Mean-Field Langevin Dynamics and Energy Landscape of Neural Networks

Kaitong Hu, Zhenjie Ren, David Siska, Lukasz Szpruch

▶ **To cite this version:**

Kaitong Hu, Zhenjie Ren, David Siska, Lukasz Szpruch. Mean-Field Langevin Dynamics and Energy Landscape of Neural Networks. 2019. hal-02292964

HAL Id: hal-02292964
https://hal.science/hal-02292964v1

Preprint submitted on 20 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Noname manuscript No.
(will be inserted by the editor)

Mean-Field Langevin Dynamics and Energy Landscape of Neural Networks

Kaitong Hu · Zhenjie Ren · David Šiška · Łukasz Szpruch

21st May 2019

Abstract We present a probabilistic analysis of the long-time behaviour of the nonlocal, diffusive equations with a gradient flow structure in 2-Wasserstein metric, namely, the Mean-Field Langevin Dynamics (MFLD). Our work is motivated by a desire to provide a theoretical underpinning for the convergence of stochastic gradient type algorithms widely used for non-convex learning tasks such as training of deep neural networks. The key insight is that the certain class of the finite dimensional non-convex problems becomes convex when lifted to infinite dimensional space of measures. We leverage this observation and show that the corresponding energy functional defined on the space of probability measures has a unique minimiser which can be characterised by a first order condition using the notion of linear functional derivative. Next, we show that the flow of marginal laws induced by the MFLD converges to the stationary distribution which is exactly the minimiser of the energy functional. We show that this convergence is exponential under conditions that are satisfied for highly regularised learning tasks. At the heart of our analysis is a pathwise perspective on Otto calculus used in gradient flow literature which is of independent interest. Our proof of convergence to stationary probability measure is novel and it relies on a generalisation of LaSalle's invariance principle. Importantly we do not assume that interaction potential of MFLD is of convolution type nor that has any particular symmetric structure. This is critical for applications. Finally, we show that the error between finite dimensional optimisation problem and its infinite dimensional limit is of order one over the number of parameters.

Keywords Mean-Field Langevin Dynamics · Gradient Flow · Neural Networks

Mathematics Subject Classification (2010) MSC 60H30 · MSC 37M25

Acknowledgement

The third and fourth authors acknowledge the support of The Alan Turing Institute under the Engineering and Physical Sciences Research Council grant EP/N510129/1.

K. Hu
CMAP, École Polytechnique

Z. Ren
CEREMADE, Université Paris-Dauphine

D. Šiška
School of Mathematics, University of Edinburgh

Ł. Szpruch
School of Mathematics, University of Edinburgh

1 Introduction

This work develops rigorous mathematical framework to study non-convex learning tasks such as training of deep neural networks. We provide a theoretical underpinning for the convergence of stochastic gradient type algorithms widely used in practice to train multi-layers neural networks. Deep neural networks trained with stochastic gradient descent algorithm proved to be extremely successful in number of applications such as computer vision, natural language processing, generative models or reinforcement learning [42]. However, complete mathematical theory that would provide theoretical guarantees for the convergence of machine learning algorithms for non-convex learning tasks has been elusive. On the contrary, empirical experiments demonstrate that classical learning theory [57] may fail to predict the behaviour of modern machine learning algorithms [60]. In fact, it has been observed that the performance of neural networks based algorithms is insensitive to the number of parameters in the hidden layers (provided that this is sufficiently large) and in practice one works with models that have number of parameters larger than the size of the training set [29, 5]. These findings motivate the study of neural networks with large number of parameters which is a subject of this work.

Furthermore while universal representation theorems ensures the existence of the optimal parameters of the network, it is in general not known when such optimal parameters can be efficiently approximated by conventional algorithms, such as stochastic gradient descent. This paper aims at revealing the intrinsic connection between the optimality of the network parameters and the dynamic of gradient-descent-type algorithm, using the perspective of the mean-field Langevin equation.

Let us first briefly recall the classical finite dimensional Langevin equation. Given a *potential* function $f : \mathbb{R}^d \to \mathbb{R}$ which is Lipschitz continuous and satisfies appropriate growth condition, the overdamped Langevin equation reads

$$\mathrm{d}X_t = -\nabla f(X_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \tag{1.1}$$

where σ is a scalar constant and W is a d-dimension Brownian motion. One can view this dynamic in two perspectives:

i) The solution to (1.1) is a time-homogeneous Markov diffusion, so under mild condition it admits a unique invariant measure $m^{\sigma,*}$, of which the density function must be in the form

$$m^{\sigma,*}(x) = \frac{1}{Z}\exp\left(-\frac{2}{\sigma^2}f(x)\right), \quad \text{for all} \quad x \in \mathbb{R}^d, \ \text{where } Z := \int_{\mathbb{R}^d} \exp\left(-\frac{2}{\sigma^2}f(x)\right)\mathrm{d}x\,.$$

ii) The dynamic (1.1) can be viewed as the path of a randomised continuous time gradient descent algorithm.

These two perspectives are unified through the variational form of the invariant measure, namely, $m^{\sigma,*}$ is the unique minimiser of the free energy function

$$V^\sigma(m) := \int_{\mathbb{R}^d} f(x)m(\mathrm{d}x) + \frac{\sigma^2}{2}H(m)$$

over all probability measure m, where H is the relative entropy with respect to the Lebesgue measure. The variational perspective has been established in [37] and [38]. Moreover, one may observe that the distribution $m^{\sigma,*}$ concentrates to the Dirac measure $\delta_{\arg\min f}$ as $\sigma \to 0$ and there is no need to assume that the function f is convex. This establishes the link between theory of statistical sampling and optimisation and show that Langevin equation plays an important role in the non-convex optimisation. This fact is well-recognized by the communities of numerical optimisation and machine learning [34, 32, 31]

This paper aims at generalising the connection between the global minimiser and the invariant measure to the case where the *potential* function is a function defined on a space of probability measures. This is motivated by the following observation on the configuration of neural network. Let us take the example of the network with 1-hidden-layer. While the universal representation theorem, [19, 2] tells us that 1-hidden-layer network can arbitrarily well approximate the continuous function on the compact time interval it

does not tell us how to find optimal parameters. One is faced with the following non-convex optimisation problem.

$$\min_{\beta_{n,i}\in\mathbb{R},\alpha_{n,i}\in\mathbb{R}^{d-1}} \left\{ \int_{\mathbb{R}\times\mathbb{R}^{d-1}} \Phi\Big(y - \frac{1}{n}\sum_{i=1}^{n} \beta_{n,i}\varphi(\alpha_{n,i}\cdot z)\Big)\nu(\mathrm{d}y,\mathrm{d}z) \right\}, \tag{1.2}$$

where $\Phi : \mathbb{R} \to \mathbb{R}$ is a convex function, $\varphi : \mathbb{R} \to \mathbb{R}$ is a bounded, continuous, non-constant activation function and ν is a measure of compact support representing the data. Let us define the empirical law of the parameters as $m^n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\{\beta_{n,i},\alpha_{n,i}\}}$. Then

$$\frac{1}{n}\sum_{i=1}^{n} \beta_{n,i}\varphi(\alpha_{n,i}\cdot z) = \int_{\mathbb{R}^d} \beta\varphi(\alpha\cdot z)\, m^n(\mathrm{d}\beta,\mathrm{d}\alpha)\,.$$

To ease notation let us use, for $x = (\beta,\alpha) \in \mathbb{R}^d$, the function $\hat{\varphi}(x,z) := \beta\varphi(\alpha\cdot z)$, and by $\mathbb{E}^m$ we denote the expectation of random variable X under the probability measure m. Now, instead of (1.2), we propose to study the following minimisation problem over the probability measures:

$$\min_m F(m), \quad \text{with} \quad F(m) := \int_{\mathbb{R}^d} \Phi\Big(y - \mathbb{E}^m\big[\hat{\varphi}(X,z)\big]\Big)\,\nu(\mathrm{d}y,\mathrm{d}z), \tag{1.3}$$

This reformulation is crucial, because the *potential* function F defined above is convex in the measure space i.e. for any probability measures m and m' it holds that

$$F((1-\alpha)m + \alpha m') \le (1-\alpha)F(m) + \alpha F(m') \ \text{ for all } \ \alpha \in [0,1]\,.$$

This example demonstrates that a non-convex minimisation problem on a finite-dimensional parameter space becomes a convex minimisation problem when lifted to the infinite dimensional space of probability measures. The key aim of this work is to provide analysis that takes advantage of this observation. We also show that this simple example generalises to certain deep neural networks architectures.

In order to build up the connection between the global minimiser of the convex potential function F and the upcoming mean-field Langevin equation, as in the classic case, we add the relative entropy H as a regulariser, but different from the classic case, we use the relative entropy with respect to a Gibbs measure of which the density is proportional to $e^{-U(x)}$. A typical choice of the Gibbs measure could be the standard Gaussian distribution. One of our main contributions is to characterise the minimiser of the free energy function

$$V^\sigma := F + \frac{\sigma^2}{2}H$$

using the *linear functional derivative* on the space of probability measures, denoted by $\frac{\delta}{\delta m}$ (introduced originally in calculus of variations and now used extensively in the theory of mean field games see, e.g. Cardaliaguet et al. [12]). Indeed, we prove the following first order condition:

$$m^* = \arg\min_m V^\sigma(m) \quad \text{if and only if} \quad \frac{\delta F}{\delta m}(m^*,\cdot) + \frac{\sigma^2}{2}\log(m^*) + \frac{\sigma^2}{2}U = \textit{constant}.$$

This condition together with the fact that m^* is a probability measure gives

$$m^*(x) = \frac{1}{Z}\exp\left(-\frac{2}{\sigma^2}\left(\frac{\delta F}{\delta m}(m^*,x) + U(x)\right)\right),$$

where Z is the normalising constant. We emphasise that throughout V and hence m^* depend on the regularisation parameter $\sigma > 0$. It is noteworthy that the variational form of the invariant measure of the classic Langevin equation is a particular example of this first order condition. Moreover, given a measure m^* satisfying the first order condition, it is formally a stationary solution to the nonlinear Fokker–Planck equation:

$$\partial_t m = \nabla\cdot\left(\Big(D_m F(m,\cdot) + \frac{\sigma^2}{2}\nabla U\Big)m + \frac{\sigma^2}{2}\nabla m\right), \tag{1.4}$$

where $D_m F$ is the *intrinsic derivative* on the probability measure space, defined as $D_m F(m,x) := \nabla \frac{\delta F}{\delta m}(m,x)$. Clearly, the particle dynamic corresponding to this Fokker-Planck equation is governed by the *mean field Langevin equation*:

$$\mathrm{d}X_t = -\Big(D_m F(m_t, X_t) + \frac{\sigma^2}{2}\nabla U(X_t)\Big)\mathrm{d}t + \sigma \mathrm{d}W_t, \quad \text{where} \;\; m_t := \mathrm{Law}(X_t). \tag{1.5}$$

Therefore, formally, we have already obtained the correspondence between the minimiser of the free energy function and the invariant measure of (1.5). In this paper, the connection is rigorously proved mainly with a probabilistic argument.

For the particular application to the neural network (1.3), it is crucial to observe that the dynamics corresponding to the mean field Langevin dynamics describes exactly the path of the randomised regularized gradient-descent algorithm. More precisely, consider the case where we are given data points $(y_m, z_m)_{m\in\mathbb{N}}$ which are i.i.d. samples from ν. If the loss function Φ is simply the square loss then a version of the (randomized, regularized) gradient descent algorithm for the evolution of parameter x_k^i will simply read as

$$x_{k+1}^i = x_k^i + 2\tau\left(\left(y_k - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}(x_k^j, z_k)\right)\nabla\hat{\varphi}(x_k^i, z^k) - \frac{\sigma^2}{2}\nabla U(x_k^i)\right) + \sigma\sqrt{\tau}\xi_k^i\,,$$

with ξ_k^i independent samples from $N(0, I_d)$ (for details we refer the reader to Section 8.2). This evolution can be viewed as a discretization of (1.5).

1.1 Theoretical Contributions and Literature Review

The study of stationary solutions to nonlocal, diffusive equations (1.4) is classical topic with it roots in statistical physics literature and with strong links to Kac's program in Kinetic theory [47]. In particular, variational approach has been developed in [14,49,56] where authors studied dissipation of entropy for granular media equations with the symmetric interaction potential of convolution type (interaction potential corresponds to term $D_m F$ in (1.4)). We also refer a reader to similar results with proofs based on particle approximation of [15,58], coupling arguments [23] and Khasminskii's technique [11,8]. All of the above results impose restrictive condition on interaction potential or/and require it to be sufficiently small. We manage to relax these assumptions allowing for the interaction potential to be arbitrary (but sufficiently regular/bounded) function of measure. Our proof is probabilisitic in nature. Using Lasalle's invariance principle and the HWI inequality from Otto and Villani [50] as the main ingredients, we prove the desired convergence. This approach, to our knowledge, is original, and it clearly justifies the solvability of the randomized/regularized gradient descent algorithm for neural networks. Furthermore, we provide probabilistic proof based on Itô calculus of chain rule for the flow of measures defined by (1.4). This can be viewed as an extension of [39, Theorem 3.1] to the McKean–Vlasov dynamics. Finally we clarify how different notions of calculus on the space of probability measures enter our framework. The calculus is critical to work with arbitrary functions of measure. We refer to [13, Chapter 5] for an overview on that topic. The calculus on the measure space enables to derive and quantify the error between finite dimensional optimisation problem and its infinite dimensional limit.

While working on this paper, other groups developed similar mean-field description of non-convex learning problems, see [46,45,18,51,35,52]. In particular the pioneering work of Mei, Misiakiewicz and Montanari [46], Chizat and Bach [18] as well as Rotskoff and Vanden-Eijnden [51] proved convergence of gradient algorithms to the minimum using the theory of gradient flow in the Wasserstein space of probability distributions [1]. Results in [46] are the closest to ours but the proofs are different. While [46] builds on ideas from [38], we provide probabilisitic perspective. We generalise and provide complete proofs of some key results such as chain rule for the flows of measures [46, Lemma 6.1] (our Theorem 2.8) and global convergence of flow of measures to the invariant measure [46, Lemma 6.12] (our Theorem 2.10). In particular we established convergence to the invariant measure in 2-Wasserstein distance and also demonstrated that for sufficiently regularised problem that convergence is exponential. Furthermore we are able to deal with general loss function. This was conjectured in Appendix B of [46] and is needed if one hopes to treat more general loss functions/network architectures.

1.2 Calculus on the Space of Probability Measures

By $\mathcal{P}(\mathbb{R}^d)$ we denote the space of probability measures on $\mathbb{R}^d$, and by $\mathcal{P}_p(\mathbb{R}^d)$ the subspace of $\mathcal{P}(\mathbb{R}^d)$ in which the measures have finite p-moment for $p \geq 1$. Note that $\pi \in \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$ is called a coupling of μ and ν in $\mathcal{P}_p(\mathbb{R}^d)$, if for any borel subset B of $\mathbb{R}^d$ we have $\pi(B, \mathbb{R}^d) = \mu(B)$ and $\pi(\mathbb{R}^d, B) = \nu(B)$. By $\mathcal{W}_p$ we denote the Wasserstein-p metric on $\mathcal{P}_p(\mathbb{R}^d)$, namely,

$$\mathcal{W}_p(\mu,\nu) := \inf\left\{ \Big(\int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^p \pi(\mathrm{d}x,\mathrm{d}y)\Big)^{\frac{1}{p}};\ \pi \text{ is a coupling of } \mu \text{ and } \nu \right\} \quad \text{for} \quad \mu,\nu \in \mathcal{P}_p(\mathbb{R}^d).$$

It is convenient to recall that

i) $\big(\mathcal{P}_p(\mathbb{R}^d), \mathcal{W}_p\big)$ is a Polish space;
ii) $\mathcal{W}_p(\mu_n, \mu) \to 0$ if and only if μ_n weakly converge to μ and $\int_{\mathbb{R}^d} |x|^p \mu^n(\mathrm{d}x) \to \int_{\mathbb{R}^d} |x|^p \mu(\mathrm{d}x)$;
iii) for $p' > p$, the set $\{\mu \in \mathcal{P}_p(\mathbb{R}^d) : \int_{\mathbb{R}^d} |x|^{p'} \mu(\mathrm{d}x) \leq C\}$ is $\mathcal{W}_p$-compact.

