Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2021, Vol. 57, No. 4, 2043–2065
https://doi.org/10.1214/20-AIHP1140

ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES

Mean-field Langevin dynamics and energy landscape of neural networks

Kaitong Hu[a], Zhenjie Ren[b], David Šiška[c] and Łukasz Szpruch[c]

[a]*CMAP, École Polytechnique, F-91128 Palaiseau Cedex, France. E-mail: hukaitong@gmail.com*
[b]*CEREMADE, Université Paris Dauphine, F-75775 Paris Cedex 16, France*
[c]*School of Mathematics, University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK*

Received 23 May 2020; revised 7 December 2020; accepted 8 December 2020

Abstract. Our work is motivated by a desire to study the theoretical underpinning for the convergence of stochastic gradient type algorithms widely used for non-convex learning tasks such as training of neural networks. The key insight, already observed in (Mei, Montanari and Nguyen (2018); Chizat and Bach (2018); Rotskoff and Vanden-Eijnden (2018)), is that a 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 study the corresponding gradient flow structure in 2-Wasserstein metric, which we call Mean-Field Langevin Dynamics (MFLD), and show that the flow of marginal laws induced by the gradient flow converges to a stationary distribution, which is exactly the minimiser of the energy functional. We observe that this convergence is exponential under conditions that are satisfied for highly regularised learning tasks. Our proof of convergence to stationary probability measure is novel and it relies on a generalisation of LaSalle's invariance principle combined with HWI inequality. Importantly, we assume neither that interaction potential of MFLD is of convolution type nor that it has any particular symmetric structure. Furthermore, we allow for the general convex objective function, unlike, most papers in the literature that focus on quadratic loss. 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.

Résumé. L'objectif de nos travaux est d'étudier le fondement théorique pour la convergence des algorithmes du type gradient stochastique, qui sont très souvent utilisés dans les problèmes d'apprentissage non-convexe, e.g. calibrer un réseau de neurones. L'observation clé, qui a déjà été remarquée dans (Mei, Montanari and Nguyen (2018); Chizat and Bach (2018); Rotskoff and Vanden-Eijnden (2018)), est qu'une certaine classe de problèmes non-convexes fini-dimensionnels devient convexe une fois injectée dans l'espace des mesures de probabilité. À l'aide de cette observation nous montrons que la fonction d'énergie correspondante définie dans l'espace des mesures de probabilité a un unique minimiser qui peut être caractérisé par une condition de premier ordre en utilisant la notion de dérivée fonctionnelle. Par la suite, nous étudions la structure de flux de gradient avec la métrique de 2-Wasserstein, que nous appelons la dynamique de Langevin au champs moyen (MFLD), et nous montrons que la loi marginale du flux de gradient converge vers une loi stationnaire qui correspond au minimiser de la même fonction d'énergie précédente. Sous certaines conditions de régularité du probléme initial, la convergence a lieu à une vitesse exponentielle. Nos preuves de la convergence vers la loi stationnaire est nouvelle, qui reposent sur le principe d'invariance de LaSalle et l'inégalité HWI. Remarquons que nous ne supposons pas que l'interaction potentielle de MFLD soit du type convolution ou symétrique. De plus, nos résultats s'appliquent aux fonctions d'objectif convexes générales contrairement aux beaucoup d'articles dans la littérature qui se limitent aux fonctions quadratiques. Enfin, nous montrons que la différence entre le probléme initial d'optimisation fini-dimensionnel et sa limite dans l'espace des mesures de probabilité est de l'ordre d'un sur le nombre de paramètres.

MSC2020 subject classifications: 60H30; 37M25

Keywords: Mean-field Langevin dynamics; Gradient flow; Neural networks

1. Introduction

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 [36]. 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 [50] may fail to predict the behaviour of modern machine learning algorithms [52]. 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 [4,24]. 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.

This work builds on the rigorous mathematical framework to study non-convex learning tasks such as training of neural networks developed in Mei, Misiakiewicz and Montanari [39], Chizat and Bach [18], Sirignano and Spiliopoulos [45] as well as Rotskoff and Vanden-Eijnden [44].

We extend some existing results and provide a novel proof technique for mathematical results which provide a theoretical underpinning for the convergence of stochastic gradient type algorithms widely used in practice to train neural networks. We demonstrate how our results apply to a situation when one aims to train one-hidden layer neural network with (noisy) stochastic gradient algorithm.

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

$$dX_t = -\nabla f(X_t)\,dt + \sigma\,dW_t, \tag{1.1}$$

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

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

(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

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

over all probability measure m, where H is the relative entropy with respect to the Lebesgue measure. The variational perspective has been established in [32] and [33]. 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 [27,28,30]

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, [3,19] 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.

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

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

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

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:

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

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

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

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.

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:

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

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

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

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:

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

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*:

$$dX_t = -\left(D_m F(m_t, X_t) + \frac{\sigma^2}{2}\nabla U(X_t)\right)dt + \sigma\, dW_t, \quad \text{where } m_t := \mathrm{Law}(X_t). \tag{1.5}$$

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

$$x_{k+1}^i = x_k^i + 2\tau\left(\left(y_k - \frac{1}{N}\sum_{j=1}^{N}\hat{\varphi}\big(x_k^j, z_k\big)\right)\nabla\hat{\varphi}\big(x_k^i, z^k\big) - \frac{\sigma^2}{2}\nabla U\big(x_k^i\big)\right) + \sigma\sqrt{\tau}\xi_k^i, \tag{1.6}$$

with ξ_k^i independent samples from $N(0, I_d)$ (for details we refer the reader to Section 3.2). This evolution is an approximation of (1.5) and can be viewed as noisy gradient decent. Indeed, in its original form, the classical stochastic gradient decent (also known as the Robins–Monroe algorithm), is given by (1.6) with $\sigma = 0$.

1.1. *Organisation of the paper*

The introduction is concluded by Section 1.2, where we compare the findings in this paper to those available in the literature, and by Section 1.3 recall some basic notions of measure derivatives. All the main results of the paper are presented in Section 2. In Section 3 we show how the results in Section 2 apply to in the case of gradient descent training of neural networks. Section 4 contains all the proofs of the results concerning the free energy function: Γ-convergence when $\sigma \to 0$, particle approximation and the first order condition. In Section 5 we prove required properties of (1.4) and (1.5), Section 6 is used to prove the convergence of the solution to (1.4) to an invariant measure which is the minimizer of the free energy function.

