup_{n:n\eta\le T}\left|\mathbb{E}\Phi(\Lambda^\eta_{n\eta}(\mu))-\mathbb{E}\Phi(\Gamma^\eta_n(\mu))\right|\le C\eta^2$$

for all $\eta \in (0,T]$, where $\Gamma_n^\eta(\mu)$, $n \in \mathbb{N}_0$, is determined by (16) with V and G given by (8). Therefore, using the equality $\Gamma_n^{M,\eta} = \Gamma_n^\eta(\mu^M)$ for all $n \in \mathbb{N}_0$, one has

$$\begin{aligned}
\sup_{n:n\eta\le T} \left|\mathbb{E}\Phi(\Lambda_{n\eta}^\eta(\mu^M)) - \mathbb{E}\Phi(\Gamma_n^{M,\eta})\right| &= \sup_{n:n\eta\le T} \left|\mathbb{E}\Phi(\Lambda_{n\eta}^\eta(\mu^M)) - \mathbb{E}\Phi(\Gamma_n^\eta(\mu^M))\right| \\
&= \sup_{n:n\eta\le T} \left|\mathbb{E}\left[\mathbb{E}\left[\Phi(\Lambda_{n\eta}^\eta(\mu^M)\middle|\mathcal{A}\right] - \mathbb{E}\left[\Phi(\Gamma_n^\eta(\mu^M))\middle|\mathcal{A}\right]\right]\right| \\
&\le \mathbb{E}\left[\sup_{n:n\eta\le T}\left|\mathbb{E}\left[\Phi(\Lambda_{n\eta}^\eta(\mu^M)\middle|\mathcal{A}\right] - \mathbb{E}\left[\Phi(\Gamma_n^\eta(\mu^M))\middle|\mathcal{A}\right]\right|\right] \\
&\le \sup_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \sup_{n:n\eta\le T} \left|\mathbb{E}\Phi(\Lambda_{n\eta}^\eta(\mu)) - \mathbb{E}\Phi(\Gamma_n^\eta(\mu))\right| \le C\eta^2
\end{aligned}$$

for all $\eta \in (0,T]$ and $M \in \mathbb{N}$, where $\mathcal{A} = \sigma(Z_0^{i,\eta},\ i \in [M])$.

We next compare $\mathbb{E}\Phi(\Lambda_{n\eta}^\eta(\mu))$ with $\mathbb{E}\Phi(\Lambda_{n\eta}^\eta(\mu^M))$. Applying Lemma 11 to $V = \Phi$ and $G = 0$, we can estimate

$$|\mathbb{E}\Phi(\Lambda_{n\eta}^\eta(\mu)) - \mathbb{E}\Phi(\Lambda_{n\eta}^\eta(\mu^M))|^2 \le \|\Phi\|_{\mathrm{C}_b^1}^2 \mathbb{E}\mathcal{W}_2^2(\Lambda_{n\eta}^\eta(\mu), \Lambda_{n\eta}^\eta(\mu^M)).$$

Since the coefficients of the SDE (10) are Lipschitz continuous, where the Lipschitz constant can be chosen independently of $\eta \in (0,T]$ due to the assumptions of the corollary and Lemma 11, we can apply Proposition 6 to bound $\mathbb{E}\mathcal{W}_2^2(\Lambda_{n\eta}^\eta(\mu), \Lambda_{n\eta}^\eta(\mu^M))$. Thus, there exists a constant $C > 0$ independent of η, M and n such that

$$\mathbb{E}\mathcal{W}_2^2(\Lambda_{n\eta}^\eta(\mu), \Lambda_{n\eta}^\eta(\mu^M)) \le C\mathbb{E}\mathcal{W}_2^2(\mu, \mu^M)$$

for all $\eta \in (0,T]$, $M \in \mathbb{N}$ and $n \in \mathbb{N}_0$ with $n\eta \le T$. This completes the proof of the first part of the corollary.