We say a function $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is in $\mathcal{C}^1$ if there exists a bounded continuous function $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^d)\times\mathbb{R}^d \to \mathbb{R}$ such that

$$F(m') - F(m) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}\big((1-\lambda)m + \lambda m', x\big)(m'-m)(\mathrm{d}x)\mathrm{d}\lambda. \tag{1.6}$$

We will refer to $\frac{\delta F}{\delta m}$ as the linear functional derivative. There is at most one $\frac{\delta F}{\delta m}$, up to a constant shift, satisfying (1.6). To avoid the ambiguity, we impose

$$\int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m,x) m(\mathrm{d}x) = 0.$$

If $(m,x) \mapsto \frac{\delta F}{\delta m}(m,x)$ is continuously differentiable in x, we define its intrinsic derivative $D_m F : \mathcal{P}(\mathbb{R}^d)\times\mathbb{R}^d \to \mathbb{R}^d$ by

$$D_m F(m,x) = \nabla\left(\frac{\delta F}{\delta m}(m,x)\right).$$

In this paper ∇ always denotes the gradient in the variable $x \in \mathbb{R}^d$.

Example 1.1 *If $F(m) := \int_{\mathbb{R}^d} \phi(x) m(\mathrm{d}x)$ for some bounded continuous function $\phi : \mathbb{R}^d \to \mathbb{R}$, we have $\frac{\delta F}{\delta m}(m,x) = \phi(x)$ and $D_m F(m,x) = \dot{\phi}(x)$.*

It is useful to see what intrinsic measure derivative look like in the special case when we consider empirical measures

$$m^N := \frac{1}{N}\sum_{i=1}^N \delta_{x^i}, \ \text{ where } x^i \in \mathbb{R}^d.$$

Then one can define $F^N : (\mathbb{R}^d)^N \to \mathbb{R}$ as $F^N(x^1,\dots,x^N) = F(m^N)$. From [16, Proposition 3.1] we know that that if $F \in \mathcal{C}^1$ then $F^N \in C^1$ and for any $i = 1,\dots,N$ and $(x^1,\dots,x^N) \in (\mathbb{R}^d)^N$ it holds that

$$\partial_{x^i} F^N(x^1,\dots,x^N) = \frac{1}{N} D_m F\left(m^N, x^i\right). \tag{1.7}$$

We remark that for notational simplicity in the proofs the constant $C > 0$ can be different from line to line.

2 Main Results

The objective of this paper is to study the minimizer(s) of a convex function $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$.

Assumption 2.1 *Assume that $F \in \mathcal{C}^1$ is convex and bounded from below.*

Instead of directly considering the minimization $\min_m F(m)$, we propose to first study the regularized version, namely, the minimization of the free energy function:

$$\min_{m\in\mathcal{P}(\mathbb{R}^d)} V^\sigma(m), \quad \text{where} \quad V^\sigma(m) := F(m) + \frac{\sigma^2}{2}H(m), \quad \text{for all} \quad m \in \mathcal{P}(\mathbb{R}^d),$$

where $H : \mathcal{P}(\mathbb{R}^d) \to [0,\infty]$ is the relative entropy (Kullback–Leibler divergence) with respect to a given Gibbs measure in $\mathbb{R}^d$, namely,

$$H(m) := \int_{\mathbb{R}^d} m(x) \log\left(\frac{m(x)}{g(x)}\right) \mathrm{d}x,$$

where

$$g(x) = e^{-U(x)} \text{ with } U \text{ s.t. } \int_{\mathbb{R}^d} e^{-U(x)}\,\mathrm{d}x = 1\,,$$

is the density of the Gibbs measure and the function U satisfies the following conditions.

Assumption 2.2 *The function $U : \mathbb{R}^d \to \mathbb{R}$ belongs to C^∞. Further,*

i) there exist constants $C_U > 0$ and $C'_U \in \mathbb{R}$ such that

$$\nabla U(x)\cdot x \geq C_U|x|^2 + C'_U \quad \textit{for all} \quad x \in \mathbb{R}^d\,.$$

ii) ∇U is Lipschitz continuous.

Immediately, we obtain that there exist $0 \leq C' \leq C$ such that for all $x \in \mathbb{R}^d$

$$C'|x|^2 - C \leq U(x) \leq C(1+|x|^2), \quad |\Delta U(x)| \leq C.$$

A typical choice of g would be the density of the d-dimensional standard Gaussian distribution. We recall that such relative entropy H has the properties: it is strictly convex when restricted to measures absolutely continuous with g, it is weakly lower semi-continuous and its sub-level sets are compact. For more details, we refer the readers to the book [20, Section 1.4]. The original minimization and the regularized one is connected through the following Γ-convergence result.

Proposition 2.3 *Assume that F is continuous in the topology of weak convergence. Then the sequence of functions $V^\sigma = F + \frac{\sigma^2}{2}H$ Γ-converges to F when $\sigma \downarrow 0$. In particular, given the minimizer $m^{*,\sigma}$ of V^σ, we have*

$$\varlimsup_{\sigma\to 0} F(m^{*,\sigma}) \;=\; \inf_{m\in\mathcal{P}_2(\mathbb{R}^d)} F(m).$$

It is a classic property of Γ-convergence that every cluster point of $\left(\arg\min_m V^\sigma(m)\right)_\sigma$ is a minimizer of F.

Moreover, when the relative entropy H is strictly convex, then so is the function V, and thus the minimizer $\operatorname{argmin}_{m\in\mathcal{P}(\mathbb{R}^d)} V(m)$, if exists, must be unique. It can be characterized by the following first order condition.

Proposition 2.4 *Under Assumption 2.1 and 2.2, the function V^σ has a unique minimizer $m^* \in \mathcal{P}_2(\mathbb{R}^d)$ which is absolutely continuous with respect to Lebesgue measure and satisfies*

$$\frac{\delta F}{\delta m}(m^*,\cdot) + \frac{\sigma^2}{2}\log(m^*) + \frac{\sigma^2}{2}U \quad \textit{is a constant, } m^*-a.s.$$

On the other hand, we have $m' = \arg\min_{m\in\mathcal{P}(\mathbb{R}^d)} V^\sigma$, if

$$m' \in \mathcal{I}_\sigma := \left\{ m \in \mathcal{P}(\mathbb{R}^d) : \quad \frac{\delta F}{\delta m}(m,\cdot) + \frac{\sigma^2}{2}\log(m) + \frac{\sigma^2}{2}U \quad \textit{is a constant} \right\}. \tag{2.1}$$

Further, we are going to approximate the minimizer of V^σ, using the marginal laws of the solution to the upcoming mean field Langevin equation. Let $\sigma \in \mathbb{R}_+$ and consider the following McKean–Vlasov SDE:

$$\mathrm{d}X_t = -\left(D_m F(m_t, X_t) + \frac{\sigma^2}{2}\nabla U(X_t)\right)\mathrm{d}t + \sigma \mathrm{d}W_t, \tag{2.2}$$

where m_t is the law of X_t and $(W_t)_{t\geq 0}$ is a standard d-dimensional Brownian motion.

Remark 2.5 *i) Let $F(m) = \int_{\mathbb{R}^d} f(x)m(\mathrm{d}x)$ for some function f in $C^1(\mathbb{R}^d, \mathbb{R})$. We know that $D_m F(m, x) = \nabla f(x)$. Hence with this choice of F and entropy regulariser with respect to the Lebesgue measure, the dynamics (2.2) becomes the standard overdamped Langevin equation (1.1).*

ii) If the Gibbs measure is chosen to be a standard Gaussian distribution, the potential of the drift of (2.2) becomes $F(m) + \frac{\sigma^2}{4}\int_{\mathbb{R}^d}|x|^2 m(\mathrm{d}x)$. This shares the same spirit as ridge regression.

Assumption 2.6 *Assume that the intrinsic derivative $D_m F : \mathcal{P}(\mathbb{R}^d)\times\mathbb{R}^d \to \mathbb{R}^d$ of the function $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ exists and satisfies the following conditions:*

i) $D_m F$ is bounded and Lipschitz continuous, i.e. there exists $C_F > 0$ such that for all $x, x \in \mathbb{R}^d$ and $m, m' \in \mathcal{P}_2(\mathbb{R}^d)$

$$|D_m F(m, x) - D_m F(m', x')| \leq C_F\big(|x - x'| + \mathcal{W}_2(m, m')\big) \tag{2.3}$$

ii) $D_m F(m, \cdot) \in \mathcal{C}^\infty(\mathbb{R}^d)$ for all $m \in \mathcal{P}(\mathbb{R}^d)$;

iii) $\nabla D_m F : \mathcal{P}(\mathbb{R}^d)\times\mathbb{R}^d \to \mathbb{R}^d\times\mathbb{R}^d$ is jointly continuous.

The well-posedness of the McKean–Vlasov SDE (2.2) under Assumption 2.2 and 2.6 on the time interval $[0, t]$, for any t, is well known, see e.g. Snitzman [53].

Proposition 2.7 *Under Assumption 2.2 and 2.6 the mean field Langevin SDE (2.2) has a unique strong solution, if $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Moreover, the solution is stable with respect to the initial law, that is, given $m_0, m_0' \in \mathcal{P}_2(\mathbb{R}^d)$, denoting by $(m_t)_{t\in\mathbb{R}^+}, (m_t')_{t\in\mathbb{R}^+}$ the marginal laws of the corresponding solutions to (2.2), we have for all $t > 0$ there is a constant $C > 0$ such that*

$$\mathcal{W}_2(m_t, m_t') \;\leq\; C\mathcal{W}_2(m_0, m_0').$$

We shall prove the process $\big(V^\sigma(m_t)\big)_t$ is decreasing and satisfies the following dynamic.

Theorem 2.8 *Let $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Under Assumption 2.2 and 2.6, we have for any $t > s > 0$*

$$V^\sigma(m_t) - V^\sigma(m_s) = -\int_s^t \int_{\mathbb{R}^d} \left| D_m F(m_r, x) + \frac{\sigma^2}{2}\frac{\nabla m_r}{m_r}(x) + \frac{\sigma^2}{2}\nabla U(x)\right|^2 m_r(x)\mathrm{d}x\mathrm{d}r. \tag{2.4}$$

Remark 2.9 *This theorem can be viewed as an extension of [39, Theorem 3.1] to the McKean–Vlasov dynamics. In order to prove (2.4), we use the Itô calculus as the main tool, similar to [39]. It is noteworthy that in [39] the authors apply the Itô calculus to the time-reversed processes, in view of the key observation in their Theorem 4.2, inherited from [25]. However, this observation no longer holds true for the McKean–Vlasov dynamics, due to the nonlinearity of the corresponding Fokker-Planck equation. Instead, we apply the Itô calculus in the conventional forward way.*

Formally, there is a clear connection between the derivative $\frac{\mathrm{d}V^\sigma(m_t)}{\mathrm{d}t}$ in (2.4) and the first order condition (2.1), and it is revealed by the following main theorem.

Theorem 2.10 *Let Assumption 2.1, 2.2 and 2.6 hold true and $m_0 \in \cup_{p>2}\mathcal{P}_p(\mathbb{R}^d)$. Denote by $(m_t)_{t\in\mathbb{R}^+}$ the flow of marginal laws of the solution to (2.2). Then, there exists an invariant measure of (2.2) equal to $m^* := \mathrm{argmin}_m V^\sigma(m)$, and $(m_t)_{t\in\mathbb{R}^+}$ converges to m^*.*

Once the convergence to the minimizer established, it is natural to look into the rate of convergence. In this paper we manage to give a partial answer under the following assumptions, which can be typically satisfied for σ big enough (in other words, in the highly regularized case). Note that some of the assumptions stated as part of Assumption 2.11 clearly overlap with Assumptions 2.2 and 2.6. They are stated in one place for reader's convenience.

Assumption 2.11 (For exponential convergence) *Let $\sigma > 0$ be fixed and the mean-field Langevin dynamics (2.2) start from $m_0 \in \mathcal{P}_p(\mathbb{R}^d)$ for some $p > 2$. Assume that there are constants $C > 0$, $C_F > 0$ and $C_U > 0$ such that for all $x, x' \in \mathbb{R}^d$ and $m, m' \in \mathcal{P}_1(\mathbb{R}^d)$ we have*

$$\begin{aligned} |D_m F(m,x) - D_m F(m',x')| &\le C_F\Big(|x-x'| + \mathcal{W}_1(m,m')\Big)\,, \\ |D_m F(m,0)| &\le C_F\Big(1 + \int_{\mathbb{R}^d} |y|\, m(\mathrm{d}y)\Big), \end{aligned} \tag{2.5}$$

and

$$\begin{aligned} (\nabla U(x) - \nabla U(x'))\cdot(x-x') &\ge C_U|x-x'|^2\,, \\ |\nabla U(x)| &\le C_U(1+|x|)\,, \end{aligned} \tag{2.6}$$

where the constants satisfy

$$\frac{\sigma^2}{2}(p-1) + 3C_F + \frac{\sigma^2}{2}|\nabla U(0)| - C_U\frac{\sigma^2}{2} < 0\,. \tag{2.7}$$

While it becomes more restrictive in the choice of σ and U, the assumption (2.5) allows more candidate functions F comparing to Assumption 2.6, for example, $(x,m) \mapsto D_m F(m,x)$ is allowed unbounded.

Theorem 2.12 *Let Assumptions 2.1 and 2.11 hold true. Then*

$$\mathcal{W}_2(m_t, m^*) \le e^{(6C_F - C_U)t}\mathcal{W}_2(m_0, m^*)\,,$$

where $(m_t)_{t\ge 0}$ is the flow of marginal laws of solution to (2.2).

Remark 2.13 *i) The contraction rate in Theorem 2.12 rests upon the condition that Lipschitz constant of $D_m F$ is sufficiently small in comparison to dissipativity constant C_U and σ in (2.7). It is a common constraint in the study of exponential convergence concerning the McKean-Vlasov dynamics, see e.g. [23], [43].*

ii) Besides using σ big enough, the condition (2.7) can be also satisfied with large enough C_U. Take $\beta > 0$ and in place of the Gibbs measure g consider

$$g^\beta(x) = e^{\frac{-2\beta}{\sigma^2}U(x)} \quad \text{and} \quad \int_{\mathbb{R}^d} e^{\frac{-2\beta}{\sigma^2}U(x)}\,\mathrm{d}x = 1\,,$$

The corresponding mean-field Langevin dynamics then becomes

$$\mathrm{d}X_t^\beta = -\Big(D_m F(m_t^\beta, X_t^\beta) + \beta\,\nabla U(X_t^\beta)\Big)\,\mathrm{d}t + \sigma \mathrm{d}W_t\,.$$

Now C_U that will appear in (2.7) is replaced by $\frac{2\beta}{\sigma^2}C_U$ and so (2.7) is replaced by the condition

$$\frac{\sigma^2}{2}(p-1) + 3C_F + \frac{\sigma^2}{2}|\nabla U(0)| - 2\beta C_U < 0$$

which can always be fulfilled by taking $\beta > 0$ sufficiently large. Hence we conclude that the flow of marginal laws $(m_t^\beta)_{t>0}$ converges exponentially to $m^{\beta,*}$. One can also note that for fixed β the law $g^\beta(x)\mathrm{d}x$ becomes singular when σ converges to 0.

3 Application to Gradient Descent of Neural Networks

Before proving the main results, we shall first apply them to study the minimization over a neural network. In particular, in Corollary 3.3 we shall show that the marginal laws of the corresponding mean-field Langevin dynamics converge to the optimal weight of the neural network with 1-hidden layer. Further we also have a discussion on the application to the deep neural network in Section 3.2.

Fix a locally Lipschitz function $\varphi : \mathbb{R} \to \mathbb{R}$ and for $l \in \mathbb{N}$ define $\varphi^l : \mathbb{R}^l \to \mathbb{R}^l$ as the function given, for $z = (z_1, \dots, z_l)^\top$ by $\varphi^l(z) = (\varphi(z_1), \dots, \varphi(z_l))^\top$. We fix $L \in \mathbb{N}$ (the number of layers), $l_k \in \mathbb{N}$, $k = 0, 1, \dots L-1$ (the size of input to layer k) and $l_L \in \mathbb{N}$ (the size of the network output). A fully connected artificial neural network is then given by $\Psi = ((\alpha^1, \beta^1), \dots, (\alpha^L, \beta^L)) \in \Pi$, where, for $k = 1, \dots, L$, we have real $l^k \times l^{k-1}$ matrices α^k and real l^k-dimensional vectors β^k. We see that $\Pi = (\mathbb{R}^{l^1 \times l^0} \times \mathbb{R}^{l^1}) \times (\mathbb{R}^{l^2 \times l^1} \times \mathbb{R}^{l^2}) \times \cdots \times (\mathbb{R}^{l^L \times l^{L-1}} \times \mathbb{R}^{l^L})$. The artificial neural network defines a reconstruction function $\mathcal{R}\Psi : \mathbb{R}^{l^0} \to \mathbb{R}^{l^L}$ given recursively, for $z_0 \in \mathbb{R}^{l^0}$, by

$$(\mathcal{R}\Psi)(z^0) = \alpha^L z^{L-1} + \beta^L\,, \quad z^k = \varphi^{l^k}(\alpha^k z^{k-1} + \beta^k)\,, k = 1, \dots, L-1\,.$$

If for each $k = 1, \dots, L-1$ we write α_i^k, β_i^k to denote the i-th row of the matrix α^k and vector β^k respectively then we can write the reconstruction of the network equivalently as

$$(\mathcal{R}\Psi)(z^0)_i = \alpha_i^L \cdot z^{L-1} + \beta_i^L\,, \quad (z^k)_i = \varphi(\alpha_i^k \cdot z^{k-1} + \beta_i^k)\,, k = 1, \dots, L-1\,. \tag{3.1}$$

We note that the number of parameters in the network is $\sum_{i=1}^{L}(l_{k-1}l_k + l_k)$.