1.2. *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 [40]. We also refer reader to excellent monographs [2] and [1]. In particular, variational approach has been developed in [14,42,49] 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 [8, 15,51], coupling arguments [22] and Khasminskii's technique [7,11]. 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 [43] 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. 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.

Other results are now available for the mean-field description of non-convex learning problems, see [18,31,38,39,44, 45]. Let us compare this paper to the key results available in the literature. There are essentially three, or, if entropic regularization is included, four key ingredients:

(i) that the finite dimensional optimisation problem is approximated by infinite dimensional problem of minimizing over measures (Theorem 2.4),
(ii) that the regularized problem approximates the original minimization problem (Proposition 2.3),
(iii) that on the space of probability measures the minimizers (or, if entropic regularization is included, the unique minimizer) satisfy a first order condition (Proposition 2.5),
(iv) and finally that on the space of probability measures we have a gradient flow that converges with time to the minimizer (Theorem 2.11).

Chizat and Bach [18] work without adding entropic regularization which means that their minimization task is convex but not strictly convex. They have results regarding (i) and (iii). They have a partial result related to (iv) in that they prove that if the gradient flow converges to a limit, as $t \to \infty$ then objective function also converge to global minimiser. To prove this final convergence result they require the assumption that the activation function is homogenous of either order 1 or order 2 and that, essentially, initial law has full support. The setting used in Chizat and Bach [18] is rather different to that of the results in the present paper: in particular we do not need to assume any homogeneity on the activation function and apart mild integrability conditions we do not make assumptions on the initial law. Since we regularize using entropy we obtain convergence of the gradient flow to the minimizer.

Rotskoff and Vanden-Eijnden [44] again work without entropic regularization and have results (i). Moreover they provide Central-Limit-Theorem-type fluctuation results. They do show that the output of the network converges to a limit as the time in the gradient flow for the parameter measure goes to infinity. However they do not prove convergence of the parameter measure itself as in (iv).

Sirignano and Spiliopoulos in [45] provide detail analysis of (i), also studying time-discretisation of continuous time gradient flow. In Section 3.2, by using links between intrinsic derivative on the space of measure and its finite dimensional projection we provide further insight on the choice of scalings needed to derive non-trival limit in [45].

The setting of our paper is the closest to that of Mei, Misiakiewicz and Montanari [39] in that they use the entropic regularization. For a square loss function Φ and a quadratic regularizer U they prove results on all of (i), (ii), (iii) and (iv). In this paper we allow a general loss function for all the above results(that this is possible is conjectured in Appendix B of [39]). Due to the special choice of the square loss function, in [39, Lemma 6.10] the authors can compute directly the dynamics of $F(m_t)$ along the flow of measures defined by (1.4). Instead, we obtain the desired dynamics for much more general F using a pathwise argument based on the Itô calculus (see Theorem 2.9). The proof of (iv) in [39, Lemma 6.12] is based on the Poincaré inequality for the Gaussian distribution and shows that the marginal law weakly converges. It is not clear whether this argument can be extended to the case with a general regularizer U, whereas this paper develops a new technique based on LaSalle's invariance principle and the HWI inequality, which allows us to prove the convergence for general U in the Wasserstein-2 metric, and moreover we observe that in the highly regularized case this convergence is exponential, see Theorem 2.11.

1.3. *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,

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

It is convenient to recall that

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

$$F\big(m'\big) - F(m) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}\big((1-\lambda)m + \lambda m', x\big)\big(m' - m\big)(dx)\, d\lambda. \tag{1.7}$$

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 (1.7). To avoid the ambiguity, we impose

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

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 $F(m) := \int_{\mathbb{R}^d} \phi(x) m(dx)$ 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

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

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

$$\partial_{x^i} F^N\big(x^1, \dots, x^N\big) = \frac{1}{N} D_m F\big(m^N, x^i\big). \tag{1.8}$$

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:

$$\min_{m \in \mathcal{P}(\mathbb{R}^d)} V^\sigma(m), \quad \text{where } V^\sigma(m) := F(m) + \frac{\sigma^2}{2} H(m), \text{ for all } m \in \mathcal{P}\big(\mathbb{R}^d\big), \tag{2.1}$$

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,

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

where

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

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

$$\nabla U(x) \cdot x \geq C_U |x|^2 + C'_U \quad \text{for all } x \in \mathbb{R}^d. \tag{2.2}$$

(ii) ∇U is Lipschitz continuous.

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

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

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*

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

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

The following theorem shows that we can control the error between the finite and infinite-dimensional optimization problems. It generalises [39, Proposition 2.1] to an arbitrary (smooth) functions of measure. It is an extension of the result from [17, Theorem 2.11].

Theorem 2.4. *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*

$$\mathbb{E}\left[\sup_{\nu\in\mathcal{P}_2(\mathbb{R}^d)}\left|\frac{\delta F}{\delta m}(\nu,\eta_1)\right|\right]+\mathbb{E}\left[\sup_{\nu\in\mathcal{P}_2(\mathbb{R}^d)}\left|\frac{\delta^2 F}{\delta m^2}(\nu,\eta_1,\eta_2)\right|\right]\le L \tag{2.3}$$

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 we have that

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

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

Proposition 2.5. *Under Assumption 2.1 and 2.2, the function V^σ has a unique minimizer absolutely continuous with respect to Lebesgue measure ℓ, and belonging to $\mathcal{P}_2(\mathbb{R}^d)$. Moreover, $m^*\in\mathcal{P}_2(\mathbb{R}^d)=\arg\min_{m\in\mathcal{P}(\mathbb{R}^d)}V^\sigma(m)$ if and only if m^* is equivalent to Lebesgue measure and*

$$\frac{\delta F}{\delta m}\big(m^*,\cdot\big)+\frac{\sigma^2}{2}\log\big(m^*\big)+\frac{\sigma^2}{2}U\quad\text{is a constant, }\ell\text{-a.s.,} \tag{2.4}$$

where we abuse the notation, still denoting by m^* the density with respect to Lebesgue measure.

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:

$$dX_t=-\left(D_mF(m_t,X_t)+\frac{\sigma^2}{2}\nabla U(X_t)\right)dt+\sigma\,dW_t, \tag{2.5}$$

where m_t is the law of X_t and $(W_t)_{t\ge0}$ is a standard d-dimensional Brownian motion.