If μ has finite pth moment for $p > 2$ such that $p \neq 4$ for $d \le 4$ and $p \neq \frac{d}{d-2}$ for $d \ge 5$, then, by Theorem 1 in Fournier and Guillin (2015),

$$\mathbb{E}\mathcal{W}_2^2(\mu, \mu^M) \le C_1 \langle \phi_p, \mu\rangle^{\frac{2}{p}} K(M),$$

where $\phi_p(x) = |x|^p$, $x \in \mathbb{R}^d$, and $C_1 > 0$ depends only on p and d. Assuming that $\frac{K(M)}{\eta^4} \le a$ for some $a > 0$, we get

$$\sup_{n:n\eta\le T} \left|\mathbb{E}\Phi(\Lambda_{n\eta}^\eta) - \mathbb{E}\Phi(\Gamma_n^{M,\eta})\right| \le C\eta^2 + C\sqrt{\mathbb{E}\mathcal{W}_2^2(\mu,\mu^M)} \le C\eta^2 + C\sqrt{aC_1}\langle \phi_p, \mu\rangle^{\frac{1}{p}}\eta^2.$$

This completes the proof of the second part of the statement. ■

Remark 17 *Assume that the measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ has all finite moments in Corollary 16. Then we can choose p so large that the first term in every case of the definition of the constant $K(M)$ dominates. Therefore, the estimate* (32) *holds for all $\eta > 0$ and $M \ge \frac{a}{\eta^q}$, where $q = 8$ for $d \le 3$, $q = 2d$ for $d \ge 5$ and any $q > 8$ for $d = 4$, since $\frac{K(M)}{\eta^4} \le a$ is satisfied for some $a > 0$ and large enough p.*

Acknowledgments

The authors were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1283/2 2021 – 317210226. BG acknowledges support by the Max Planck Society through the Research Group "Stochastic Analysis in the Sciences (SAiS)". The third author thanks the Max Planck Institute for Mathematics in the Sciences and the University of Bielefeld for their warm hospitality, where a part of this research was carried out.

References

Stefan Ankirchner and Stefan Perko. Towards diffusion approximations for stochastic gradient descent without replacement. *hal-03527878*, 2022. URL `https://hal.science/hal-03527878`.

Stefan Ankirchner and Stefan Perko. A comparison of continuous-time approximations to stochastic gradient descent. *hal-03262396*, 2023. URL `https://hal.science/hal-03262396`.

Maria A. Belozerova. Asymptotic behavior of solutions to stochastic differential equations with interaction. *Theory Stoch. Process.*, 25(2):1–8, 2020. ISSN 0321-3900.

Pierre Cardaliaguet. Notes on mean field games. *P.-L. Lions lectures at College de France.*, 2013. Online at https://www.ceremade.dauphine.fr/∼cardaliaguet/MFG20130420.pdf.

René Carmona and François Delarue. *Probabilistic theory of mean field games with applications. I*, volume 83 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 2018. ISBN 978-3-319-56437-1; 978-3-319-58920-6. Mean field FBSDEs, control, and games.

René Carmona, François Delarue, and Daniel Lacker. Mean field games with common noise. *Ann. Probab.*, 44(6):3740–3803, 2016. ISSN 0091-1798. doi: 10.1214/15-AOP1060. URL `https://doi.org/10.1214/15-AOP1060`.

Fan Chen, Zhenjie Ren, and Songbo Wang. Uniform-in-time propagation of chaos for mean field langevin dynamics. *arXiv:2212.03050*, 2022. URL `https://arxiv.org/abs/2212.03050`.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

Andrey A. Dorogovtsev. Measure-valued Markov processes and stochastic flows on abstract spaces. *Stoch. Stoch. Rep.*, 76(5):395–407, 2004. ISSN 1045-1129. doi: 10.1080/10451120422331292216. URL `https://doi.org/10.1080/10451120422331292216`.

Andrey A. Dorogovtsev. *Measure-valued Processes and Stochastic Flows.* De Gruyter, Berlin, Boston, 2024. ISBN 9783110986518. doi: doi:10.1515/9783110986518. URL `https://doi.org/10.1515/9783110986518`.