In supervised learning the task is to find the parameters Ψ such that the artificial neural network provides a good approximation to a real world problem. In practice this means that given a potential function Φ and training data $(y^j, z^j)_{j=1}^N$, $(y_j, z_j) \in \mathbb{R}^d$ one approximates the optimal parameters by finding

$$\operatorname*{argmin}_{\Psi \in \Pi} \frac{1}{N} \sum_{j=1}^{N} \Phi\Big(y^j - (\mathcal{R}\Psi)(z^j)\Big)\,.$$

Since the typical machine learning task involves huge data sets it makes sense to invoke the law of large numbers, postulate that the training data are distributed according to some measure ν which has a compact support and instead frame the problem as

$$\operatorname*{argmin}_{\Psi \in \Pi} \int_{\mathbb{R}^d} \Phi\Big(y - (\mathcal{R}\Psi)(z)\Big)\,\nu(\mathrm{d}y, \mathrm{d}z)\,. \tag{3.2}$$

This is a non-convex minimization problem, so in general hard to solve. Theoretically, the following universal representation theorem ensures that the minimum value should attain 0, provided that $y = f(z)$ with a continuous function f.

Theorem 3.1 (Universal Representation Theorem) *If an activation function φ is bounded, continuous and non-constant, then for any compact set $K \subset \mathbb{R}^d$ the set*

$$\Big\{(\mathcal{R}\Psi) : \mathbb{R}^d \to \mathbb{R} : (\mathcal{R}\Psi) \text{ given by } (3.1) \text{ with } L = 2 \text{ for some } n \in \mathbb{N}, \alpha_j^2, \beta_j^1 \in \mathbb{R}, \alpha_j^1 \in \mathbb{R}^d, j = 1, \dots, n\Big\}$$

is dense in $C(K)$.

For an elementary proof, we refer the readers to [33, Theorem 2].

3.1 Fully connected 1-hidden layer neural network

Take $L = 2$, fix $d \in \mathbb{N}$ and $n \in \mathbb{N}$ and consider the following 1-hidden layer neural network for approximating functions from $\mathbb{R}^d$ to $\mathbb{R}$: let $l_0 = d$, let $l_1 = n$, let $\beta^2 = 0 \in \mathbb{R}$, $\beta^1 = 0 \in \mathbb{R}^n$, $\alpha^1 \in \mathbb{R}^{n\times d}$. We will denote, for $i \in \{1, \dots, l^0\}$, its i-th row by $\alpha_i^1 \in \mathbb{R}^{1\times d}$. Let $\alpha^2 = (\frac{c_1}{n}, \cdots, \frac{c_n}{n})^\top$, where $c_i \in \mathbb{R}$. The neural network is $\Psi^n = \Big((\alpha^1, \beta^1), (\alpha^2, \beta^2)\Big)$ (where we emphasise the that the size of the hidden layer is n). For $z \in \mathbb{R}^{l^0}$, its reconstruction can be written as

$$(\mathcal{R}\Psi^n)(z) = \alpha^2 \varphi^{l^1}(\alpha^1 z) = \frac{1}{n}\sum_{i=1}^n c_i \varphi(\alpha_i^1 \cdot z)\,.$$

The key observation is to note that, due to law of large numbers (and under appropriate technical assumptions) $\frac{1}{n}\sum_{j=1}^n c_j\varphi(\alpha_j^1 \cdot z) \to \mathbb{E}^m[B\varphi(A \cdot z)]$ as $n \to \infty$, where m is the law of the pair of random variables (B, A) and $\mathbb{E}^m$ is the expectation under the measure m. Therefore, another way (indeed a more intrinsic way regarding to the universal representation theorem) to formulate the minimizaiton problem (3.2) is:

$$\min_{m\in\mathcal{P}(\mathbb{R}^d\times\mathbb{R})} \tilde{F}(m)\,, \quad \text{where} \quad \tilde{F}(m) := \int_{\mathbb{R}^d} \Phi\big(y - \mathbb{E}^m[B\varphi(A \cdot z)]\big)\,\nu(\mathrm{d}z, \mathrm{d}y)\,.$$

For technical reason, we introduce a truncation function $\ell : \mathbb{R} \to K$ (K denotes again some compact set), and consider the truncated version of the minimization:

$$F(m) := \int_{\mathbb{R}^d} \Phi\big(y - \mathbb{E}^m[\ell(B)\varphi(A \cdot z)]\big)\,\nu(\mathrm{d}z, \mathrm{d}y).$$

It is crucial to note that in the reformulation the objective function F becomes a convex function on $\mathcal{P}(\mathbb{R}^d)$, provided that Φ is convex.

Assumption 3.2 *We apply the following assumptions on the coefficients Φ, μ, φ, ℓ:*

i) the function Φ is convex, smooth and $0 = \Phi(0) = \min_{a\in\mathbb{R}} \Phi(a)$;
ii) the data measure μ is of compact support;
iii) the truncation function $\ell \in C_b^\infty(\mathbb{R}^d)$ such that $\dot{\ell}$ and $\ddot{\ell}$ are bounded;
iv) the activation function $\varphi \in C_b^\infty(\mathbb{R}^d)$ such that $\dot{\varphi}$ and $\ddot{\varphi}$ are bounded.

Corollary 3.3 *Under Assumption 3.2, the function F satisfies Assumption 2.1, 2.6. In particular, with a Gibbs measure of which the function U satisfies Assumption 2.2, the corresponding mean field Langevin equation* (2.2) *admits a unique strong solution, given $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Moreover, the flow of marginal laws of the solution, $(m_t)_{t\in\mathbb{R}^+}$, satisfies*

$$\lim_{t\to+\infty} \mathcal{W}_2\left(m_t, \operatorname*{argmin}_{m\in\mathcal{P}(\mathbb{R}^d)} V^\sigma(m)\right) = 0.$$

Proof Let us define, for $x = (\beta, \alpha) \in \mathbb{R}^d$, $\beta \in \mathbb{R}$, $\alpha \in \mathbb{R}^{d-1}$ and $z \in \mathbb{R}^{d-1}$ the function $\hat{\varphi}(x, z) := \ell(\beta)\varphi(\alpha \cdot z)$. Then

$$\begin{aligned} \tfrac{\delta F}{\delta m}(m, x) &= -\textstyle\int_{\mathbb{R}^d\times\mathbb{R}} \dot{\Phi}\big(y - \mathbb{E}^m[\hat{\varphi}(X, z)]\big)\hat{\varphi}(x, z)\,\nu(\mathrm{d}z, \mathrm{d}y) \\ \text{and} \quad D_m F(m, x) &= -\textstyle\int_{\mathbb{R}^d\times\mathbb{R}} \dot{\Phi}\big(y - \mathbb{E}^m[\hat{\varphi}(X, z)]\big)\nabla\hat{\varphi}(x, z)\,\nu(\mathrm{d}z, \mathrm{d}y)\,. \end{aligned}$$

Then it becomes straightforward to verify that F satisfies both Assumption 2.1, 2.6. The rest of the result is direct from Proposition 2.7 and Theorem 2.10. ∎

3.2 Deep neural network

As we have seen, in the 1-hidden layer case, we linearize the problem by lifting the minimization problem to the measure space. We present two examples of deep artificial neural networks. Example 3.4 shows that this linearization technique cannot be applied to all deep fully connected artificial neural networks. However Example 3.5 shows that there are deep artificial neural networks where the linearization technique applies.

Example 3.4 (Fully connected 2-hidden layers neural network) *Take $L = 3$. Let $\beta^3 = \beta^2 = \beta^1 = 0$, $\alpha^1 \in \mathbb{R}^{l^1 \times l^0}$, $\alpha^2 \in \mathbb{R}^{l^2 \times l^1}$. Let $\alpha^3 = (\frac{c_1^3}{l^3}, \cdots, \frac{c_{l^3}^3}{l^3})^\top$, where $c_i^3 \in \mathbb{R}$. Let $\alpha_{ij}^2 := \frac{c_{ij}^2}{l^2}$. The reconstruction for $\Psi^{l^3,l^2} = \Big((\alpha^1,\beta^1),(\alpha^2,\beta^2),(\alpha^3,\beta^3)\Big)$ can be written as*

$$(\mathcal{R}\Psi^{l^3,l^2})(z) = \alpha^3\varphi^{l_2}(\alpha^2\varphi^{l_1}(\alpha^1 z)) = \frac{1}{l^3}\sum_{i=1}^{l^3} c_i^3\varphi\Big(\frac{1}{l^2}\sum_{j=1}^{l^2} c_{ij}^2\varphi(\alpha_j^1 z)\Big)\,.$$

Let $\mu^{I,l^2} := \frac{1}{l^2}\sum_{j=1}^{l^2}\delta_{\{\alpha^2,\alpha^1\}}$ be a conditional empirical law, conditioning on a random variable I uniformly distributed on $\{c_1^3, c_2^3, \ldots, c_{l^3}^3\}$. Then we may write

$$(\mathcal{R}\Psi^{l^3,l^2})(z) = \frac{1}{l^3}\sum_{i=1}^{l^3} c_i^3\varphi\Big(\frac{1}{l^2}\sum_{j=1}^{l^2} c_{ij}^2\varphi(\alpha_j^1 z)\Big) = \mathbb{E}^I\left[c^I\varphi\left(\int_{\mathbb{R}\times\mathbb{R}^{l_0}} y^2\varphi(y^1 z)\mu^{I,l_1}(\mathrm{d}y^1,\mathrm{d}y^2)\right)\right].$$

In general this will not longer be a convex function of measure.

Below we present an example of deep neural network where the last layer is an average of output of fully connected deep neural networks.

Example 3.5 (Averaging fully connected deep artificial neural networks) *Consider $n \in \mathbb{N}$ fully connected artificial neural networks with L hidden layers and identical architectures denoted*

$$\Psi^{(i)} = \Big(\big(\alpha^{(i,1)},\beta^{(i,1)}\big),\big(\alpha^{(i,2)},\beta^{(i,2)}\big),\ldots,\big(\alpha^{(i,L)},\beta^{(i,L)}\big)\Big) \in \Pi, i = 1,\ldots,n\,.$$

We will now construct an artificial neural network Ψ which will be the average of the above n networks. To that end let

$$\alpha^1 := \big(\alpha^{(1,1)},\alpha^{(2,1)},\ldots,\alpha^{(n,1)}\big)^\top \in \mathbb{R}^{n\,l^1\times l^0}\,,\quad \alpha^L := \frac{1}{n}\big(\alpha^{(1,L)},\alpha^{(2,L)},\ldots,\alpha^{(n,L)}\big) \in \mathbb{R}^{l^L\times n\,l^{L-1}}$$

and

$$\alpha^i := diag\big(\alpha^{(1,j)},\alpha^{(2,j)},\ldots,\alpha^{(n,j)}\big) \in \mathbb{R}^{n\,l^j\times n\,l^{j-1}}\,,\ \text{for } j = 2,\ldots,L-1\,.$$

Moreover let

$$\beta^i := \big(\beta^{(1,j)},\beta^{(2,j)},\ldots,\beta^{(n,j)}\big)^\top \in \mathbb{R}^{l^L\times n\,L^{L-1}}\,,\ \text{for } j = 1,\ldots,L-1$$

and

$$\beta^L := \frac{1}{n}\sum_{i=1}^n \beta^{(i,L)}\,.$$

Let $\Psi^n := \Big(\big(\alpha^1,\beta^1\big),\big(\alpha^2,\beta^2\big),\ldots,\big(\alpha^L,\beta^L\big)\Big) \in \Pi^n$ (we use n to emphasise the dependence on the number of networks we are averaging over). One may check that for any $z^0 \in \mathbb{R}^{l^0}$ we have

$$(\mathcal{R}\Psi^n)(z^0) = \frac{1}{n}\sum_{i=1}^n(\mathcal{R}\Psi^{(i)})(z_0)\,.$$

Let $m^n = \frac{1}{n}\sum_{i=1}^n \delta_{\left(\alpha^{(i,1)},\beta^{(i,1)}\right),\left(\alpha^{(i,2)},\beta^{(i,2)}\right),\ldots,\left(\alpha^{(i,L)},\beta^{(i,L)}\right)} \in \mathcal{P}(\Pi)$ be the empirical measure over the parameter space fully describing the network Ψ^n. Then for any $z \in \mathbb{R}^{l^0}$ we may write

$$(\mathcal{R}\Psi^n)(z) = \int_\Pi (\mathcal{R}x)(z)\,m^n(\mathrm{d}x)\,.$$

We note that $m^n \mapsto \int_\Pi(\mathcal{R}x)(z)\,m^n(\mathrm{d}x)$ is a linear (and so convex) function of the measure m^n.

In Section 9 we shall show a numerical example comparing the performance of the averaged network and the conventional fully-connected network.

4 Free Energy Function

In this section, we study the properties concerning the minimizer of the free energy function V^σ. First, we prove that V^σ is an approximation of F in the sense of Γ-convergence.

Proof of Proposition 2.3 Let $(\sigma_n)_{n\in\mathbb{N}}$ be a positive sequence decreasing to 0. On the one hand, since F is continuous and $H(m) \geq 0$, for all $m_n \to m$, we have

$$\varliminf_{n\to+\infty} V^{\sigma_n}(m_n) \geq \lim_{n\to+\infty} F(m_n) = F(m).$$

On the other hand, given $m \in \mathcal{P}_2(\mathbb{R}^d)$, since the function

$$h(x) := x\log(x) \tag{4.1}$$

is convex, it follows Jensen's inequality that

$$\int_{\mathbb{R}^d} h(m * f_n)\mathrm{d}x \leq \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h\big(f_n(x-y)\big)m(\mathrm{d}y)\mathrm{d}x = \int_{\mathbb{R}^d} h\big(f_n(x)\big)\mathrm{d}x = \int_{\mathbb{R}^d} h\big(f(x)\big)\mathrm{d}x - d\log(\sigma_n),$$

where f is the heat kernel and $f_n(x) = \sigma_n^{-d} f(x/\sigma_n)$. Besides, we have

$$\int_{\mathbb{R}^d} (m * f_n)\log(g)\mathrm{d}x = -\int_{\mathbb{R}^d} m(\mathrm{d}y)\int_{\mathbb{R}^d} f_n(x)U(x-y)\mathrm{d}x \geq -C\Big(1 + \int_{\mathbb{R}^d} |y|^2 m(\mathrm{d}y)\Big).$$

The last inequality is due to the quadratic growth of U. Therefore

$$\varlimsup_{n\to+\infty} V^{\sigma_n}(m * f_n) \leq F(m) + \varlimsup_{n\to+\infty} \frac{\sigma_n^2}{2}\Big\{\int_{\mathbb{R}^d} h(m * f_n)\mathrm{d}x - \int_{\mathbb{R}^d} (m * f_n)\log(g)\mathrm{d}x\Big\} \leq F(m). \tag{4.2}$$

In particular, given a minimizer $m^{*,\sigma}$ of V^σ, by (4.2) we have

$$\varlimsup_{n\to\infty} F(m^{*,\sigma_n}) \;\leq\; \varlimsup_{n\to\infty} V^\sigma(m^{*,\sigma_n}) \;\leq\; \varlimsup_{n\to+\infty} V^{\sigma_n}(m * f_n) \;\leq\; F(m), \quad \text{for all} \quad m \in \mathcal{P}_2(\mathbb{R}^d).$$

■

In the rest of the section, we shall discuss the first order condition for the minimizer of the function V^σ. We first show an elementary lemma for convex functions on $\mathcal{P}(\mathbb{R}^d)$.