Remark 2.6.

(i) Let $F(m)=\int_{\mathbb{R}^d}f(x)m(dx)$ for some function f in $C^1(\mathbb{R}^d,\mathbb{R})$. We know that $D_mF(m,x)=\nabla f(x)$. Hence with this choice of F and entropy regulariser with respect to the Lebesgue measure, the dynamics (2.5) 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 (2.5) becomes $F(m)+\frac{\sigma^2}{4}\int_{\mathbb{R}^d}|x|^2m(dx)$. This shares the same spirit as ridge regression.

Assumption 2.7. Assume that the intrinsic derivative $D_mF:\mathcal{P}(\mathbb{R}^d)\times\mathbb{R}^d\to\mathbb{R}^d$ of the function $F:\mathcal{P}(\mathbb{R}^d)\to\mathbb{R}$ exists and satisfies the following conditions:

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

$$\big|D_mF(m,x)-D_mF\big(m',x'\big)\big|\le C_F\big(|x-x'|+\mathcal{W}_2\big(m,m'\big)\big) \tag{2.6}$$

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

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

The well-posedness of the McKean–Vlasov SDE (2.5) under Assumption 2.2 and 2.7 on the time interval $[0,t]$, for any t, is well known, see e.g. Snitzman [46], so the proof of the following proposition is omitted.

Proposition 2.8. *Under Assumption 2.2 and 2.7 the mean field Langevin SDE (2.5) 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 (2.5), we have for all $t > 0$ there is a constant $C > 0$ such that*

$$\mathcal{W}_2\big(m_t, m_t'\big) \le C\mathcal{W}_2\big(m_0, m_0'\big).$$

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

Theorem 2.9. *Let $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Under Assumption 2.2 and 2.7, we have for any $t > s > 0$*

$$V^\sigma(m_t) - V^\sigma(m_s) = -\int_s^t \int_{\mathbb{R}^d} \left| D_m F(m_r, x) + \frac{\sigma^2}{2}\frac{\nabla m_r}{m_r}(x) + \frac{\sigma^2}{2}\nabla U(x) \right|^2 m_r(x)\, dx\, dr. \tag{2.7}$$

Remark 2.10. In order to prove (2.7), we use the generalized Itô calculus as the main tool. Alternative proofs of Theorem 2.9 can be obtained under the comparable assumptions using the theory of gradient flows, see e.g. the monograph [1]. Also note that our path-wise argument shares the spirit with the recent work [34], which recovers the results of gradient flows for probability measure on the Euclidean space using the Itô calculus but only for linear functional F.

Formally, there is a clear connection between the derivative $\frac{dV^\sigma(m_t)}{dt}$ in (2.7) and the first order condition (2.4), and it is revealed by the following main theorem.

We call a measure m an invariant measure of (2.5), if $\mathrm{Law}(X_t) = m$ for all $t \ge 0$.

Theorem 2.11. *Let Assumption 2.1, 2.2 and 2.7 hold true and $m_0 \in \bigcup_{p>2} \mathcal{P}_p(\mathbb{R}^d)$. Denote by $(m_t)_{t\in\mathbb{R}^+}$ the flow of marginal laws of the solution to (2.5). There exists an invariant measure of (2.5) equal to $m^* := \arg\min_m V^\sigma(m)$, and $\lim_{t\to\infty} \mathcal{W}_2(m_t, m^*) = 0$.*

Remark 2.12. As mentioned, the main contribution of this paper is to prove the $\mathcal{W}_2$-convergence of the marginal laws of (2.5) towards the invariant measure under the mild conditions (Assumption 2.1, 2.2 and 2.7). Note that it is possible to obtain exponential convergence result with extra conditions on the coefficients. More precisely, given the constants C_F, ρ_F, C_U such that

$$\begin{aligned}
&\big|D_m F(m,x) - D_m F\big(m', x'\big)\big| \le C_F\big|x - x'\big| + \rho_F \mathcal{W}_1\big(m, m'\big) \quad \text{for all } x, x' \in \mathbb{R}^d \text{ and } m, m' \in \mathcal{P}\big(\mathbb{R}^d\big),\\
&\big(x - x'\big)\cdot\big(\nabla U(x) - \nabla U\big(x'\big)\big) \ge C_U\big|x - x'\big|^2 \quad \text{for } \big|x - x'\big| \text{ big enough},
\end{aligned}$$

Eberle, Guillin and Zimmer [22] proved that there exists a constant γ depending on σ, C_F and C_U such that

$$\mathcal{W}_1\big(m_t, m_t'\big) \le Ce^{-(\gamma-\rho_F)t}\mathcal{W}_1\big(m_0, m_0'\big), \tag{2.8}$$

where $(m_t')_{t\in\mathbb{R}^+}, (m_t')_{t\in\mathbb{R}^+}$ are the flows of marginal laws of the solutions to (2.5) with the initial law m_0, m_0', respectively. In particular,

(i) the result (2.8) only implies the exponential contraction provided that ρ_F is small enough, that is, the mean field dependence must be small;
(ii) the constant γ is increasing in σ and C_U, so γ is big only if σ or/and C_U are large, that is, the optimization (2.1) is over-regularized.

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.

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

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

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

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

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

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

$$\arg\min_{\Psi\in\Pi}\frac{1}{N}\sum_{j=1}^{N}\Phi\big(y^j-(\mathcal{R}\Psi)\big(z^j\big)\big). \tag{3.2}$$

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*

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

is dense in $C(K)$.