Andrey A. Dorogovtsev and Peter Kotelenez. *Smooth stationary solutions of quasilinear stochastic differential equations. Finite mass.* Preprint No. 97. Department of Mathematics Case Western Reserv University Cleveland, Ohio, 1997.

Andrey A. Dorogovtsev and Olena V. Ostapenko. Large deviations for flows of interacting Brownian motions. *Stoch. Dyn.*, 10(3):315–339, 2010. ISSN 0219-4937. doi: 10.1142/S0219493710002978. URL `http://dx.doi.org/10.1142/S0219493710002978`.

Weinan E, Chao Ma, and Lei Wu. Machine learning from a continuous viewpoint, I. *Sci. China Math.*, 63(11):2233–2266, 2020. ISSN 1674-7283. doi: 10.1007/s11425-020-1773-8. URL `https://doi.org/10.1007/s11425-020-1773-8`.

Yuanyuan Feng, Lei Li, and Jian-Guo Liu. Semigroups of stochastic gradient descent and online principal component analysis: properties and diffusion approximations. *Commun. Math. Sci.*, 16(3):777–789, 2018. ISSN 1539-6746,1945-0796. doi: 10.4310/CMS.2018.v16.n3.a8. URL `https://doi.org/10.4310/CMS.2018.v16.n3.a8`.

Yuanyuan Feng, Tingran Gao, Lei Li, Jian-Guo Liu, and Yulong Lu. Uniform-in-time weak error analysis for stochastic gradient descent algorithms via diffusion approximation. *Commun. Math. Sci.*, 18(1):163–188, 2020. ISSN 1539-6746,1945-0796. doi: 10.4310/CMS.2020.v18.n1.a7. URL `https://doi.org/10.4310/CMS.2020.v18.n1.a7`.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738, 2015. ISSN 0178-8051. doi: 10.1007/s00440-014-0583-7. URL `https://doi.org/10.1007/s00440-014-0583-7`.

Leszek Gawarecki and Vidyadhar Mandrekar. *Stochastic differential equations in infinite dimensions with applications to stochastic partial differential equations.* Probability and its Applications (New York). Springer, Heidelberg, 2011.

Benjamin Gess and Sebastian Kassing. Convergence rates for momentum stochastic gradient descent with noise of machine learning type. *arXiv:2302.03550*, 2023. URL `https://arxiv.org/abs/2302.03550`.

Benjamin Gess, Rishabh S. Gvalani, and Vitalii Konarovskyi. Conservative SPDEs as fluctuating mean field limits of stochastic gradient descent. *arXiv:2207.05705*, 2022. URL `https://arxiv.org/abs/2207.05705`.

Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications*, 4(1), 2019.

Adel Javanmard, Marco Mondelli, and Andrea Montanari. Analysis of a two-layer neural network via displacement convexity. *Ann. Statist.*, 48(6):3619–3642, 2020. ISSN 0090-5364. doi: 10.1214/20-AOS1945. URL `https://doi.org/10.1214/20-AOS1945`.

Hiroshi Kunita. *Stochastic flows and stochastic differential equations*, volume 24 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1990. ISBN 0-521-35050-6.

Thomas G. Kurtz and Jie Xiong. Particle representations for a class of nonlinear SPDEs. *Stochastic Process. Appl.*, 83(1):103–126, 1999. ISSN 0304-4149. doi: 10.1016/S0304-4149(99)00024-1. URL https://doi.org/10.1016/S0304-4149(99)00024-1.

Lei Li and Yuliang Wang. On uniform-in-time diffusion approximation for stochastic gradient descent. *arXiv:2207.04922*, 2022. URL https://arxiv.org/abs/2207.04922.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2017.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: mathematical foundations. *J. Mach. Learn. Res.*, 20: Paper No. 40, 2019. ISSN 1532-4435.

Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). In *Advances in Neural Information Processing Systems*, volume 34, pages 12712–12725. Curran Associates, Inc., 2021a.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. In *International Conference on Learning Representations*, 2021b.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA*, 115(33):E7665–E7671, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1806579115. URL https://doi.org/10.1073/pnas.1806579115.

Céline Moucer, Adrien Taylor, and Francis Bach. A systematic approach to Lyapunov analyses of continuous-time models in convex optimization. *SIAM J. Optim.*, 33(3): 1558–1586, 2023. ISSN 1052-6234,1095-7189. doi: 10.1137/22M1498486. URL https://doi.org/10.1137/22M1498486.

Phan-Minh Nguyen. Mean field limit of the learning dynamics of multilayer neural networks. *arXiv:1902.02880*, 2019. URL https://arxiv.org/abs/1902.02880.

Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv:1712.05438*, 2017. URL https://arxiv.org/abs/1712.05438.

Stefan Perko. Unlocking optimal batch size schedules using continuous-time control and perturbation theory. *arXiv:2312.01898*, 2023. URL https://arxiv.org/abs/2312.01898.

Andrey Yu. Pilipenko. Support theorem on stochastic flows with interaction. *Theory Stoch. Process.*, 12(1-2):127–141, 2006. ISSN 0321-3900.

Panpan Ren and Feng-Yu Wang. Space-distribution PDEs for path independent additive functionals of McKean-Vlasov SDEs. *Infin. Dimens. Anal. Quantum Probab. Relat. Top.*, 23(3):2050018, 15, 2020. ISSN 0219-0257,1793-6306. doi: 10.1142/S0219025720500186. URL https://doi.org/10.1142/S0219025720500186.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177729586.

Grant Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv:1805.00915*, 2018a. URL http://arxiv.org/abs/1805.00915.

Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: an interacting particle system approach. *Comm. Pure Appl. Math.*, 75(9): 1889–1935, 2022. ISSN 0010-3640.

Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. In *International Conference on Machine Learning*, pages 5508–5517. PMLR, 2019.

Yuzuru Sato, Daiji Tsutsui, and Akio Fujiwara. Noise-induced degeneration in online learning. *Physica D: Nonlinear Phenomena*, 430:133095, 2022. ISSN 01672789. doi: 10.1016/j.physd.2021.133095.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: a central limit theorem. *Stochastic Process. Appl.*, 130(3):1820–1852, 2020a. ISSN 0304-4149. doi: 10.1016/j.spa.2019.06.003. URL https://doi.org/10.1016/j.spa.2019.06.003.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: a law of large numbers. *SIAM J. Appl. Math.*, 80(2):725–752, 2020b. ISSN 0036-1399. doi: 10.1137/18M1192184. URL https://doi.org/10.1137/18M1192184.

Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Convergence of mean-field langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. *arXiv:2306.07221*, 2023. URL https://arxiv.org/abs/2306.07221.

Feng-Yu Wang. Image-dependent conditional McKean-Vlasov SDEs for measure-valued diffusion processes. *J. Evol. Equ.*, 21(2):2009–2045, 2021. ISSN 1424-3199. doi: 10.1007/s00028-020-00665-z. URL https://doi.org/10.1007/s00028-020-00665-z.

Stephan Wojtowytsch. Stochastic Gradient Descent with Noise of Machine Learning Type Part II: Continuous Time Analysis. *J. Nonlinear Sci.*, 34(1):Paper No. 16, 2024. ISSN 0938-8974,1432-1467. doi: 10.1007/s00332-023-09992-0. URL https://doi.org/10.1007/s00332-023-09992-0.

Lei Wu, Chao Ma, and Weinan E. How SGD selects the global minima in over-parameterized learning: a dynamical stability perspective. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2020.

Jim Zhao, Aurelien Lucchi, Frank Norbert Proske, Antonio Orvieto, and Hans Kersting. Batch size selection by stochastic optimal control. In *Has it Trained Yet? NeurIPS 2022 Workshop*, 2022.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning*, pages 7654–7663. PMLR, 2019.