Lemma 4.1 *Under Assumption 2.1, given $m, m' \in \mathcal{P}(\mathbb{R}^d)$, we have*

$$F(m') - F(m) \geq \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m, x)(m' - m)(\mathrm{d}x). \tag{4.3}$$

Proof Define $m^\varepsilon := (1-\varepsilon)m + \varepsilon m'$. Since F is convex, we have

$$\varepsilon\Big(F(m') - F(m)\Big) \geq F(m^\varepsilon) - F(m) = \int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m^s, x)(m' - m)(\mathrm{d}x)\mathrm{d}s$$

Since $\frac{\delta F}{\delta m}$ is bounded and continuous, we obtain (4.3) by the dominant convergence theorem. ■

Proof of Proposition 2.4: *Step 1.* We first prove the existence of minimizer. Clearly there exists $\bar{m} \in \mathcal{P}(\mathbb{R}^d)$ such that $V^\sigma(\bar{m}) < +\infty$. Denote

$$\mathcal{S} := \left\{m : \frac{\sigma^2}{2}H(m) \leq V^\sigma(\bar{m}) - \inf_{m'\in\mathcal{P}(\mathbb{R}^d)} F(m')\right\}.$$

As a sublevel set of the relative entropy H, $\mathcal{S}$ is compact. Together with the lower semi-continuity of V^σ, the minimum of V^σ on $\mathcal{S}$ is attained. Notice that for all $m \notin \mathcal{S}$, we have $V^\sigma(m) \geq V^\sigma(\bar{m})$, so the minimum of V^σ on $\mathcal{S}$ coincides with the global minimum. Further, since V^σ is strictly convex, the minimizer is unique.

Moreover, given $m^* = \arg\min_{m\in\mathcal{P}(\mathbb{R}^d)} V^\sigma(m)$, we know $m^* \in \mathcal{S}$, and thus $H(m^*) < \infty$. Therefore, m^* is absolutely continuous with respect to the Gibbs measure, so also absolutely continuous with respect to the Lebesgue measure.

Step 2. Sufficient condition: Let $m^* \in \mathcal{I}_\sigma$ (defined in (2.1)), in particular, m^* is equivalent to the Lebesgue measure. Let $m \in \mathcal{P}(\mathbb{R}^d)$ such that m is equivalent to the Lebesgue measure (otherwise $V^\sigma(m) = +\infty$), and thus equivalent to the measure m^*. Let

$$f := \frac{\mathrm{d}m}{\mathrm{d}m^*}$$

be the Radon-Nikodym derivative. Let $m^\varepsilon := (1-\varepsilon)m^* + \varepsilon m = (1+\varepsilon(f-1))m^*$ for $\varepsilon > 0$. Recall the function h in (4.1) and note that $h(y) \geq y - 1$ for all $y \in \mathbb{R}^+$. Using (4.3), we obtain

$$\frac{F(m^\varepsilon) - F(m^*)}{\varepsilon} \geq \frac{1}{\varepsilon}\int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m^*,\cdot)(m^\varepsilon - m^*)\,\mathrm{d}x = \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m^*,\cdot)\,(f-1)m^*\,\mathrm{d}x\,.$$

Moreover

$$\begin{aligned}
\frac{\sigma^2}{2\varepsilon}\Big(H(m^\varepsilon) - H(m^*)\Big) &= \frac{\sigma^2}{2\varepsilon}\int_{\mathbb{R}^d}\left(m^\varepsilon \log\frac{m^\varepsilon}{g} - m^*\log\frac{m^*}{g}\right)\mathrm{d}x \\
&= \frac{\sigma^2}{2\varepsilon}\int_{\mathbb{R}^d}(m^\varepsilon - m^*)\log\frac{m^*}{g}\,\mathrm{d}x + \frac{\sigma^2}{2\varepsilon}\int_{\mathbb{R}^d} m^\varepsilon\left(\log\frac{m^\varepsilon}{g} - \log\frac{m^*}{g}\right)\mathrm{d}x \\
&= \frac{\sigma^2}{2}\int_{\mathbb{R}^d}(f-1)m^*\log\frac{m^*}{g}\,\mathrm{d}x + \frac{\sigma^2}{2\varepsilon}\int_{\mathbb{R}^d} m^\varepsilon\log\frac{m^\varepsilon}{m^*}\,\mathrm{d}x \\
&= \frac{\sigma^2}{2}\int_{\mathbb{R}^d}(f-1)m^*\big(\log m^* + U\big)\,\mathrm{d}x + \frac{\sigma^2}{2\varepsilon}\int_{\mathbb{R}^d} h(1+\varepsilon(f-1))m^*\,\mathrm{d}x \\
&\geq \frac{\sigma^2}{2}\int_{\mathbb{R}^d}(f-1)m^*\big(\log m^* + U\big)\,\mathrm{d}x + \frac{\sigma^2}{2}\int_{\mathbb{R}^d}(f-1)m^*\,\mathrm{d}x \\
&= \frac{\sigma^2}{2}\int_{\mathbb{R}^d}(f-1)m^*\big(\log m^* + U\big)\,\mathrm{d}x
\end{aligned}$$

since $\int_{\mathbb{R}^d}(f-1)m^*\,\mathrm{d}x = \int_{\mathbb{R}^d}(m-m^*)\,\mathrm{d}x = 0$. Hence

$$\frac{V^\sigma(m^\varepsilon) - V^\sigma(m^*)}{\varepsilon} \geq \int_{\mathbb{R}^d}\left(\frac{\delta F}{\delta m}(m^*,\cdot) + \frac{\sigma^2}{2}\log m^* + \frac{\sigma^2}{2}U\right)(f-1)m^*\,\mathrm{d}x = 0$$

due to the fact that $m^* \in \mathcal{I}_\sigma$ and $(f-1)m^* = m - m^*$.

Step 3. Necessary condition: Let m^* be a minimizer of V^σ. Since $H(m^*) < \infty$, we have $\mathbb{E}^{m^*}[U(X)] < \infty$ and thus $m^* \in \mathcal{P}_2(\mathbb{R}^d)$. Let m a probability measure absolutely continuous with respect to m^* such that the Radon-Nikodym derivative $f := \frac{dm}{dm^*}$ is bounded. By the same computation as in the proof for the sufficient condition, we have

$$\begin{aligned}
&\frac{V^\sigma(m^\varepsilon) - V^\sigma(m^*)}{\varepsilon} \\
&= \frac{F(m^\varepsilon) - F(m^*)}{\varepsilon} + \frac{\sigma^2}{2\varepsilon}\int_{\mathbb{R}^d}\left(m^\varepsilon(x)\log\left(\frac{m^\varepsilon(x)}{g(x)}\right) - m^*(x)\log\left(\frac{m^*(x)}{g(x)}\right)\right)\mathrm{d}x \\
&= \frac{1}{\varepsilon}\int_0^\varepsilon\int_{\mathbb{R}^d}\frac{\delta F}{\delta m}(m^s,x)(f-1)m^*(x)\mathrm{d}x\mathrm{d}s \\
&\qquad + \frac{\sigma^2}{2}\int_{\mathbb{R}^d}\left(\log\left(\frac{m^*(x)}{g(x)}\right)(f-1)m^*(x) + m^*(x)\frac{h(1+\varepsilon(f-1))}{\varepsilon}\right)\mathrm{d}x.
\end{aligned}$$

Since f is bounded, $\frac{\delta F}{\delta m}$ is bounded and $m^* \in \mathcal{P}_2(\mathbb{R}^d)$, by the dominated convergence theorem

$$0 \leq \int_{\mathbb{R}^d}(f-1)m^*(x)\left(\frac{\delta F}{\delta m}(m^*,x) + \frac{\sigma^2}{2}\log(m^*(x)) + \frac{\sigma^2}{2}U(x) + \frac{\sigma^2}{2}\right)\mathrm{d}x.$$

Since f is an arbitrary bounded density function, we prove the necessary condition. ∎

5 Mean Field Langevin Equations

Recall that

$$b(x,m) := D_m F(m,x) + \frac{\sigma^2}{2}\nabla U(x).$$

Due to Assumption 2.6 and 2.2, the function b is of linear growth.

Lemma 5.1 *Under Assumption 2.2 and 2.6, let X be the strong solution to (2.2). If $m_0 \in \mathcal{P}_p(\mathbb{R}^d)$ for some $p \geq 2$, we have*

$$\mathbb{E}\Big[\sup_{t\leq T}|X_t|^p\Big] \leq C, \qquad \text{for some } C \text{ depending on } p,\sigma,T. \tag{5.1}$$

If $m_0 \in \mathcal{P}_p(\mathbb{R}^d)$ for some $p \geq 2$, we have

$$\sup_{t\in\mathbb{R}^+}\mathbb{E}\Big[|X_t|^p\Big] \leq C, \qquad \text{for some } C \text{ depending on } p,\sigma. \tag{5.2}$$

In particular, if $m_0 \in \cup_{p>2}\mathcal{P}_p(\mathbb{R}^d)$, then $(m_t)_{t\in\mathbb{R}^+}$ belong to a $\mathcal{W}_2$-compact subset of $\mathcal{P}_2(\mathbb{R}^d)$.

Proof Since b is of linear growth, we have

$$|X_t| \leq |X_0| + \int_0^t C(1+|X_t|)\mathrm{d}t + |\sigma W_t|\,.$$

Therefore,

$$\sup_{t\leq s}|X_t|^p \leq C\left(|X_0|^p + 1 + \int_0^s \sup_{t\leq r}|X_t|^p \mathrm{d}r + \sup_{t\leq s}|\sigma W_t|^p\right).$$

Note that $\mathbb{E}\left[\sup_{t\leq s}|\sigma W_t|^p\right] \leq Cs^{p/2}$. Then (5.1) follows from the Gronwall inequality.

For the second estimate, we apply the Itô formula and obtain

$$\mathrm{d}|X_t|^p = |X_t|^{p-2}\Big(-pX_t\cdot b(X_t,m_t) + \frac{p(p-1)}{2}\sigma^2\Big)\mathrm{d}t + p\sigma|X_t|^{p-2}X_t\cdot\mathrm{d}W_t.$$

Since D_mF is bounded and $\nabla U(x)\cdot x \geq C|x|^2 + C'$, we have

$$\begin{aligned}\mathrm{d}|X_t|^p &\leq |X_t|^{p-2}\Big(C''|X_t| - \frac{p\sigma^2}{2}(C|X_t|^2 + C') + \frac{p(p-1)}{2}\sigma^2\Big)\mathrm{d}t + p\sigma|X_t|^{p-2}X_t\cdot\mathrm{d}W_t\\ &\leq |X_t|^{p-2}\Big(C - \varepsilon|X_t|^2 + \frac{p(p-1)}{2}\sigma^2\Big)\mathrm{d}t + p\sigma|X_t|^{p-2}X_t\cdot\mathrm{d}W_t, \text{ for some } 0<\varepsilon<\tfrac{p\sigma^2C}{2}.\end{aligned}$$

The last inequality is due to the Young inequality. Again by the Itô formula we have

$$\mathrm{d}\big(e^{\varepsilon t}|X_t|^p\big) \;\leq\; e^{\varepsilon t}\left(|X_t|^{p-2}\Big(C + \frac{p(p-1)}{2}\sigma^2\Big)\mathrm{d}t + p\sigma|X_t|^{p-2}X_t\cdot\mathrm{d}W_t\right) \tag{5.3}$$

Further, define the stopping time $\tau_m := \inf\{t\geq 0 : |X_t| \geq m\}$. By taking expectation on both sides of (5.3), we have

$$\mathbb{E}[e^{\varepsilon(\tau_m\wedge t)}|X_{\tau_m\wedge t}|^p] \leq \mathbb{E}[|X_0|^p] + \mathbb{E}\left[\int_0^{\tau_m\wedge t} e^{\varepsilon s}|X_s|^{p-2}\Big(C + \frac{p(p-1)}{2}\sigma^2\Big)\mathrm{d}s\right]. \tag{5.4}$$

In the case $p=2$, it follows from the Fatou lemma and the monotone convergence theorem that

$$\mathbb{E}[|X_t|^2] \leq e^{-\varepsilon t}\mathbb{E}[|X_0|^2] + \int_0^t e^{\varepsilon(s-t)}\Big(C+\sigma^2\Big)\mathrm{d}s \leq C\left(e^{-\varepsilon t} + \varepsilon^{-1}(1-e^{-\varepsilon t})\right),$$

and thus $\sup_{t\in\mathbb{R}^+}\mathbb{E}[|X_t|^2] < \infty$. For $p>2$, we again obtain from (5.4) that

$$\mathbb{E}[|X_t|^p] \leq e^{-\varepsilon t}\mathbb{E}[|X_0|^p] + \int_0^t e^{\varepsilon(s-t)}\mathbb{E}[|X_s|^{p-2}]\Big(C + \frac{p(p-1)}{2}\sigma^2\Big)\mathrm{d}s.$$

Then (5.2) follows from induction. ∎

Proposition 5.2 *Let Assumption 2.2 and 2.6 hold true and assume $m_0 \in \mathcal{P}_p(\mathbb{R}^d)$ for some $p \geq 2$. The law m of the solution X to (2.2) is the unique solution to Fokker-Planck equation:*

$$\partial_t m = \nabla \cdot \left(b(x, m)m + \frac{\sigma^2}{2} \nabla m \right), \tag{5.5}$$

such that $t \mapsto m_t$ in weakly continuous on $[0, +\infty)$, the joint density function $(t, x) \mapsto m(t, x)$ exists and $m \in C^{1,\infty}\left((0, \infty) \times \mathbb{R}^d, \mathbb{R}\right)$.

Proof Let $(t, x) \mapsto \phi(t, x)$ be a smooth test function of compact support. By applying the Itô formula on $\phi(t, X_t)$, we can verify that m is a weak solution to (5.5). Next, define $\tilde{b}(x, t) := b(x, m_t)$. Obviously, m can be regarded as a weak solution to the linear PDE:

$$\partial_t m = \nabla \cdot \left(\tilde{b}m + \frac{\sigma^2}{2} \nabla m \right). \tag{5.6}$$

Then the regularity result follows from a standard argument through L^p_{loc}-estimate. For details, we refer the readers to the seminal paper [38, p.14-p.15] or the classic book [41, Chapter IV]. ■

6 Convergence Towards the Invariant Measure

Now we are going to show that under mild conditions, the flow of marginal law $(m_t)_{t \in \mathbb{R}^+}$ converges toward the invariant measure which coincides with the minimizer of V^σ.

Lemma 6.1 *Suppose Assumption 2.2 and 2.6 hold true and $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Let m be the law of the solution to the mean field Langevin equation (2.2). Denote by $\mathbb{P}_{\sigma,w}$ the scaled Wiener measure[1] with initial distribution m_0. Then,*

i) For any $T > 0$, $\mathbb{P}_{\sigma,w}$ is equivalent to m on $\mathcal{F}_T$, where $\{\mathcal{F}_t\}$ is the filtration generated by X, and the relative entropy

$$\mathbb{E}^m \left[\log \left(\frac{\mathrm{d}m}{\mathrm{d}\mathbb{P}_{\sigma,w}} \Big|_{\mathcal{F}_T} \right) \right] < \infty. \tag{6.1}$$

ii) For all $t > 0$, the marginal law m_t admits density such that $m_t > 0$ and $H(m_t) < \infty$.

Proof i) We shall prove in the Appendix in Lemma 10.1 that due to the linear growth in x of the drift b, $\mathbb{P}_{\sigma,w}$ is equivalent to m. Also by the linear growth of coefficient, we have

$$\mathbb{E}^m \left[\log \left(\frac{\mathrm{d}m}{\mathrm{d}\mathbb{P}_{\sigma,w}} \Big|_{\mathcal{F}_T} \right) \right] = \mathbb{E}^m \left[\frac{1}{\sigma^2} \int_0^T |b(X_t, m_t)|^2 \mathrm{d}t \right] \leq C\mathbb{E}^m \left[1 + \sup_{t \leq T} |X_t|^2 \right] < \infty.$$

The last inequality is due to Lemma 5.1.

ii) Since $\mathbb{P}_{\sigma,w}$ is equivalent to m, we have $m_t > 0$. Denote $f_{\sigma,t}$ the density function of the marginal law of a standard Brownian motion multiplied by σ with initial distribution m_0. It follows from the conditional Jensen inequality that for all $t \in [0, T]$

$$\int_{\mathbb{R}^d} m_t \log \left(\frac{m_t(x)}{f_{\sigma,t}(x)} \right) \mathrm{d}x \leq \mathbb{E}^m \left[\log \left(\frac{\mathrm{d}m}{\mathrm{d}\mathbb{P}_{\sigma,w}} \Big|_{\mathcal{F}_T} \right) \right] < +\infty. \tag{6.2}$$

Further, by the fact $f_{\sigma,t}(x) \leq \frac{1}{(2\pi t)^{d/2}\sigma}$, we have

$$\int_{\mathbb{R}^d} m_t(x) \log(f_{\sigma,t}(x)) \mathrm{d}x \leq -\frac{d}{2} \log(2\pi t \sigma^2).$$

[1] Under the scaled Wiener measure $\mathbb{P}_{\sigma,w}$, if we denote X as the canonical process, $\frac{X}{\sigma}$ is a standard Brownian motion.

Finally, note that

$$-\int_{\mathbb{R}^d} m_t(x)\log\big(g(x)\big)\mathrm{d}x = \int_{\mathbb{R}^d} m_t(x)U(x)\mathrm{d}x \le C\int_{\mathbb{R}^d} m_t(x)|x|^2\mathrm{d}x < \infty.$$

Together with (6.2), we have $H(m_t) < \infty$. ■

Next, we introduce an interesting result of [24, Theorem 3.10 and Remark 4.13].