For an elementary proof, we refer the readers to [29, 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 $\alpha^2=(\frac{c_1}{n},\dots,\frac{c_n}{n})^\top$, where $c_i\in\mathbb{R}$. The neural network is $\Psi^n=((\alpha^1,\beta^1),(\alpha^2,\beta^2))$ (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

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

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:

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

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:

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

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, 2.7. In particular, with a Gibbs measure of which the function U satisfies Assumption 2.2, the corresponding mean field Langevin equation (2.5) 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*

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

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

$$\frac{\delta F}{\delta m}(m, x) = -\int_{\mathbb{R}^d} \dot\Phi\big(y - \mathbb{E}^m\big[\hat\varphi(X, z)\big]\big)\hat\varphi(x, z)\nu(dz, dy) \quad \text{and}$$

$$D_m F(m, x) = -\int_{\mathbb{R}^d} \dot\Phi\big(y - \mathbb{E}^m\big[\hat\varphi(X, z)\big]\big)\nabla\hat\varphi(x, z)\nu(dz, dy).$$

Then it becomes straightforward to verify that F satisfies both Assumption 2.1, 2.7. The rest of the result is direct from Proposition 2.8 and Theorem 2.11. □

3.2. *Gradient descent*

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 (2.5) by its empirical law we arrive at the following interacting particle system

$$\begin{cases} dX_t^i = -(D_m F(m_t^N, X_t^i) + \frac{\sigma^2}{2}\nabla U(X_t^i))\,dt + \sigma\, dW_t^i, \quad i = 1, \dots, N, \\ m_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}. \end{cases} \tag{3.3}$$

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 (1.8) and for any $(x^1, \dots, x^N) \in (\mathbb{R}^d)^N$ let $F^N(x^1, \dots, x^N) = F(\frac{1}{N}\sum_{i=1}^N \delta_{x^i})$. Then

$$dX_t^i = -\left(N\partial_{x_i} F^N\big(X_t^1, \dots, X_t^N\big) + \frac{\sigma^2}{2}\nabla U\big(X_t^i\big)\right)dt + \sigma\, dW_t^i.$$

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

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

Hence

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

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

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

We thus see that (3.3) corresponds to

$$dX_t^i = \left(\int_{\mathbb{R}^d} \dot\Phi\left(y - \frac{1}{N}\sum_{j=1}^N \hat\varphi\big(X_t^j, z\big)\right)\nabla\hat\varphi\big(X_t^i, z\big)\nu(dz, dy) - \frac{\sigma^2}{2}\nabla U\big(X_t^i\big)\right)dt + \sigma\, dW_t^i.$$

This is classical Langevin dynamics (1.1) on $(\mathbb{R}^d)^N$. One may reasonably expect that the a version of Theorem 2.4 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 (3.3) and, at least for the finite time, the error bounds are rather well understood [9,10,37,47,48].

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

$$X^{\tau,i}_{t_{k+1}} - X^{\tau,i}_{t_k} = \left(\int_{\mathbb{R}^d} \dot{\Phi}\left(y - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}\big(X^{\tau,j}_{t_k}, z\big)\right)\nabla\hat{\varphi}\big(X^{\tau,i}_{t_k}, z\big)\nu(dz, dy) - \frac{\sigma^2}{2}\nabla U\big(X^{\tau,i}_{t_k}\big)\right)\tau + \sigma\big(W^i_{t_{k+1}} - W^i_{t_k}\big).$$

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^i_k will simply read as

$$x^i_{k+1} = x^i_k + 2\tau\left(\left(y_k - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}\big(x^j_k, z_k\big)\right)\nabla\hat{\varphi}\big(x^i_k, z^k\big) - \frac{\sigma^2}{2}\nabla U\big(x^i_k\big)\right) + \sigma\sqrt{\tau}\xi^i_k, \quad \text{with } \xi^i_k \sim N(0, I_d) \text{ independent.}$$

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

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

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

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

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

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

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

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

Proof of Theorem 2.4. 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 $m^N_t = \mu + t(\mu_N - \mu)$, $t \in [0, 1]$. 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

$$\begin{aligned}\mathbb{E}\big[F(\mu_N)\big] - F(\mu) &= \mathbb{E}\bigg[\int_0^1 \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}\big(m_t^N, v\big)(\mu_N - \mu)(dv)\,dt\bigg] \\ &= \int_0^1 \frac{1}{N}\sum_{i=1}^N \bigg(\mathbb{E}\bigg[\frac{\delta F}{\delta m}\big(m_t^N, X_1\big)\bigg] - \mathbb{E}\bigg[\frac{\delta F}{\delta m}\big(m_t^N, \tilde{X}_1\big)\bigg]\bigg)\,dt \\ &= \int_0^1 \mathbb{E}\bigg[\frac{\delta F}{\delta m}\big(m_t^N, X_1\big) - \frac{\delta F}{\delta m}\big(m_t^N, \tilde{X}_1\big)\bigg]\,dt. \end{aligned} \tag{4.3}$$

We introduce the (random) measures

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

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

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

Therefore,

$$\begin{aligned}\mathbb{E}\big[F(\mu_N) - F(\mu)\big] &= \int_0^1 \mathbb{E}\bigg[\frac{\delta F}{\delta m}\big(\tilde{m}_t^N, \tilde{X}_1\big) - \frac{\delta F}{\delta m}\big(m_t^N, \tilde{X}_1\big)\bigg]\,dt \\ &= \int_0^1 \mathbb{E}\bigg[\int_0^1 \int_{\mathbb{R}^d} \frac{\delta^2 F}{\delta m^2}\big(m_{t,t_1}^N, \tilde{X}_1, y_1\big)\big(\tilde{m}_t^N - m_t^N\big)(dy_1)\,dt_1\bigg]\,dt \\ &= \frac{1}{N}\mathbb{E}\bigg[\int_0^1 \int_0^1 \int_{\mathbb{R}^d} t\frac{\delta^2 F}{\delta m^2}\big(m_{t,t_1}^N, \tilde{X}_1, y_1\big)(\delta_{\tilde{X}_1} - \delta_{X_1})(dy_1)\,dt_1\,dt\bigg]. \end{aligned} \tag{4.4}$$

To conclude, we observe that

$$\begin{aligned}&\mathbb{E}\bigg[\bigg|\int_{\mathbb{R}^d} \frac{\delta^2 F}{\delta m^2}\big(m_{t,t_1}^N\big)(\tilde{X}_1, y_1)(\delta_{\tilde{X}_1} - \delta_{X_1})(dy_1)\bigg|\bigg] \\ &\quad= \mathbb{E}\bigg[\bigg|\int_{\mathbb{R}^d} \frac{\delta^2 F}{\delta m^2}\big(m_{t,t_1}^N\big)(\tilde{X}_1, y_1)\delta_{\tilde{X}_1}(dy_1) - \int_{\mathbb{R}^d} \frac{\delta^2 F}{\delta m^2}\big(m_{t,t_1}^N\big)(\tilde{X}_1, y_1)\delta_{X_1}(dy_1)\bigg|\bigg] \\ &\quad\le \mathbb{E}\bigg[\sup_{\nu\in\mathcal{P}_2(\mathbb{R}^d)}\bigg|\frac{\delta^2 F}{\delta m^2}(\nu)(\tilde{X}_1, \tilde{X}_1)\bigg| + \sup_{\nu\in\mathcal{P}_2(\mathbb{R}^d)}\bigg|\frac{\delta^2 F}{\delta m^2}(\nu)(\tilde{X}_1, X_1)\bigg|\bigg] \le 2L, \end{aligned}$$

by (2.3). 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

$$\big|\mathbb{E}\big[F(\mu_N)\big] - F(\mu)\big| \le \frac{2L}{N}. \tag{4.5}$$

From (4.5) with i.i.d. $(X_i^*)_{i=1}^N$ such that $X_i^* \sim m^*$, $i = 1, \dots, N$ we get that