Lemma 6.2 *Let m be a measure equivalent to the scaled Wiener measure $\mathbb{P}_{\sigma,w}$ such that the relative entropy is finite as in* (6.1)*. Then,*

i) for any $0 < t < T$ we have $\int_t^T \int_{\mathbb{R}^d} |\nabla \log(m_s)|^2 m_s \mathrm{d}x\mathrm{d}s < +\infty$.

ii) given $t \ge t_0 > 0$ such that the Doléans-Dade exponential $\mathcal{E}^b(X) := e^{-\int_{t-t_0}^t \frac{b_s}{\sigma^2}\mathrm{d}X_s - \int_{t-t_0}^t \frac{1}{2}|\frac{b_s}{\sigma}|^2 ds}$ [2] is conditionally $\mathbb{L}^2$-differentiable on the interval $[t-t_0, t]$ [3], we have

$$\nabla \log\big(m_t(x)\big) = -\frac{1}{t_0}\mathbb{E}\Big[\int_0^{t_0}\big(1 + s\nabla b(X_{t-t_0+s}, m_{t-t_0+s})\big)\mathrm{d}W_s^{t-t_0}\Big|X_t = x\Big], \tag{6.4}$$

where $W_s^{t-t_0} := W_{t-t_0+s} - W_{t-t_0}$ and W is the Brownian motion in (2.2)*.*

We shall prove in the Appendix, Lemma 10.2, that under Assumption 2.2 and 2.6, $\mathcal{E}^b$ is conditionally $\mathbb{L}^2$-differentiable on $[t-t_0, t]$ for all $t \ge t_0 > 0$.

The estimate (i) leads to some other integrability results.

Lemma 6.3 *Suppose Assumption 2.2 and 2.6 hold true and $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. We have*

$$\int_t^T\int_{\mathbb{R}^d} |\nabla m_t(x)|\mathrm{d}x\mathrm{d}t < \infty, \quad \int_t^T\int_{\mathbb{R}^d} |x\cdot\nabla m_t(x)|\mathrm{d}x\mathrm{d}t < \infty, \quad \textit{and} \quad \int_t^T\int_{\mathbb{R}^d} |\Delta m_t(x)|\mathrm{d}x\mathrm{d}t < \infty.$$

Proof By the Young inequality, we have

$$|\nabla m_t| \le m_t + \left|\frac{\nabla m_t}{m_t}\right|^2 m_t \quad \text{and} \quad |x\cdot\nabla m_t| \le x^2 m_t + \left|\frac{\nabla m_t}{m_t}\right|^2 m_t.$$

All terms on the right hand sides are integrable, due to Lemma 6.2, therefore so are ∇m and $x\cdot\nabla m$. Next, in order to prove the integrability of Δm, we apply Itô's formula:

$$\mathrm{d}\log(m_t(X_t)) = \left(\tfrac{\partial_t m_t(X_t)}{m_t(X_t)} - \tfrac{\nabla m_t(X_t)}{m_t(X_t)}\cdot b_t(X_t, m_t) - \tfrac{\sigma^2}{2}\left|\tfrac{\nabla m_t}{m_t}(X_t)\right|^2 + \tfrac{\sigma^2}{2}\tfrac{\Delta m_t(X_t)}{m_t(X_t)}\right)\mathrm{d}t + \sigma\tfrac{\nabla m_t(X_t)}{m_t(X_t)}\cdot\mathrm{d}W_t.$$

Together with the Fokker-Planck equation (5.5), we have

$$\begin{aligned}&\log(m_t(X_t)) - \log(m_s(X_s))\\ &\qquad = \int_s^t\left(\sigma^2\frac{\Delta m_r}{m_r}(X_r) - \frac{\sigma^2}{2}\left|\frac{\nabla m_r}{m_r}(X_r)\right|^2 + \nabla\cdot b_r(X_r, m_r)\right)\mathrm{d}r + \int_s^t \frac{\nabla m_r(X_r)}{m_r(X_r)}\sigma\mathrm{d}W_r.\end{aligned} \tag{6.5}$$

[2] Again, we slightly abuse the notation, using X to denote the canonical process of the Wiener space.

[3] Denote by $\mathbb{P}_{\sigma,w}^{t-t_0,x_0}$ the conditional probability of $\mathbb{P}_{\sigma,w}$ given $X_{t-t_0} = x_0$. $\mathcal{E}^b(X)$ is conditionally $\mathbb{L}^2$-differentiable on the interval $[t-t_0, t]$, if there exists an absolutely continuous process $D\mathcal{E}^b := \int_{t-t_0}^{\cdot} D\mathcal{E}_s^b ds$ with $D\mathcal{E}_s^b \in \mathbb{L}^2(\mathbb{P}_{\sigma,w}^{t-t_0,x_0})$ for all $x_0 \in \mathbb{R}^d$ such that for any $h := \int_{t-t_0}^{\cdot}\dot{h}_s ds$ with bounded predictable $\dot{h}$, we have

$$\lim_{\varepsilon\to 0}\Big|\frac{\mathcal{E}^b(X+\varepsilon h) - \mathcal{E}^b(X)}{\varepsilon} - \langle D\mathcal{E}^b(X), h\rangle\Big| = 0, \quad \text{in } \mathbb{L}^2(\mathbb{P}_{\sigma,w}^{t-t_0,x_0}) \text{ for all } x_0 \in \mathbb{R}^d, \tag{6.3}$$

where $\langle D\mathcal{E}^b(X), h\rangle = \int_{t-t_0}^t \dot{h}_s D\mathcal{E}_s^b(X) ds$.

By Lemma 6.2, we have

$$\mathbb{E}\left[\int_s^t \frac{\nabla m_t(X_t)}{m_t(X_t)}\sigma \mathrm{d}W_t\right] = 0.$$

Also recall that $\nabla \cdot b$ is of linear growth. Taking expectation on both side of (6.5), we obtain

$$\mathbb{E}\left[\int_s^t \sigma^2 \frac{|\Delta m_r|}{m_r}(X_r)\right] dr \le H(m_t) + H(m_s) + \mathbb{E}\left[\int_s^t \left(\frac{\sigma^2}{2}\left|\frac{\nabla m_r}{m_r}(X_r)\right|^2 + C(1+|X_r|)\right)\mathrm{d}r\right].$$

By Lemma 6.1 and Lemma 6.2, the right hand side is finite. ∎

Based on the previous integrability results, the next lemma follows from the integration by part.

Lemma 6.4 *Let $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Under Assumption 2.2 and 2.6 we have for Leb-a.s. t that*

$$\int_{\mathbb{R}^d} \mathrm{Tr}(\nabla D_m F(m_t, x)) m_t \mathrm{d}x = -\int_{\mathbb{R}^d} D_m F(m_t, x) \cdot \nabla m_t \mathrm{d}x,$$
$$\int_{\mathbb{R}^d} \nabla \cdot \frac{\nabla m_t}{m_t} m_t \mathrm{d}x = -\int_{\mathbb{R}^d} \left|\frac{\nabla m_t}{m_t}\right|^2 m_t \mathrm{d}x \quad \textit{and} \quad \int_{\mathbb{R}^d} \Delta U(x) m_t(x) \mathrm{d}x = -\int_{\mathbb{R}^d} \nabla U(x) \cdot \nabla m_t(x) \mathrm{d}x.$$

Proof of Theorem 2.8 By the Itô-type formula given by [13, Theorem 4.14] and Lemma 6.4, we have

$$\begin{aligned}\mathrm{d}F(m) &= \int_{\mathbb{R}^d}\left(-|D_m F(m_t,x)|^2 - \frac{\sigma^2}{2} D_m F(m_t,x)\cdot \nabla U(x) + \frac{\sigma^2}{2}\mathrm{Tr}(\nabla D_m F(m_t,x))\right) m_t \mathrm{d}x\mathrm{d}t \\ &= \int_{\mathbb{R}^d}\left(-|D_m F(m_t,x)|^2 - \frac{\sigma^2}{2} D_m F(m_t,x)\cdot \left(\nabla U(x) + \frac{\nabla m_t}{m_t}\right)\right) m_t \mathrm{d}x\mathrm{d}t. \end{aligned} \tag{6.6}$$

On the other hand, by (6.5), Itô's formula and Lemma 6.4, we have

$$\begin{aligned}&\mathrm{d}\big(\log(m_t(X_t)) + U(X_t)\big) - \mathrm{d}M_t \\ &= \left(\sigma^2 \frac{\Delta m_t}{m_t}(X_t) - \frac{\sigma^2}{2}\left|\frac{\nabla m_t}{m_t}(X_t)\right|^2 + \nabla \cdot b_t(X_t, m_t) - \nabla U(X_t)\cdot b(X_t, m_t) + \frac{\sigma^2}{2}\Delta U(X_t)\right)\mathrm{d}t, \\ &= \left(\sigma^2 \nabla \cdot \frac{\nabla m_t}{m_t}(X_t) + \frac{\sigma^2}{2}\left|\frac{\nabla m_t}{m_t}(X_t)\right|^2 + \nabla \cdot b_t(X_t, m_t) - \nabla U(X_t)\cdot b(X_t, m_t) + \frac{\sigma^2}{2}\Delta U(X_t)\right)\mathrm{d}t.\end{aligned}$$

where $dM_t = \left(\frac{\nabla m_t}{m_t}(X_t) + X_t\right)\cdot \sigma \mathrm{d}W_t$ is a martingale on $[s,T]$ for any $0 < s < T$. By taking expectation on both sides and using Lemma 6.4, we obtain for $t > 0$:

$$\begin{aligned}&\mathrm{d}H(m_t) \\ &= \mathbb{E}\left[-\frac{\sigma^2}{2}\left|\frac{\nabla m_t}{m_t}(X_t)\right|^2 - b_t(X_t,m_t)\cdot \frac{\nabla m_t}{m_t}(X_t) - \nabla U(X_t)\cdot b(X_t,m_t) - \frac{\sigma^2}{2}\nabla U(X_t)\cdot \frac{\nabla m_t}{m_t}(X_t)\right]\mathrm{d}t \\ &= \mathbb{E}\left[-\frac{\sigma^2}{2}\left|\frac{\nabla m_t}{m_t}(X_t) + \nabla U(X_t)\right|^2 - D_m F(X_t, m_t)\cdot\left(\frac{\nabla m_t}{m_t}(X_t) + \nabla U(X_t)\right)\right]\mathrm{d}t \\ &= \int_{\mathbb{R}^d}\left(-\frac{\sigma^2}{2}\left|\frac{\nabla m_t}{m_t} + \nabla U(x)\right|^2 - D_m F(m_t,x)\cdot\left(\frac{\nabla m_t}{m_t} + \nabla U(x)\right)\right) m_t(x)\mathrm{d}x\end{aligned} \tag{6.7}$$

Summing up the equation (6.6) and (6.7), we obtain (2.4). ∎

In order to prove there exists an invariant measure of (2.2) equal to the minimizer of V^σ, we shall apply Lasalle's invariance principle. Now we simply recall it in our context. Let $(m_t)_{t\in\mathbb{R}^+}$ be the flow of marginal laws of the solution to (2.2), given an initial law m_0. Define a dynamic system $S(t)[m_0] := m_t$. We shall consider the so-called w-limit set:

$$w(m_0) := \Big\{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \quad \text{there exist } t_n \to \infty \text{ such that } \ \mathcal{W}_2\Big(S(t_n)[m_0], \mu\Big) \to 0\Big\}$$

Proposition 6.5 *[Invariance Principle] Let Assumption 2.6 hold true and assume that $m_0 \in \cup_{p>2}\mathcal{P}_p(\mathbb{R}^d)$. Then the set $w(m_0)$ is nonempty, compact and invariant, that is,*

i) for any $\mu \in w(m_0)$, we have $S(t)[\mu] \in w(m_0)$ for all $t \in \mathbb{R}^+$.
ii) for any $\mu \in w(m_0)$ and all $t \in \mathbb{R}^+$, there exits $\mu' \in w(m_0)$ such that $S(t)[\mu'] = \mu$.

Proof Under the upholding assumptions, it follows from Proposition 2.7 that $S(t)$ is continuous with respect to the $\mathcal{W}_2$-topology. By Lemma 5.1, we have (5.2) with $p > 2$, and thus $\big(S(t)[m_0]\big)_{t\in\mathbb{R}^+} = (m_t)_{t\in\mathbb{R}^+}$ live in a $\mathcal{W}_2$-compact subset of $\mathcal{P}_2(\mathbb{R}^d)$. The desired result follows from the invariance principle, see e.g. [30, Theorem 4.3.3]. In order to keep the paper self-contained, we state the proof as follows.

First, for any $t \geq 0$, $(m_s)_{s\geq t}$ is relatively compact, hence $\overline{(m_s)_{s\geq t}}$ is compact. Since the arbitrary intersection of closed sets is closed, the set

$$w(m_0) = \bigcap_{t\geq 0} \overline{(m_s)_{s\geq t}}$$

is compact.

Next, let $\mu \in w(m_0)$, by definition we know that there exists a sequence $(t_N)_{N>0}$ such that $S(t_N)[m_0] \to \mu$. Let $t \in \mathbb{R}^+$, by the continuity of $S(t) : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$, we have $S(t + t_N)[m_0] \to S(t)[\mu]$ and therefore $S(t)[\mu] \in w(m_0)$.

Finally, for the second point, let $t \in \mathbb{R}^+$ and consider the sequence $\big(S(t_N - t)[m_0]\big)_N$. Since $(m_t)_{t\in\mathbb{R}^+}$ live in a $\mathcal{W}_2$-compact subset of $\mathcal{P}_2(\mathbb{R}^d)$, there exists a subsequence $(t_{N'})$ and $\mu' \in w(m_0)$ such that $S(t_{N'} - t)[m_0] \to \mu'$. Again, by the continuity of $S(t)$, we have $S(t)[\mu'] = \lim_{N'\to\infty} S(t_{N'} - t + t)m_0 = \mu$. ∎

Proof of Theorem 2.10 *Step 1.* We first prove the existence of a convergent subsequence. Since $w(m_0)$ is compact, there exists $\tilde{m} \in \arg\min_{m\in w(m_0)} V^\sigma(m)$. By Proposition 6.5, for $t > 0$ there exists a probability measure $\mu \in w(m_0)$ such that $S(t)[\mu] = \tilde{m}$. By Theorem 2.8, for any $s > 0$ we have

$$V^\sigma\big(S(t + s)[\mu]\big) \leq V^\sigma(\tilde{m}).$$

Since $w(m_0)$ is invariant, $S(t + s)[\mu] \in w(m_0)$ and thus $V^\sigma\big(S(t + s)[\mu]\big) = V^\sigma(\tilde{m})$. Again by Theorem 2.8, we obtain

$$0 = \frac{dV^\sigma\big(S(t)[\mu]\big)}{dt} = -\int_{\mathbb{R}^d} \left| D_m F(\tilde{m}, x) + \frac{\sigma^2}{2}\frac{\nabla\tilde{m}}{\tilde{m}}(x) + \frac{\sigma^2}{2}\nabla U(x)\right|^2 \tilde{m}(x)\mathrm{d}x.$$

Since $\tilde{m} = S(t)[\mu]$ is equivalent to the Lebesgue measure (Proposition 6.1), we have

$$D_m F(\tilde{m}, \cdot) + \frac{\sigma^2}{2}\frac{\nabla\tilde{m}}{\tilde{m}} + \frac{\sigma^2}{2}\nabla U = 0. \tag{6.8}$$

The probability measure $\tilde{m}$ is an invariant measure of (2.2), because it is a stationary solution to the Fokker-Planck equation (5.5). Meanwhile, by Proposition 2.4 we have $\tilde{m} = m^*$.

Step 2. Let $(m_{t_n})_n$ be the subsequence converging to m^*. We are going to prove that $V^\sigma(m^*) = \lim_{n\to\infty} V^\sigma(m_{t_n})$. It is enough to prove $\int_{\mathbb{R}^d} m^* \log(m^*)\mathrm{d}x = \lim_{n\to\infty} \int_{\mathbb{R}^d} m_{t_n} \log(m_{t_n})\mathrm{d}x$. By the lower-semicontinuity of entropy, it is sufficient to prove that

$$\int_{\mathbb{R}^d} m^* \log(m^*)\mathrm{d}x \geq \overline{\lim_{n\to\infty}} \int_{\mathbb{R}^d} m_{t_n} \log(m_{t_n})\mathrm{d}x \tag{6.9}$$

By (6.8), we know that $-\log m^*$ is semi-convex, so we may apply the HWI inequality in [50, Theorem 3]:

$$\int_{\mathbb{R}^d} m_{t_n}\Big(\log(m_{t_n}) - \log(m^*)\Big)\mathrm{d}x \leq \mathcal{W}_2(m_{t_n}, m^*)\Big(\sqrt{I_n} + C\mathcal{W}_2(m_{t_n}, m^*)\Big), \tag{6.10}$$

where I_n is the relative Fisher information defined as

$$\begin{aligned} I_n &:= \mathbb{E}\left[\left|\nabla \log\Big(m_{t_n}(X_{t_n})\Big) - \nabla \log\Big(m^*(X_{t_n})\Big)\right|^2\right] \\ &= \mathbb{E}\left[\left|\nabla \log\Big(m_{t_n}(X_{t_n})\Big) + \frac{2}{\sigma^2} D_m F(m^*, X_{t_n}) + \nabla U(X_{t_n})\right|^2\right]. \end{aligned} \tag{6.11}$$

We are going to show that $\sup_n I_n < \infty$. First, since $D_m F$ is bounded and ∇U is of linear growth, by Lemma 5.1 we have

$$\sup_n \mathbb{E}\left[\left|\frac{2}{\sigma^2} D_m F(m^*, X_{t_n}) + \nabla U(X_{t_n})\right|^2\right] < \infty. \tag{6.12}$$

Next, since ∇b is bounded, by Lemma 10.2 and (6.4) we have for all n

$$\begin{aligned} \mathbb{E}\left[\left|\nabla \log\Big(m_{t_n}(X_{t_n})\Big)\right|^2\right] &\le \inf_{0<s\le t_n} \frac{1}{s^2}\int_0^s C(1+r^2)\,\mathrm{d}r \\ &= \inf_{0<s\le t_n} C\Big(\frac{1}{s} + \frac{s}{3}\Big) \le \frac{2C}{\sqrt{3}}, \qquad \text{for } t_n > \sqrt{3}, \end{aligned} \tag{6.13}$$

where the constant C does not depend on n. Combining (6.11), (6.12) and (6.13) we obtain $\sup_n I_n < \infty$. Now the HWI inequality (6.10) reads

$$\int_{\mathbb{R}^d} m_{t_n}\Big(\log(m_{t_n}) - \log(m^*)\Big)\mathrm{d}x \le C\mathcal{W}_2(m_{t_n}, m^*)\Big(1 + \mathcal{W}_2(m_{t_n}, m^*)\Big).$$

By letting $n \to \infty$, since $\mathcal{W}_2(m_{t_n}, m^*) \to 0$, we obtain (6.9).

Step 3. Finally we prove the convergence of the whole sequence $(m_t)_{t\in\mathbb{R}^+}$. Since $V^\sigma(m_t)$ is non-increasing in t, there is a constant $c := \lim_{t\to\infty} V^\sigma(m_t)$. A subsequence of a convergent sequence converges to the same limit, by the result of *Step 2*, $c = V^\sigma(m^*)$. For any $\mu \in w(m_0)$ there is a subsequence $(m_{t_n})_n$ converging to $\mu \in w(m_0)$ and by lower-semicontinuity we have $V^\sigma(\mu) \le \varliminf_{n\to\infty} V(m_{t_n}) = c$. Because $m^* = \tilde{m} = \operatorname{argmin}_{m\in w(m_0)} V^\sigma(m)$, we have

$$V^\sigma(\mu) = V^\sigma(m^*) = c, \quad \text{for all } \mu \in w(m_0),$$

and thus $w(m_0) = \{m^*\}$, that is, $\lim_{t\to\infty} \mathcal{W}_2(m_t, m^*) = 0$. ∎

7 Exponential Convergence

In this section, we show the exponential convergence under Assumption 2.11.

Lemma 7.1 *Fix $p > 2$. If Assumption 2.11 holds true, then there is a unique solution to (2.2) for all $t \ge 0$ such that*

$$\sup_{t\ge 0} \mathbb{E}|X_t|^p < \infty \tag{7.1}$$

and moreover if $(X_t)_{t\ge0}$ and $(X'_t)_{t\ge0}$ are two solutions to (2.2) then

$$\mathbb{E}\Big[|X_t - X'_t|^2\Big] \le e^{(6C_F - C_U)t}\mathbb{E}\Big[|X_0 - X'_0|^2\Big]. \tag{7.2}$$

This lemma is proved by verifying existence of a suitable Lyapunov function and then applying results from [27]. The proof can be found in the Appendix.

Proof of Theorem 2.12. We can argue as in *Step 1* of the proof of Theorem 2.10 that there exists $m^* \in \mathcal{P}_2(\mathbb{R}^d)$ which is the unique minimizer of the free energy function as well as an invariant measure of

(2.2), i.e. $S(t)[m^*] = m^*$. Let $(X'_t)_{t\geq 0}$ denote the solution of (2.2) starting from $X'_0 \sim m^*$, then $X'_t \sim m^*$ for all $t \geq 0$. By Lemma 7.1 we get that

$$\mathcal{W}_2(m_t, m^*) \leq \mathbb{E}\Big[|X_t - X'_t|^2\Big] \leq e^{(6C_F - C_U)t}\mathbb{E}\Big[|X_0 - X'_0|^2\Big]$$

which completes the proof by taking infimum over all couplings of m_0 and m^*. ■

Remark 7.2 *Let Assumption 2.11 hold but with the global monotonicity condition* (2.6) *replaced with the following monotonicity at infinity condition: fix $R > 0$ and assume that there exits a constant $M > 0$ such that for any $x, x' \notin [-R, R]^d$*

$$\big(\nabla U(x) - \nabla U(x')\big) \cdot (x - x') \geq M|x - x'|^2 . \tag{7.3}$$

For fixed $\sigma > 0$, let $Y_t = X_{t/\sigma^2}$ for all $t \geq 0$. It's clear that $\lim_{t\to\infty} \mathrm{Law}(Y_t) = \lim_{t\to\infty} \mathrm{Law}(X_t)$. Dynamics of Y reads

$$\mathrm{d}Y_t = -\Big(\frac{1}{\sigma^2} D_m F(m_t^Y, Y_t) + \frac{1}{2}\nabla U(Y_t)\Big)\mathrm{d}t + \mathrm{d}W_t, \quad \textit{where} \;\; m_t^Y := \mathrm{Law}(Y_t).$$

Our aim is to prove the exponential convergence, applying [23, Theorem 2.1]. Fix $\delta > 0$. Let $\kappa : (0, \infty) \to [0, \infty)$ be a piecewise-linear continuous function such that $\kappa(r) = -M$ for $r > 2R$ and $\kappa(r) = C$ for $r \leq 2R - \delta$. Clearly $\overline{\lim}_{r\to\infty} \kappa(r) < 0$. We see that Assumption 2.1 in [23, Theorem 2.1] holds with κ as defined above. In addition (7.3) *corresponds to Assumption 2.6 in [23, Theorem 2.1]. While authors state Assumption 2.6 for linear function of measure, it is easy to see that Lipchitz continuity suffices. Note that this has been done, in the case of Levy noise, in [43, Theorem 1.3]. Consequently there exits $\sigma^* > 0$ such that for all $\sigma \in [\sigma^*, \infty)$ there exist positive constants c and λ, that are independent of time, such that*

$$\mathcal{W}_1(m_t, m^*) \leq ce^{-\frac{\lambda}{\sigma^2}t}\left(1 + \int_{\mathbb{R}^d} |x|\, m_0(\mathrm{d}x)\right).$$

Finally note that the contraction rate in the above observation typically degenerates with the increase of dimension d, see [23, Section 3.3].

8 Particle System Approximation

8.1 Static case

Next we aim to generalise [46, Proposition 2.1] to arbitrary (smooth) functions of measure. This is slight generalisation of the result from [17, Theorem 2.11] but we provide the proof here for readers' convenience.

Theorem 8.1 *We assume that the 2nd order linear functional derivative of F exists, is jointly continuous in both variables and that there is $L > 0$ such that for any random variables η_1, η_2 such that $\mathbb{E}[|\eta_i|^2] < \infty$, $i = 1, 2$, it holds that*

$$\mathbb{E}\left[\sup_{\nu\in\mathcal{P}_2(\mathbb{R}^d)}\left|\frac{\delta F}{\delta m}(\nu, \eta_1)\right|\right] + \mathbb{E}\left[\sup_{\nu\in\mathcal{P}_2(\mathbb{R}^d)}\left|\frac{\delta^2 F}{\delta m^2}(\nu, \eta_1, \eta_2)\right|\right] \leq L \tag{8.1}$$

If there is an $m^* \in \mathcal{P}_2(\mathbb{R}^d)$ such that $F(m^*) = \inf_{m\in\mathcal{P}_2(\mathbb{R}^d)} F(m)$ then with i.i.d $(X_i^*)_{i=1}^N$ such that $X_i^* \sim m^*$, $i = 1, \ldots, N$ we have that

$$\left|\mathbb{E}\left[F\Big(\frac{1}{N}\sum_{i=1}^N \delta_{X_i^*}\Big)\right] - F(m^*)\right| \leq \frac{2L}{N} \quad \textit{and} \quad \left|\inf_{(x_i)_{i=1}^N \subset \mathbb{R}^d} F\left(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}\right) - F(m^*)\right| \leq \frac{2L}{N}.$$

Proof Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ be arbitrary. Let $(X_i)_{i=1}^N$ be i.i.d. with law μ. Let $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ and $m_t^N = \mu + t(\mu_N - \mu)$, $t \in [0,1]$. Further let $(\tilde{X}_i)_{i=1}^N$ be consider i.i.d., independent of $(X_i)_{i=1}^N$ with law μ.
By the definition of linear functional derivatives, we have

$$\begin{aligned}\mathbb{E}[F(\mu_N)] - F(\mu) &= \mathbb{E}\left[\int_0^1 \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m_t^N, v)\,(\mu_N - \mu)(\mathrm{d}v)\mathrm{d}t\right] \\ &= \int_0^1 \frac{1}{N}\sum_{i=1}^N \left(\mathbb{E}\left[\frac{\delta F}{\delta m}(m_t^N, X_1)\right] - \mathbb{E}\left[\frac{\delta F}{\delta m}(m_t^N, \tilde{X}_1)\right]\right)\mathrm{d}t \\ &= \int_0^1 \mathbb{E}\left[\frac{\delta F}{\delta m}(m_t^N, X_1) - \frac{\delta F}{\delta m}(m_t^N, \tilde{X}_1)\right]\mathrm{d}t\,. \end{aligned} \tag{8.2}$$

We introduce the (random) measures

$$\tilde{m}_t^N := m_t^N + \frac{t}{N}(\delta_{\tilde{X}_1} - \delta_{X_1}) \quad \text{and} \quad m_{t,t_1}^N := (\tilde{m}_t^N - m_t^N)t_1 + m_t^N, \quad t, t_1 \in [0,1],$$

and notice that due to independence of $(X_i)_{i=1}^N$ and $(\tilde{X}_i)_{i=1}^N$ we have that

$$\mathbb{E}\left[\frac{\delta F}{\delta m}(\tilde{m}_t^N, \tilde{X}_1)\right] = \mathbb{E}\left[\frac{\delta F}{\delta m}(m_t^N, X_1)\right].$$

Therefore,

$$\begin{aligned}\mathbb{E}[F(\mu_N) - F(\mu)] &= \int_0^1 \mathbb{E}\left[\frac{\delta F}{\delta m}(\tilde{m}_t^N, \tilde{X}_1) - \frac{\delta F}{\delta m}(m_t^N, \tilde{X}_1)\right]\mathrm{d}t \\ &= \int_0^1 \mathbb{E}\left[\int_0^1 \int_{\mathbb{R}^d} \frac{\delta^2 F}{\delta m^2}(m_{t,t_1}^N, \tilde{X}_1, y_1)(\tilde{m}_t^N - m_t^N)(\mathrm{d}y_1)\,\mathrm{d}t_1\right]\mathrm{d}t \\ &= \frac{1}{N}\mathbb{E}\left[\int_0^1 \int_0^1 \int_{\mathbb{R}^d} t\frac{\delta^2 F}{\delta m^2}(m_{t,t_1}^N, \tilde{X}_1, y_1)(\delta_{\tilde{X}_1} - \delta_{X_1})(\mathrm{d}y_1)\,\mathrm{d}t_1\mathrm{d}t\right]. \end{aligned} \tag{8.3}$$

To conclude, we observe that

$$\begin{aligned}&\mathbb{E}\left[\left|\int_{\mathbb{R}^d} \frac{\delta^2 F}{\delta m^2}(m_{t,t_1}^N)(\tilde{X}_1, y_1)(\delta_{\tilde{X}_1} - \delta_{X_1})(\mathrm{d}y_1)\right|\right] \\ &= \mathbb{E}\left[\left|\int_{\mathbb{R}^d} \frac{\delta^2 F}{\delta m^2}(m_{t,t_1}^N)(\tilde{X}_1, y_1)\delta_{\tilde{X}_1}(\mathrm{d}y_1) - \int_{\mathbb{R}^d} \frac{\delta^2 F}{\delta m^2}(m_{t,t_1}^N)(\tilde{X}_1, y_1)\delta_{X_1}(\mathrm{d}y_1)\right|\right] \\ &\le \mathbb{E}\left[\sup_{\nu\in\mathcal{P}_2(\mathbb{R}^d)}\left|\frac{\delta^2 F}{\delta m^2}(\nu)(\tilde{X}_1, \tilde{X}_1)\right| + \sup_{\nu\in\mathcal{P}_2(\mathbb{R}^d)}\left|\frac{\delta^2 F}{\delta m^2}(\nu)(\tilde{X}_1, X_1)\right|\right] \le 2L\,, \end{aligned}$$

by (8.1). We have thus shown that for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, for all i.i.d. $(X_i)_{i=1}^N$ with law μ and with $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ it holds that

$$|\mathbb{E}[F(\mu_N)] - F(\mu)| \le \frac{2L}{N}\,. \tag{8.4}$$

From (8.4) with i.i.d $(X_i^*)_{i=1}^N$ such that $X_i^* \sim m^*$, $i = 1, \dots, N$ we get that

$$\left|\mathbb{E}\left[F\left(\frac{1}{N}\sum_{i=1}^N \delta_{X_i^*}\right)\right] - F(m^*)\right| \le \frac{2L}{N}\,.$$

Let $(X_i^*)_{i=1}^N$ be i.i.d. such that $X_i^* \sim m^*$, $i = 1, \dots, N$. Note that

$$F(m^*) \le \inf_{(x_i)_{i=1}^N \subset \mathbb{R}^d} F\left(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}\right) \le \mathbb{E}\left[F\left(\frac{1}{N}\sum_{i=1}^N \delta_{X_i^*}\right)\right].$$

From this and (8.4) we then obtain

$$0 \le \inf_{(x_i)_{i=1}^N \subset \mathbb{R}^d} F\left(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}\right) - F(m^*) \le \frac{2L}{N}\,.$$

■

8.2 Dynamic case

Consider independent random variables $(X_0^i)_{i=1}^N$, $X_0^i \sim m_0$ and independent Brownian motions $(W^i)_{i=1}^N$. By approximating the law of the process (2.2) by its empirical law we arrive at the following interacting particle system

$$\begin{cases} \mathrm{d}X_t^i = & -\Big(D_m F(m_t^N, X_t^i) + \frac{\sigma^2}{2}\nabla U(X_t^i)\Big)\mathrm{d}t + \sigma \mathrm{d}W_t^i \;\; i = 1, \dots, N\,, \\ m_t^N = & \frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}\,. \end{cases} \tag{8.5}$$

Note that particles $(X^i)_{i=1}^N$ are not independent, but their laws are exchangeable. Recall the link between partial derivatives and measure derivative given by (1.7) and for any $(x^1, \dots, x^N) \in (\mathbb{R}^d)^N$ let $F^N(x^1, \dots, x^N) = F\left(\frac{1}{N}\sum_{i=1}^N \delta_{x^i}\right)$. Then

$$\mathrm{d}X_t^i = -\Big(N\partial_{x_i} F^N(X_t^1, \dots, X_t^N) + \frac{\sigma^2}{2}\nabla U(X_t^i)\Big)\mathrm{d}t + \sigma \mathrm{d}W_t^i\,.$$

Let us define, for $x = (\beta, \alpha) \in \mathbb{R}^d$, $\beta \in \mathbb{R}$, $\alpha \in \mathbb{R}^{d-1}$ and $z \in \mathbb{R}^{d-1}$ the function $\hat{\varphi}(x, z) := \ell(\beta)\varphi(\alpha \cdot z)$. Then for $(x^i)_{i=1}^N$ we have

$$F^N(x) = \int_{\mathbb{R}^d} \Phi\Big(y - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}(x^j, z)\Big)\,\nu(\mathrm{d}z, \mathrm{d}y).$$

Hence

$$\partial_{x^i} F^N(x^1, \dots, x^N) = -\frac{1}{N}\int_{\mathbb{R}^d} \dot{\Phi}\Big(y - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}(x^j, z)\Big)\nabla\hat{\varphi}(x^i, z)\nu(\mathrm{d}z, \mathrm{d}y)\,,$$

where we denote for all $z \in \mathbb{R}^{d-1}$

$$\nabla\hat{\varphi}(x^i, z) = \nabla_{(\beta^i, \alpha^i)}[\ell(\beta^i)\varphi(\alpha^i z)] = \begin{pmatrix} \dot{\ell}(\beta^i)\varphi(\alpha^i \cdot z) \\ \ell(\beta^i)\dot{\varphi}(\alpha^i \cdot z)z \end{pmatrix}.$$

We thus see that (8.5) corresponds to

$$\mathrm{d}X_t^i = \left(\int_{\mathbb{R}^d} \dot{\Phi}\Big(y - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}(X_t^j, z)\Big)\nabla\hat{\varphi}(X_t^i, z)\nu(\mathrm{d}z, \mathrm{d}y) - \frac{\sigma^2}{2}\nabla U(X_t^i)\right)\mathrm{d}t + \sigma \mathrm{d}W_t^i.$$

This is classical Langevin dynamics (1.1) on $(\mathbb{R}^d)^N$. One may reasonably expect that the a version of Theorem (8.1) can be proved in this dynamical setup. This has been done for finite time horizon problem in [17]. The extension to the infinite horizon requires uniform in time regularity of the corresponding PDE on Wasserstein space $(\mathcal{W}_2, \mathcal{P}_2)$ and we leave it for a future research. However rate for uniform propagation of chaos in $\mathcal{W}_1$ under structural condition on the drift has been proved in [21]. We also remark that for the implementable algorithm one works with time discretisation of (8.5) and, at least for the finite time, the error bounds are rather well understood [10, 9, 44, 54, 55].