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

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

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

From this and (4.5) we then obtain

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

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*

$$F\big(m'\big) - F(m) \ge \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m,x)\big(m'-m\big)(dx). \tag{4.6}$$

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

$$\varepsilon\big(F\big(m'\big) - F(m)\big) \ge F\big(m^\varepsilon\big) - F(m) = \int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}\big(m^s, x\big)\big(m'-m\big)(dx)\,ds$$

Since $\frac{\delta F}{\delta m}$ is bounded and continuous, we obtain (4.6) by the dominant convergence theorem. $\square$

Proof of Proposition 2.5. *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

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

As a sublevel set of the relative entropy H, $\mathcal{S}$ is weakly compact, see e.g. [20, Lemma 1.4.3]. Together with the weak 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 $V^\sigma(m) \ge V^\sigma(\bar m)$, 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 we have $H(m^*) < \infty$ as well as $\mathbb{E}^{m^*}[U(X)] < \infty$. Therefore, $m^* \in \mathcal{P}_2(\mathbb{R}^d)$ is absolutely continuous with respect to the Gibbs measure, so also absolutely continuous with respect to the Lebesgue measure.

Step 2. Sufficient condition: Let $m^* \in \mathcal{P}_2(\mathbb{R}^d)$ satisfy (2.4), in particular, m^* is equivalent to the Lebesgue measure. Let $m \in \mathcal{P}(\mathbb{R}^d)$ such that m is absolutely continuous with respect to the Lebesgue measure (otherwise $V^\sigma(m) = +\infty$). Let

$$f := \frac{dm}{dm^*}$$

be the Radon–Nikodym derivative. Let $m^\varepsilon := (1-\varepsilon)m^* + \varepsilon m = (1+\varepsilon(f-1))m^*$ for $\varepsilon > 0$. For the simplicity of the notations, denote $m^\varepsilon(x)$ and $m^*(x)$ the respective density function of m^ε and m^* with respect to Lebesgue measure. Recall the function h in (4.1) and note that $h(y) \ge y - 1$ for all $y \in \mathbb{R}^+$. Using (4.6), we obtain

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

Moreover

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

$$\geq \frac{\sigma^2}{2}\int_{\mathbb{R}^d}(f-1)m^*\big(\log m^*+U\big)\,dx+\frac{\sigma^2}{2}\int_{\mathbb{R}^d}(f-1)m^*\,dx$$
$$=\frac{\sigma^2}{2}\int_{\mathbb{R}^d}(f-1)m^*\big(\log m^*+U\big)\,dx$$

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

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

Step 3. Necessary condition: Let m^* be the minimizer of V^σ. Let m a probability measure such that $H(m)<\infty$, in particular m is also absolutely continuous with respect to Lebesgue measure ℓ. As above, denote $m(x)$ and $m^*(x)$ the respective density function of m and m^* with respect to Lebesgue measure and we have

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

Since h is convex, we note that for all $\varepsilon\in(0,1)$

$$\frac{1}{\varepsilon}\big(h\big(m^\varepsilon(x)\big)-h\big(m^*(x)\big)-\log\big(g(x)\big)\big(m^\varepsilon(x)-m^*(x)\big)\big)\leq m(x)\log\left(\frac{m(x)}{g(x)}\right)-m^*(x)\log\left(\frac{m^*(x)}{g(x)}\right).$$

Since $H(m)$ and $H(m^*)$ are both finite, the right hand side of the above inequality is integrable. Therefore by Fatou's Lemma we obtain

$$0\leq\overline{\lim_{\varepsilon\to0}}\frac{V^\sigma(m^\varepsilon)-V^\sigma(m^*)}{\varepsilon}\leq\int_{\mathbb{R}^d}\left(\frac{\delta F}{\delta m}\big(m^*,x\big)+\frac{\sigma^2}{2}\log\big(m^*(x)\big)+\frac{\sigma^2}{2}U(x)\right)\big(m(x)-m^*(x)\big)\,dx.\tag{4.7}$$

Since m is arbitrary, we first obtain

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

Now suppose that m^* is not equivalent to Lebesgue measure. There exists a set $\mathcal{K}\subset\mathbb{R}^d$ such that $m^*(\mathcal{K})=0$ and $\ell(\mathcal{K})>0$. It follows from (4.7) that $0\leq C-\int_{\mathcal{K}}\infty\,dm$. Since we may choose m having positive mass on $\mathcal{K}$, it is a contradiction. Therefore, m^* is equivalent to Lebesgue measure and we have

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

□

5. Mean field Langevin equations

Recall that

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

Due to Assumption 2.7 and 2.2, the function b is of linear growth.

Lemma 5.1. *Under Assumption 2.2 and 2.7, let X be the strong solution to (2.5). If $m_0\in\mathcal{P}_p(\mathbb{R}^d)$ for some $p\geq2$, we have*

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

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

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

In particular, if $m_0 \in \bigcup_{p>2} \mathcal{P}_p(\mathbb{R}^d)$, 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

$$|X_t| \leq |X_0| + \int_0^t C\big(1 + |X_t|\big)\, dt + |\sigma W_t|.$$

Therefore,

$$\sup_{t\leq s} |X_t|^p \leq C\bigg(|X_0|^p + 1 + \int_0^s \sup_{t\leq r} |X_t|^p\, dr + \sup_{t\leq s} |\sigma W_t|^p\bigg).$$

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

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

$$d|X_t|^p = |X_t|^{p-2}\bigg(-pX_t \cdot b(X_t, m_t) + \frac{p(p-1)}{2}\sigma^2\bigg)\, dt + p\sigma |X_t|^{p-2} X_t \cdot dW_t.$$

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

$$\begin{aligned} d|X_t|^p &\leq |X_t|^{p-2}\bigg(C''|X_t| - \frac{p\sigma^2}{2}\big(C|X_t|^2 + C'\big) + \frac{p(p-1)}{2}\sigma^2\bigg)\, dt + p\sigma |X_t|^{p-2} X_t \cdot dW_t \\ &\leq |X_t|^{p-2}\bigg(C - \varepsilon |X_t|^2 + \frac{p(p-1)}{2}\sigma^2\bigg)\, dt + p\sigma |X_t|^{p-2} X_t \cdot dW_t, \quad \text{for some } 0 < \varepsilon < \frac{p\sigma^2 C}{2}. \end{aligned}$$

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

$$d\big(e^{\varepsilon t}|X_t|^p\big) \leq e^{\varepsilon t}\bigg(|X_t|^{p-2}\bigg(C + \frac{p(p-1)}{2}\sigma^2\bigg)\, dt + p\sigma |X_t|^{p-2} X_t \cdot dW_t\bigg) \tag{5.3}$$

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

$$\mathbb{E}\big[e^{\varepsilon(\tau_m \wedge t)}|X_{\tau_m\wedge t}|^p\big] \leq \mathbb{E}\big[|X_0|^p\big] + \mathbb{E}\bigg[\int_0^{\tau_m\wedge t} e^{\varepsilon s}|X_s|^{p-2}\bigg(C + \frac{p(p-1)}{2}\sigma^2\bigg)\, ds\bigg]. \tag{5.4}$$

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

$$\mathbb{E}\big[|X_t|^2\big] \leq e^{-\varepsilon t}\mathbb{E}\big[|X_0|^2\big] + \int_0^t e^{\varepsilon(s-t)}\big(C + \sigma^2\big)\, ds \leq C\big(e^{-\varepsilon t} + \varepsilon^{-1}\big(1 - e^{-\varepsilon t}\big)\big),$$

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

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

Then (5.2) follows from induction. □

Proposition 5.2. *Let Assumption 2.2 and 2.7 hold true and assume $m_0 \in \mathcal{P}_p(\mathbb{R}^d)$ for some $p \geq 2$. The marginal law m of the solution X to (2.5) is a weak solution to Fokker–Planck equation*:

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

in the sense that for all C^∞-function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that $\phi, \nabla\phi, \nabla^2\phi$ decay to 0 at infinity, we have

$$\int_{\mathbb{R}^d} \phi(x)(m_t - m_s)(dx) = \int_s^t \int_{\mathbb{R}^d} \left(-\nabla\phi(x) b(x, m_u) m_u(dx) + \frac{\sigma^2}{2} \Delta\phi(x) m_u(dx) \right) du.$$

Moreover, the mapping $t \mapsto m_t$ is weakly continuous on $[0, +\infty)$, the joint density function $(t, x) \mapsto m(t, x)$ exists and $m \in C^{1,\infty}((0, \infty) \times \mathbb{R}^d, \mathbb{R})$. In particular, a stationary solution to the Fokker–Planck equation (5.5) is an invariant measure to (2.5).

Proof. 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 [33, p.14–p.15] or the classic book [35, Chapter IV].

Let m^* be a stationary solution to (5.5), and X be the strong solution to the SDE:

$$dX_t = -b\big(X_t, m^*\big)\, dt + \sigma\, dW_t.$$

It is easy to verify that given $\text{Law}(X_0) = m^*$ we have $\text{Law}(X_t) = m^*$ for all $t \geq 0$. Therefore X is the solution to mean-field Langevin equation (2.5) and m^* is an invariant measure. □

6. Convergence to 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 2.7 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 (2.5). 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*

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

(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 A.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

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

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 $t \in [0, T]$

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

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

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

[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

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

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

Next, we introduce an interesting result of [23, 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 $\int_t^T \int_{\mathbb{R}^d} |\nabla \log(m_s)|^2 m_s\,dx\,ds < +\infty$.*

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

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

where $W_s^{t-t_0} := W_{t-t_0+s} - W_{t-t_0}$ and W is the Brownian motion in (2.5).

We shall prove in the Appendix, Lemma A.2, that under Assumption 2.2 and 2.7, $\mathcal{E}^b$ 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 2.7 hold true and $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. We have*

$$\int_t^T \int_{\mathbb{R}^d} \big|\nabla m_t(x)\big|\,dx\,dt < \infty \quad \text{and} \quad \int_t^T \int_{\mathbb{R}^d} \big|x\cdot\nabla m_t(x)\big|\,dx\,dt < \infty.$$

Proof. By the Young inequality, we have

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

Since all terms on the right hand sides are integrable, due to Lemma 6.2, so are ∇m and $x\cdot\nabla m$. □

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 2.7 we have for Leb-a.s. t that*

$$\int_{\mathbb{R}^d} \mathrm{Tr}\big(\nabla D_m F(m_t, x)\big)m_t\,dx = -\int_{\mathbb{R}^d} D_m F(m_t, x)\cdot\nabla m_t\,dx, \quad \text{and}$$
$$\int_{\mathbb{R}^d} \Delta U(x) m_t(x)\,dx = -\int_{\mathbb{R}^d} \nabla U(x)\cdot\nabla m_t(x)\,dx.$$

Again using the estimate (i) in Lemma 6.2, together with Theorem 2.1 of Haussmann and Pardoux [25], we directly obtain the following result concerning the time reverse process $\widetilde{X}_t := X_{T-t}$ for a given $T > 0$ and $t \le T$.