For a fixed time step $\tau > 0$ fixing a grid of time points $t_k = k\tau$, $k = 0, 1, \dots$ we can then write the explicit Euler scheme

$$\begin{aligned} &X_{t_{k+1}}^{\tau,i} - X_{t_k}^{\tau,i} \\ &\qquad = \left(\int_{\mathbb{R}^d} \dot{\Phi}\Big(y - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}(X_{t_k}^{\tau,j}, z)\Big)\nabla\hat{\varphi}(X_{t_k}^{\tau,i}, z)\nu(\mathrm{d}z, \mathrm{d}y) - \frac{\sigma^2}{2}\nabla U(X_{t_k}^{\tau,i})\right)\tau + \sigma(W_{t_{k+1}}^i - W_{t_k}^i)\,. \end{aligned}$$

To relate this to the gradient descent algorithm we consider the case where we are given data points $(y_m, z_m)_{n\in\mathbb{N}}$ which are i.i.d. samples from ν. If the loss function Φ is simply the square loss then a version of the (regularized) gradient descent algorithm for the evolution of parameter x_k^i will simply read as

$$x_{k+1}^i = x_k^i + 2\tau\left(\Big(y_k - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}(x_k^j, z_k)\Big)\nabla\hat{\varphi}(x_k^i, z^k) - \frac{\sigma^2}{2}\nabla U(x_k^i)\right) + \sigma\sqrt{\tau}\xi_k^i\,,$$

with ξ_k^i independent samples from $N(0, I_d)$.

9 Numerical Example of Averaging Deep Artificial Neural Networks

We have discussed in Section 3.2 that averaging deep neural networks fit in our theoretical framework while the fully connected neural networks do not. Here we present a numerical example to compare the performances of these two different architectures. We aim at justifying that the averaging neural network is a reasonable alternative to the fully connected one.

Artificial neural networks are an effective tool for approximating partial differential equations (PDEs) in high dimensions. See, for example, Beck et al. [3] and [4], E et al. [22], Han et al. [28] or Sabate et al. [59]. In fact part of the reason for this success is that deep artificial neural networks approximate solutions to, in particular, parabolic PDEs to an arbitrary accuracy without suffering from the curse of dimensionality. The first mathematically rigorous proof is given in Grohs et al. [26]. See also Jentzen at al. [36]. Examining [26] one can see that the resulting network has precisely the architecture of Example 3.5.

We now provide a sketch of how such architecture arises. Consider

$$\begin{cases} \partial_t v + \operatorname{tr}(a\partial_x^2 v) + b\partial_x v = 0 \ \text{ in } [0,T)\times D\,, \\ v(T,\cdot) = g \ \text{ on } D\,. \end{cases} \tag{9.1}$$

Here $a := \frac{1}{2}\sigma\sigma^*$ and b, σ, and g are suitable given functions such that $v \in C^{1,2}([0,T]\times D)$. See e.g. [40]. From Feynman–Kac formula we know that with a Markov process X, that is given as the solution to the SDE,

$$\mathrm{d}X_s = b(X_s)\,\mathrm{d}s + \sigma(X_s)\,\mathrm{d}W_s \ \ t\in[t,T]\,, \ \ X_t = x \tag{9.2}$$

we have that

$$v(t,x) := \mathbb{E}[g(X_T)|X_t = x]\,.$$

Take a partition of $[0,T]$ denoted by

$$\pi := \{t = t_0 < \cdots < t_{N_{\text{steps}}} = T\}\,, \tag{9.3}$$

let $\Delta W_{t_k} := W_{t_k} - W_{t_{k-1}}$ for $k = 1,\dots,N_{\text{steps}}$, let $W^\pi := (W_{t_i})_{i=1}^{N_{\text{steps}}}$ and consider an approximation of (9.2) by $(X^\pi_{t_i})_{i=0}^{N_{\text{steps}}}$ which itself arises as

$$X^\pi_{t_{k+1}} = G(X^\pi_{t_k}, \Delta W_{t_{k+1}})$$

where $G : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is a measurable function. For example in the case of the Euler scheme this would be

$$G(x,y) := x + b(x)h + \sigma(x)y\,.$$

We now make a simplifying assumption that there exist deep networks $\Psi^{(b)}$ and $\Psi^{(\sigma)}$ such that for all $x\in\mathbb{R}^d$ we have $b(x) = (\mathcal{R}\Psi^{(b)})(x)$ and $\sigma(x) = (\mathcal{R}\Psi^{(\sigma)})(x)$. It can be shown that sums of deep neural networks is again a deep neural network and composition of deep neural networks is again a deep neural network and the identity function can be approximated by a neural network (all having the same activation function). See e.g. Grohs et al. [26] for details. Hence, given W^π one can construct a deep neural network $\Psi^{(W^\pi)}$ such $X^{\Pi}_T = (\mathcal{R}\Psi^{(W^\pi)})(X_0)$. Note that (some of) the weights in $\Psi^{(W^\pi)}$ depend on W^π by construction.

Assume further that there is a network $\Psi^{(g)}$ such that for all $x\in\mathbb{R}^d$ we have $g(x) = (\mathcal{R}\Psi^{(g)})(x)$. An approximation for $v(t,x)$ is then given as follows: first note that

$$v(t,x) = \mathbb{E}[g(X_T)] \approx \mathbb{E}[g(X^\pi_T)] = \mathbb{E}[(\mathcal{R}\Psi^{(g)})(X^\pi_T)] = \mathbb{E}[(\mathcal{R}\Psi^{(g)})\circ(\mathcal{R}\Psi^{(W^\pi)})(x)]\,.$$

Consider now n i.i.d. samples $(W^{\pi,(j)})_{j=1}^n$ from W^π. Then

$$v(t,x) = \frac{1}{n}\sum_{j=1}^n (\mathcal{R}\Psi^{(g)})\circ(\mathcal{R}\Psi^{(W^{\pi,(j)})})(x)\,.$$

This is exactly of the form of Example 3.5.

We now provide a numerical experiment based on Sabate et al. [59], see Example 6.1. therein. An artificial neural network is used for approximating a solution to a PDE arising in mathematical finance (pricing of an exchange option).

The experiment here is the following: given a fixed computational budget (i.e. same overall memory requirements, same number of iterations of gradient descent algorithm) is it significantly better to:

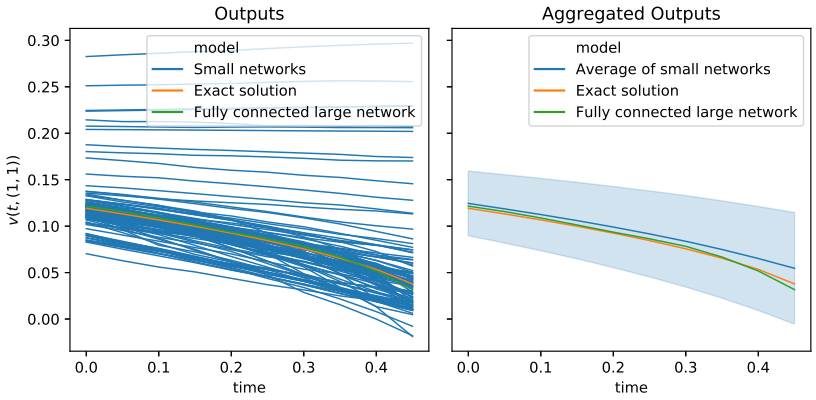

Fig. 1 One complex network trained for longer or an average of many trained with the same computational budget?

i) train a larger fully-connected neural network for longer or
ii) take the average of many (in this case 100) smaller networks that are trained for shorter time?

In Figure 1 we can see that while individual small networks trained for short period of time don't always perform well on their own, their average is a very good approximation of the true solution (and similar to that provided by one complex network that has been trained for a long time).

10 Appendix

The following result regarding to the change of measure in the Wiener space is classic, see e.g. [6]. For readers' convenience, we provide a transparent proof as follow. Our argument is largely inspired by the one in [7, Lemma 4.1.1].

Lemma 10.1 *Let a function $(t,x)\mapsto b(t,x)$ be Lipschitz continuous and of linear growth in x, and a process X be the strong solution to the SDE:*

$$\mathrm{d}X_t = b(t,X_t)\mathrm{d}t + \sigma \mathrm{d}W_t.$$

Define the following Doléan-Dade exponential for all $t\in\mathbb{R}^+$

$$\rho_t := \exp\left(\frac{1}{\sigma}\int_0^t b(s,X_s)\mathrm{d}W_s - \frac{1}{2\sigma^2}\int_0^t |b(s,X_s)|^2\mathrm{d}s\right). \tag{10.1}$$

Then we have $\mathbb{E}[\rho_t]=1$ and thus ρ is a martingale on any finite horizon.

Proof First, we shall prove that there exists $C>0$ such that for all $t\in\mathbb{R}^+$, we have

$$\mathbb{E}[\rho_t|X_t|^2] < C. \tag{10.2}$$

By Itô's formula, we have

$$\mathrm{d}|X_t|^2 = (2X_t b(t,X_t)+\sigma^2)\mathrm{d}t + 2X_t\sigma\mathrm{d}W_t,$$

and

$$\mathrm{d}(\rho_t|X_t|^2) = \rho_t\Big(4X_t b(t,X_t)+\sigma^2\Big)\mathrm{d}t + \rho_t\left(\frac{1}{\sigma}|X_t|^2 b(t,X_t)+2X_t\sigma\right)\mathrm{d}W_t,$$

and further

$$\begin{aligned}\mathrm{d}\frac{\rho_t|X_t|^2}{1+\varepsilon\rho_t|X_t|^2} &= \frac{\rho_t}{(1+\varepsilon\rho_t|X_t|^2)^2}\left(\frac{1}{\sigma}|X_t|^2 b(t,X_t)+2X_t\sigma\right)\mathrm{d}W_t\\ &\quad+\frac{\rho_t}{(1+\varepsilon\rho_t|X_t|^2)^2}\Big(4X_t b(t,X_t)+\sigma^2\Big)\mathrm{d}t\\ &\quad-\frac{\varepsilon\rho_t^2}{(1+\varepsilon\rho_t|X_t|^2)^3}\left|\frac{1}{\sigma}|X_t|^2 b(t,X_t)+2X_t\sigma\right|^2\mathrm{d}t.\end{aligned}$$

Note that the integrand of the stochastic integral on the right hand side above is bounded, so the stochastic integral is actually a real martingale. Therefore, by taking the expectation on both sides and using the fact that b has linear growth in x, we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\frac{\rho_t|X_t|^2}{1+\varepsilon\rho_t|X_t|^2}\right] &\le \mathbb{E}\left[\frac{\rho_t}{(1+\varepsilon\rho_t|X_t|^2)^2}\Big(4X_t b(t,X_t)+\sigma^2\Big)\right]\\ &\le K\mathbb{E}\left[\frac{\rho_t|X_t|^2}{1+\varepsilon\rho_t|X_t|^2}+1\right].\end{aligned}$$

By Grönwall inequality, we get

$$\mathbb{E}\left[\frac{\rho_t|X_t|^2}{1+\varepsilon\rho_t|X_t|^2}\right]\le C,$$

for some constant C which does not depend on ε. By Fatou's lemma, we get (10.2).

Next, by Itô's formula, we have

$$\mathrm{d}\frac{\rho_t}{1+\varepsilon\rho_t}=\frac{\rho_t b(t,X_t)}{(1+\varepsilon\rho_t)^2}\mathrm{d}W_t-\frac{\varepsilon\rho_t^2 b(t,X_t)^2}{(1+\varepsilon\rho_t)^3}\mathrm{d}t.$$

By (10.2), the stochastic integral above is a martingale, so taking the expectation on both sides, we get

$$\mathbb{E}\left[\frac{\rho_t}{1+\varepsilon\rho_t}\right]=\frac{1}{1+\varepsilon}-\int_0^t\mathbb{E}\left[\frac{\varepsilon\rho_s^2 b(s,X_s)^2}{(1+\varepsilon\rho_s)^3}\right]\mathrm{d}s.$$

Due to the linear growth of b, the term inside the expectation on the right hand side is bounded by $C\rho_s(|X_s|^2+1)$ for some constant $C>0$ independent of ε. By the dominated convergence theorem, we get

$$\lim_{\varepsilon\to 0}\mathbb{E}\left[\frac{\rho_t}{1+\varepsilon\rho_t}\right]=1.$$

To conclude, one only needs to note that $\lim_{\varepsilon\to 0}\mathbb{E}\left[\frac{\rho_t}{1+\varepsilon\rho_t}\right]=\mathbb{E}[\rho_t]$.

Lemma 10.2 *Under Assumption 2.2 and 2.6, the exponential martingale $\mathcal{E}(b)$ is conditionally $\mathbb{L}^2$-differentiable on $[t-t_0,t]$, i.e. the equation* (6.3) *holds true, for all $t\ge t_0>0$.*

Proof Without loss of generality, we may assume $t=t_0$. Under the upholding assumptions, the process $(b_t)_{t\in[0,t_0]}$ is $\mathbb{L}^2$-differentiable. By [48, Lemma 1.3.4], we know that $\zeta(X):=-\int_0^{t_0}\frac{b_s}{\sigma^2}\mathrm{d}X_s-\int_0^{t_0}\frac{1}{2}|\frac{b_s}{\sigma}|^2 ds$ is $\mathbb{L}^2$-differentiable for any $t_0>0$, namely there exists $D\zeta$ such that

$$\frac{\zeta(X+\varepsilon h)-\zeta(X)}{\varepsilon}-\langle D\zeta(X),h\rangle\to 0,\quad \text{in } \mathbb{L}^2(\mathbb{P}^{0,x}_{\sigma,w}) \text{ for all } x\in\mathbb{R}^d, \text{ as } \varepsilon\to 0.$$

By Proposition 1.3.8 and Proposition 1.3.11 from [48], we may compute $D\zeta$ explicitly:

$$D\zeta(X)=-\int_0^{t_0}\left(\frac{b_s}{\sigma^2}+\int_s^{t_0}\frac{\nabla b_r}{\sigma^2}(\mathrm{d}X_r+b_r\mathrm{d}r)\right)\mathrm{d}s. \tag{10.3}$$

Note that $\mathcal{E}^b = e^{\zeta}$. Therefore, we have

$$\mathcal{E}^b(X+\varepsilon h) - \mathcal{E}^b(X) = \int_0^{\varepsilon} \langle \mathcal{E}^b(X+sh) D\zeta(X+sh), h\rangle \mathrm{d}s, \quad \mathbb{P}^{0,x}_{\sigma,w}\text{-a.s. for all } x \in \mathbb{R}^d.$$

In order to prove (6.3), it is sufficient to prove that for all $x \in \mathbb{R}^d$

$$\sup_{s\le 1} \mathbb{E}^{\mathbb{P}^{0,x}_{\sigma,w}}\left[\left| \langle \mathcal{E}^b(X+sh) D\zeta(X+sh), h\rangle \right|^p \right] < \infty, \quad \text{for some } p > 2.$$

By the form (10.3), we have $\langle D\zeta(X+sh), h\rangle \in \cap_{q>1} \mathbb{L}^q(\mathbb{P}^{0,x}_{\sigma,w})$, so it is enough to show

$$\mathbb{E}^{\mathbb{P}^{0,x}_{\sigma,w}}\left[\left| \mathcal{E}^b(X) \right|^p \right] < \infty, \quad \text{for some } p > 2. \tag{10.4}$$

Further, note that

$$\left| \mathcal{E}^b(X) \right|^p = e^{-p\int_0^{t_0} \left(\sigma^{-2} D_m F(m_s, X_s) + \nabla U(X_s)\right) \mathrm{d}X_s - \frac{p}{2}\int_0^{t_0} \frac{|b_s|^2}{\sigma^2} \mathrm{d}s}.$$

Since $D_m F$ is bounded, in order to prove (10.4), it is enough to show that

$$\mathbb{E}^{\mathbb{P}^{0,x}_{\sigma,w}}\left[e^{-p\int_0^{t_0} \nabla U(X_s) \mathrm{d}X_s} \right] < \infty, \quad \text{for some } p > 2.$$

By Itô formula, we obtain

$$\mathbb{E}^{\mathbb{P}^{0,x}_{\sigma,w}}\left[e^{-p\int_0^{t_0} \nabla U(X_s) \mathrm{d}X_s} \right] = \mathbb{E}^{\mathbb{P}^{0,x}_{\sigma,w}}\left[e^{-p\left(U(X_{t_0}) - U(x) - \int_0^{t_0} \frac{\sigma^2}{2} \Delta U(X_s) \mathrm{d}s \right)} \right] < \infty,$$

where we use the fact that $U \ge -C$ for some $C > 0$ and ΔU is bounded. ∎

Proof of Lemma 7.1. All integrals in this proof are taken over $\mathbb{R}^d$ unless stated otherwise. We will first show existence of a weak solution. Note that (2.5) and (2.6) imply that for $m \in \mathcal{P}(\mathbb{R}^d), x \in \mathbb{R}^d$ that

$$|D_m F(m,x)| \le C_F\Big(1 + \int |y|\, m(\mathrm{d}y) + |x|\Big) \text{ and } \quad -x \cdot \nabla U(x) \le -C_U |x|^2 + |x||\nabla U(0)|\,. \tag{10.5}$$

Let L denote the diffusion generator for (2.2), that is for any $v \in C^2(\mathbb{R}^d)$ we have

$$L(m,\cdot)v = \frac{\sigma^2}{2}\Delta v - \nabla v \cdot \left(D_m F(m,\cdot) + \frac{\sigma^2}{2} \nabla U(x) \right).$$

Fix $p \ge 2$ and consider now the Lyapunov function $x \mapsto |x|^p$. Note that in this case

$$L(m,x)v(x) = \frac{\sigma^2}{2} p(p-1)|x|^{p-2} - p|x|^{p-2} x \cdot \left(D_m F(m,x) + \frac{\sigma^2}{2} \nabla U(x) \right).$$

Recalling that for any $x \in \mathbb{R}^d$ we have $|x|^{p-2} \le (1 + |x|^p)$ we hence get, for any $m \in \mathcal{P}(\mathbb{R}^d)$, that

$$\begin{aligned}
&\int L(m,x)v(x)\, m(\mathrm{d}x) \\
&\le \frac{\sigma^2}{2} p(p-1) + p \int \Big[\frac{\sigma^2}{2}(p-1)|x|^p - |x|^{p-2} x \cdot D_m F(m,x) - \frac{\sigma^2}{2} |x|^{p-2} x \cdot \nabla U(x) \Big]\, m(\mathrm{d}x)\,.
\end{aligned}$$

Now we observe that due to (10.5) and Hölder's inequality we have

$$\begin{aligned}
p\int |x|^{p-1} |D_m F(m,x)|\, m(\mathrm{d}x) &\le pC_F \int |x|^{p-1} \left(1 + \int |y|\, m(\mathrm{d}y) + |x|\right) m(\mathrm{d}x) \\
&\le pC_F \left(\left(\int |x|^p\, m(\mathrm{d}x) \right)^{(p-1)/p} + 2\int |x|^p\, m(\mathrm{d}x) \right) \\
&\le pC_F \left(1 + 3\int |x|^p\, m(\mathrm{d}x) \right).
\end{aligned}$$

Moreover

$$p\frac{\sigma^2}{2}|\nabla U(0)|\int |x|^{p-1}\,m(\mathrm{d}x) \le p\frac{\sigma^2}{2}|\nabla U(0)|\left(\int |x|^p\,m(\mathrm{d}x)\right)^{(p-1)/p} \le p\frac{\sigma^2}{2}|\nabla U(0)|\left(1+\int |x|^p\,m(\mathrm{d}x)\right).$$

Hence for any $m \in \mathcal{P}(\mathbb{R}^d)$, that

$$\int L(m,x)v(x)\,m(\mathrm{d}x) \le \frac{\sigma^2}{2}p(p-1)+pC_F+p\frac{\sigma^2}{2}|\nabla U(0)| + p\int\Big[\frac{\sigma^2}{2}(p-1)+3C_F+\frac{\sigma^2}{2}|\nabla U(0)|-C_U\frac{\sigma^2}{2}\Big]|x|^p\,m(\mathrm{d}x)\,.$$

This shows that Assumption 2.2 from [27] on the integrated Lyapunov condition holds. Our assumptions also ensure that the initial condition satisfies Assumption 2.3 [27] and our continuity assumptions ensure that Assumption 2.5 from [27] holds. Assumptions 2.6 and 2.7 from [27] hold due to our assumption of linear growth. Using (2.7) and applying Theorem 2.10 from [27] shows that there is a weak solution to (2.2) which satisfies (7.1).

To show the continuous dependence on initial data consider the Lyapunov function $\bar{v}(x)=x^2$ and define

$$L(m,m',x,x')\bar{v}(x-x') = -\Big(D_mF(m,x)-D_mF(m',x')+\nabla U(x)-\nabla U(x')\Big)\cdot(\partial_x\bar{v})(x-x')\,.$$

Let $\pi^{m,m'}$ be any coupling of m,m'. We can see that due to (10.5) we have

$$\begin{aligned}-2\iint &\Big(D_mF(m,x)-D_mF(m',x')\Big)\cdot(x-x')\,\pi^{m,m'}(\mathrm{d}x,\mathrm{d}x')\\ &\le 2C_F\iint\Big(1+|x-x'|+\mathcal{W}_1(m,m')\Big)|x-x'|\,\pi^{m,m'}(\mathrm{d}x,\mathrm{d}x')\\ &\le 6C_F\iint|x-x'|^2\,\pi^{m,m'}(\mathrm{d}x,\mathrm{d}x')\end{aligned}$$

and that due to (2.6) we have

$$\begin{aligned}-\iint &\Big(\nabla U(x)-\nabla U(x')\Big)\cdot(\partial_x\bar{v})(x-x')\pi^{m,m'}(\mathrm{d}x,\mathrm{d}x')\\ &\le -C_U\iint|x-x'|^2\,\pi^{m,m'}(\mathrm{d}x,\mathrm{d}x')\,.\end{aligned}$$

Hence for any coupling $\pi^{m,m'}$ be of m,m' it holds

$$\iint L(m,m',x,x')\bar{v}(x-x')\,\pi^{m,m'}(\mathrm{d}x,\mathrm{d}x') \le (6C_F-C_U)\iint\bar{v}(x-x')\,\pi^{m,m'}(\mathrm{d}x,\mathrm{d}x')\,.$$

This shows that Assumption 3.2 (Integrated Global Lyapunov condition) in [27] holds. Also note that for $\bar{p}=p/2$, $\bar{q}:=\bar{p}/(\bar{p}-1)$ we have

$$|2(x-x')|^{2\bar{p}}+2\sigma^{2\bar{q}} \le c_p(1+|x|^p+|x'|^p) = c_p(1+v(x)+v(x'))\,.$$

Hence we may apply Theorem 3.3 in [27] to conclude that if $(x_t)_{t\ge0}$ and $(x'_t)_{t\ge0}$ are two (weak) solutions to (2.2) then

$$\mathbb{E}\Big[|x_t-x'_t|^2\Big] \le e^{(6C_F-C_U)t}\mathbb{E}\Big[|x_0-x'_0|^2\Big]$$

i.e. we get (7.2). From this the pathwise uniqueness of solutions to (2.2) follows which in turn (together with the weak existence) implies existence of a strong solution by the Yamada-Watanabe principle. ■

References

1. Ambrosio, L., Gigli, N., Savaré, G.: Gradient flows: in metric spaces and in the space of probability measures. Springer (2008)
2. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information theory **39**(3), 930–945 (1993)
3. Beck, C., Becker, S., Grohs, P., Jaafari, N., Jentzen, A.: Solving stochastic differential equations and Kolmogorov equations by means of deep learning. arXiv:1806.00421 (2018)
4. Beck, C., E, W., Jentzen, A.: Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. arXiv:1709.05963 (2017)
5. Belkin, M., Rakhlin, A., Tsybakov, A.B.: Does data interpolation contradict statistical optimality? arXiv:1806.09471 (2018)
6. Benes, V.: Existence of optimal stochastic control laws. Society for Industrial and Applied Mathematics pp. 446–472 (1970). DOI 10.1137/0309034
7. Bensoussan, A.: Stochastic Control of Partially Observable Systems. Cambridge University Press (1992)
8. Bogachev, V.I., Röckner, M., Shaposhnikov, S.V.: Convergence in variation of solutions of nonlinear Fokker–Planck–Kolmogorov equations to stationary measures. Journal of Functional Analysis (2019)
9. Bossy, M., Jourdain, B., et al.: Rate of convergeance of a particle method for the solution of a 1d viscous scalar conservation law in a bounded interval. The Annals of Probability **30**(4), 1797–1832 (2002)
10. Bossy, M., Talay, D.: A stochastic particle method for the McKean–Vlasov and the Burgers equation. Mathematics of Computation of the American Mathematical Society **66**(217), 157–192 (1997)
11. Butkovsky, O.: On ergodic properties of nonlinear Markov chains and stochastic McKean–Vlasov equations. Theory of Probability & Its Applications **58**(4), 661–674 (2014)
12. Cardaliaguet, P., Delarue, F., Lasry, J.M., Lions, P.L.: The master equation and the convergence problem in mean field games. arXiv:1509.02505 (2015)
13. Carmona, R., Delarue, F.: Probabilistic Theory of Mean Field Games with Applications II. Springer (2018)
14. Carrillo, J.A., McCann, R.J., Villani, C.: Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. Revista Matematica Iberoamericana **19**(3), 971–1018 (2003)
15. Cattiaux, P., Guillin, A., Malrieu, F.: Probabilistic approach for granular media equations in the non-uniformly convex case. Probability Theory and Related Fields **140**(1-2), 19–40 (2008)
16. Chassagneux, J.F., Crisan, D., Delarue, F.: A probabilistic approach to classical solutions of the master equation for large population equilibria. arXiv:1411.3009 (2014)
17. Chassagneux, J.F., Szpruch, L., Tse, A.: Weak quantitative propagation of chaos via differential calculus on the space of measures. arXiv:1901.02556 (2019)
18. Chizat, L., Bach, F.: On the global convergence of gradient descent for over-parameterized models using optimal transport. In: Advances in neural information processing systems, pp. 3040–3050 (2018)
19. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems **2**(4), 303–314 (1989)
20. Dupuis, P., Ellis, R.S.: A Weak Convergence Approach to the Theory of Large Deviations. Wiley (1997)
21. Durmus, A., Eberle, A., Guillin, A., Zimmer, R.: An elementary approach to uniform in time propagation of chaos. arX:1805.11387 (2018)
22. E, W., Han, J., Jentzen, A.: Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. Commun. Math. Stat. **5**(4), 349–380 (2017)
23. Eberle, A., Guillin, A., Zimmer, R.: Quantitative Harris-type theorems for diffusions and McKean–Vlasov processes. Transactions of the American Mathematical Society **371**(10), 7135–7173 (2019)
24. Föllmer, H.: Time reversal on Wiener space. In: S.A. Albeverio, P. Blanchard, L. Streit (eds.) Stochastic Processes - Mathematics and Physics, pp. 119–129. Springer (1986)
25. Fontbona, J., Jourdain, B.: A trajectorial interpretation of the dissipations of entropy and Fisher information for stochastic differential equations. Ann. Probab. **44**(1), 131–170 (2016)
26. Grohs, P., Hornung, F., Jentzen, A., von Wurstemberger, P.: A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. arXiv:1809.02362 (2018)
27. Hammersley, W., Šiška, D., Szpruch, L.: McKean–Vlasov SDEs under measure dependent Lyapunov conditions. arXiv:1802.03974 (2018)
28. Han, J., Jentzen, A., E, W.: Solving high-dimensional partial differential equations using deep learning. Proceedings of the National Academy of Sciences (2018)
29. Hastie, T., Montanari, A., Rosset, S., Tibshirani, R.J.: Surprises in high-dimensional ridgeless least squares interpolation. arXiv:1903.08560 (2019)
30. Henry, D.: Geometric Theory of Semilinear Parabolic Equations. Springer (1981)
31. Holley, R., Stroock, D.: Simulated annealing via Sobolev inequalities. Communications in Mathematical Physics **115**(4), 553–569 (1988)
32. Holley, R.A., Kusuoka, S., Stroock, D.W.: Asymptotics of the spectral gap with applications to the theory of simulated annealing. Journal of functional analysis **83**(2), 333–347 (1989)
33. Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural Networks **4**(2), 251–257 (1991)
34. Hwang, C.R., et al.: Laplace's method revisited: weak convergence of probability measures. The Annals of Probability **8**(6), 1177–1182 (1980)
35. Javanmard, A., Mondelli, M., Montanari, A.: Analysis of a two-layer neural network via displacement convexity. arXiv:1901.01375 (2019)

36. Jentzen, A., Salimova, D., Welti, T.: A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of Kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients. arXiv:1809.07321 (2018)
37. Jordan, R., Kinderlehrer, D.: An extended variational principle. In: Partial differential equations and applications: collected papers in honor of Carlo Pucci. CRC (1996)
38. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker–Planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (1998)
39. Karatzas, I., Schachermayer, W., Tschiderer, B.: Pathwise Otto calculus. arXiv:1811.08686 (2019)
40. Krylov, N.V.: On Kolmogorov's equations for finite dimensional diffusions. In: Stochastic PDE's and Kolmogorov Equations in Infinite Dimensions, pp. 1–63. Springer (1999)
41. Ladyzenskaja, O.A., Solonnikov, V.A., Ural'ceva, N.N.: Linear and quasi-linear equations of parabolic type. Translations of Mathematical Monographs. AMS (1968)
42. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
43. Liang, M., Majka, M.B., Wang, J.: Exponential ergodicity for SDEs and McKean–Vlasov processes with Lévy noise. arXiv:1901.11125 (2019)
44. Malrieu, F., et al.: Convergence to equilibrium for granular media equations and their Euler schemes. Annals of Applied Probability **13**(2), 540–560 (2003)
45. Mei, S., Misiakiewicz, T., Montanari, A.: Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. arXiv:1902.06015 (2019)
46. Mei, S., Montanari, A., Nguyen, P.M.: A mean field view of the landscape of two-layer neural networks. Proceedings of the National Academy of Sciences **115**(33), E7665–E7671 (2018)
47. Mischler, S., Mouhot, C.: Kac's program in kinetic theory. Inventiones Mathematicae **193**(1), 1–147 (2013)
48. Nualart, D.: The Malliavin Calculus and Related Topics. Springer (2006)
49. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation (2001)
50. Otto, F., Villani, C.: Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. Journal of Functional Analysis **173**, 361–400 (2000)
51. Rotskoff, G.M., Vanden-Eijnden, E.: Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. arXiv:1805.00915 (2018)
52. Sirignano, J., Spiliopoulos, K.: Mean field analysis of neural networks. arXiv:1805.01053 (2018)
53. Sznitman, A.S.: Topics in propagation of chaos. Springer (1991)
54. Szpruch, L., Tan, S., Tse, A.: Iterative particle approximation for McKean–Vlasov sdes with application to multilevel Monte Carlo estimation. To appear in Annals of Applied Probability (2019)
55. Szpruch, L., Tse, A.: Antithetic multilevel particle system sampling method for McKean–Vlasov SDEs. arXiv:1903.07063 (2019)
56. Tugaut, J.: Convergence to the equilibria for self-stabilizing processes in double-well landscape. The Annals of Probability **41**(3A), 1427–1460 (2013)
57. Vapnik, V.: The nature of statistical learning theory. Springer (2013)
58. Veretennikov, A.Y.: On ergodic measures for McKean–Vlasov stochastic equations. In: Monte Carlo and Quasi-Monte Carlo Methods 2004, pp. 471–486. Springer (2006)
59. Vidales, M.S., Šiška, D., Szpruch, L.: Martingale functional control variates via deep learning. arXiv:1810.05094 (2018)
60. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv:1611.03530 (2016)