---

[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

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

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

Lemma 6.5. *Under Assumption 2.2 and 2.7, there exists a Brownian motion $\widetilde{W}_t$ such that $(\widetilde{X}, \widetilde{W})$ is a weak solution to the SDE*:

$$d\widetilde{X}_t = \big(b(\widetilde{X}_t, m_{T-t}) + \sigma^2 \nabla \log m_{T-t}(\widetilde{X}_t)\big)\, dt + \sigma\, d\widetilde{W}_t.$$

Proof of Theorem 2.9. By the Itô formula and the Fokker–Plank equation (5.5), we have

$$\begin{aligned} d\log m_{T-t}(\widetilde{X}_t) &= \bigg(-\frac{\partial_t m_{T-t}}{m_{T-t}}(\widetilde{X}_t) + \nabla \log\big(m_{T-t}(\widetilde{X}_t)\big) \cdot \big(b(\widetilde{X}_t, m_{T-t}) + \sigma^2 \nabla \log m_{T-t}(\widetilde{X}_t)\big) \\ &\quad + \frac{1}{2}\sigma^2 \Delta \log\big(m_{T-t}(\widetilde{X}_t)\big)\bigg)\, dt + \nabla \log\big(m_{T-t}(\widetilde{X}_t)\big) \cdot d\widetilde{W}_t \\ &= \bigg(\frac{\sigma^2}{2}\bigg|\frac{\nabla m_{T-t}}{m_{T-t}}(\widetilde{X}_t)\bigg|^2 - \nabla \cdot b(\widetilde{X}_t, m_{T-t})\bigg)\, dt + \nabla \log\big(m_{T-t}(\widetilde{X}_t)\big) \cdot d\widetilde{W}_t. \end{aligned}$$

Next by the Itô formula we obtain

$$dU(X_t) = \bigg(-\nabla U(X_t) \cdot b(X_t, m_t) + \frac{\sigma^2}{2}\Delta U(X_t)\bigg)\, dt + \nabla U(X_t)\, dW_t.$$

Note that $dH(m_t) = d\mathbb{E}[\log m_t(\widetilde{X}_{T-t}) + U(X_t)]$. Therefore, it follows from Lemma 6.4 that

$$\begin{aligned} &dH(m_t) \\ &= \mathbb{E}\bigg[-\frac{\sigma^2}{2}\bigg|\frac{\nabla m_t}{m_t}(X_t)\bigg|^2 - b(X_t, m_t) \cdot \frac{\nabla m_t}{m_t}(X_t) - \nabla U(X_t) \cdot b(X_t, m_t) - \frac{\sigma^2}{2}\nabla U(X_t) \cdot \frac{\nabla m_t}{m_t}(X_t)\bigg]\, dt \\ &= \mathbb{E}\bigg[-\frac{\sigma^2}{2}\bigg|\frac{\nabla m_t}{m_t}(X_t) + \nabla U(X_t)\bigg|^2 - D_m F(X_t, m_t) \cdot \bigg(\frac{\nabla m_t}{m_t}(X_t) + \nabla U(X_t)\bigg)\bigg]\, dt \\ &= \int_{\mathbb{R}^d}\bigg(-\frac{\sigma^2}{2}\bigg|\frac{\nabla m_t}{m_t} + \nabla U(x)\bigg|^2 - D_m F(m_t, x) \cdot \bigg(\frac{\nabla m_t}{m_t} + \nabla U(x)\bigg)\bigg) m_t(x)\, dx \end{aligned} \tag{6.5}$$

Further, by the Itô-type formula given by [13, Theorem 4.14] and Lemma 6.4, we have

$$\begin{aligned} dF(m_t) &= \int_{\mathbb{R}^d}\bigg(-\big|D_m F(m_t, x)\big|^2 - \frac{\sigma^2}{2} D_m F(m_t, x) \cdot \nabla U(x) + \frac{\sigma^2}{2}\operatorname{Tr}\big(\nabla D_m F(m_t, x)\big)\bigg) m_t\, dx\, dt \\ &= \int_{\mathbb{R}^d}\bigg(-\big|D_m F(m_t, x)\big|^2 - \frac{\sigma^2}{2} D_m F(m_t, x) \cdot \bigg(\nabla U(x) + \frac{\nabla m_t}{m_t}\bigg)\bigg) m_t\, dx\, dt. \end{aligned} \tag{6.6}$$

Finally, summing up the equation (6.5) and (6.6), we obtain (2.7). □

In order to prove there exists an invariant measure of (2.5) 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 (2.5), 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:

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

Proposition 6.6. [*Invariance Principle*] *Let Assumption 2.7 hold true and assume that $m_0 \in \bigcup_{p>2}\mathcal{P}_p(\mathbb{R}^d)$. 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 2.8 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 $(S(t)[m_0])_{t\in\mathbb{R}^+} = (m_t)_{t\in\mathbb{R}^+}$ 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. [26, 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 $(S(t_N - t)[m_0])_N$. 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 2.11. *Step* 1. We first prove that $m^* \in w(m_0)$. Since $w(m_0)$ is compact, there exists $\tilde{m} \in \arg\min_{m\in w(m_0)} V^\sigma(m)$. By Proposition 6.6, for $t > 0$ there exists a probability measure $\mu \in w(m_0)$ such that $S(t)[\mu] = \tilde{m}$. By Theorem 2.9, 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 $V^\sigma(S(t + s)[\mu]) = V^\sigma(\tilde{m})$. Again by Theorem 2.9, we obtain

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

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

$$D_m F(\tilde{m}, \cdot) + \frac{\sigma^2}{2}\frac{\nabla\tilde{m}}{\tilde{m}} + \frac{\sigma^2}{2}\nabla U = 0. \tag{6.7}$$

The probability measure $\tilde{m}$ is an invariant measure of (2.5), because it is a stationary solution to the Fokker–Planck equation (5.5). Meanwhile, by Proposition 2.5 we have $\tilde{m} = m^*$. Therefore, $m^* \in w(m_0)$.

Step 2. Since $m^* \in w(m_0)$, there exists a subsequence, denoted by $(m_{t_n})_{n\in\mathbb{N}}$, 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 $\int_{\mathbb{R}^d} m^* \log(m^*)\,dx = \lim_{n\to\infty} \int_{\mathbb{R}^d} m_{t_n} \log(m_{t_n})\,dx$. By the lower-semicontinuity of entropy, it is sufficient to prove that

$$\int_{\mathbb{R}^d} m^* \log\big(m^*\big)\,dx \geq \varlimsup_{n\to\infty} \int_{\mathbb{R}^d} m_{t_n} \log(m_{t_n})\,dx \tag{6.8}$$

By (6.7), we know that $-\log m^*$ is semi-convex, so we may apply the HWI inequality in [43, Theorem 3]:

$$\int_{\mathbb{R}^d} m_{t_n}\big(\log(m_{t_n}) - \log\big(m^*\big)\big)\,dx \leq \mathcal{W}_2\big(m_{t_n}, m^*\big)\big(\sqrt{I_n} + C\mathcal{W}_2\big(m_{t_n}, m^*\big)\big), \tag{6.9}$$

where I_n is the relative Fisher information defined as

$$\begin{aligned} I_n &:= \mathbb{E}\big[\big|\nabla\log\big(m_{t_n}(X_{t_n})\big) - \nabla\log\big(m^*(X_{t_n})\big)\big|^2\big] \\ &= \mathbb{E}\left[\left|\nabla\log\big(m_{t_n}(X_{t_n})\big) + \frac{2}{\sigma^2} D_m F\big(m^*, X_{t_n}\big) + \nabla U(X_{t_n})\right|^2\right]. \end{aligned} \tag{6.10}$$

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

$$\sup_n \mathbb{E}\left[\left|\frac{2}{\sigma^2} D_m F\big(m^*, X_{t_n}\big) + \nabla U(X_{t_n})\right|^2\right] < \infty. \tag{6.11}$$

Next, since ∇b is bounded, by Lemma A.2 and (6.4) we have for all n

$$\begin{aligned} \mathbb{E}\big[\big|\nabla\log\big(m_{t_n}(X_{t_n})\big)\big|^2\big] &\leq \inf_{0<s\leq t_n} \frac{1}{s^2}\int_0^s C\big(1 + r^2\big)\,dr \\ &= \inf_{0<s\leq t_n} C\left(\frac{1}{s} + \frac{s}{3}\right) \leq \frac{2C}{\sqrt{3}}, \quad \text{for } t_n > \sqrt{3}, \end{aligned} \tag{6.12}$$

where the constant C does not depend on n. Combining (6.10), (6.11) and (6.12) we obtain $\sup_n I_n < \infty$. Now the HWI inequality (6.9) reads

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

By letting $n \to \infty$, since $\mathcal{W}_2(m_{t_n}, m^*) \to 0$, we obtain (6.8).

Step 3. Finally we prove the convergence of the whole sequence $(m_t)_{t\in\mathbb{R}^+}$ towards m^*, by showing that the set $w(m_0)$ is a singleton, namely $w(m_0) = \{m^*\}$. Since $V^\sigma(m_t)$ is non-increasing in t, there is a constant $c := \lim_{t\to\infty} V^\sigma(m_t)$. Recall that in *Step 2* we proved $V^\sigma(m^*) = \lim_{n\to\infty} V^\sigma(m_{t_n})$, so we obtain $c = V^\sigma(m^*)$. On the other hand, for any $\mu \in w(m_0)$ there is a subsequence $(m_{t'_n})_{n\in\mathbb{N}}$ converging to μ and by the weak lower-semicontinuity of V^σ we have $V^\sigma(\mu) \le \underline{\lim}_{n\to\infty} V^\sigma(m_{t'_n}) = c$. Using the fact that $m^* = \arg\min_{m\in w(m_0)} V^\sigma(m)$, we have

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

Finally by the uniqueness of the minimiser of V^σ, we have $w(m_0) = \{m^*\}$. □

Appendix

The following result regarding to the change of measure in the Wiener space is classic, see e.g. [5]. For readers' convenience, we provide a transparent proof as follow. Our argument is largely inspired by the one in [6, Lemma 4.1.1].

Lemma A.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*:

$$dX_t = b(t, X_t)\,dt + \sigma\,dW_t.$$

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

$$\rho_t := \exp\bigg(\frac{1}{\sigma}\int_0^t b(s, X_s)\,dW_s - \frac{1}{2\sigma^2}\int_0^t \big|b(s, X_s)\big|^2\,ds\bigg). \tag{A.1}$$

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

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

By Itô's formula, we have

$$d|X_t|^2 = \big(2X_t b(t, X_t) + \sigma^2\big)\,dt + 2X_t\sigma\,dW_t,$$

and

$$d\big(\rho_t|X_t|^2\big) = \rho_t\big(4X_t b(t, X_t) + \sigma^2\big)\,dt + \rho_t\bigg(\frac{1}{\sigma}|X_t|^2 b(t, X_t) + 2X_t\sigma\bigg)\,dW_t,$$

and further

$$\begin{aligned} d\frac{\rho_t|X_t|^2}{1+\varepsilon\rho_t|X_t|^2} &= \frac{\rho_t}{(1+\varepsilon\rho_t|X_t|^2)^2}\bigg(\frac{1}{\sigma}|X_t|^2 b(t, X_t) + 2X_t\sigma\bigg)\,dW_t \\ &\quad + \frac{\rho_t}{(1+\varepsilon\rho_t|X_t|^2)^2}\big(4X_t b(t, X_t) + \sigma^2\big)\,dt \\ &\quad - \frac{\varepsilon\rho_t^2}{(1+\varepsilon\rho_t|X_t|^2)^3}\bigg|\frac{1}{\sigma}|X_t|^2 b(t, X_t) + 2X_t\sigma\bigg|^2\,dt. \end{aligned}$$

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

$$\frac{d}{dt}\mathbb{E}\left[\frac{\rho_t|X_t|^2}{1+\varepsilon\rho_t|X_t|^2}\right] \le \mathbb{E}\left[\frac{\rho_t}{(1+\varepsilon\rho_t|X_t|^2)^2}\big(4X_t b(t,X_t)+\sigma^2\big)\right]$$
$$\le K\mathbb{E}\left[\frac{\rho_t|X_t|^2}{1+\varepsilon\rho_t|X_t|^2}+1\right].$$

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 (A.2).

Next, by Itô's formula, we have

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

By (A.2), the stochastic integral above is a martingale, so taking the expectation on both sides, we get

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

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

Lemma A.2. *Under Assumption 2.2 and 2.7, 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 [41, Lemma 1.3.4], we know that $\zeta(X) := -\int_0^{t_0}\frac{b_s}{\sigma^2}\,dX_s - \int_0^{t_0}\frac{1}{2}|\frac{b_s}{\sigma}|^2\,ds$ is $\mathbb{L}^2$-differentiable for any $t_0>0$, namely there exists $D\zeta$ such that

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

By Proposition 1.3.8 and Proposition 1.3.11 from [41], we may compute $D\zeta$ explicitly:

$$D\zeta(X) = -\int_0^{t_0}\left(\frac{b_s}{\sigma^2}+\int_s^{t_0}\frac{\nabla b_r}{\sigma^2}(dX_r+b_r dr)\right)ds. \tag{A.3}$$

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

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

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

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

By the form (A.3), we have $\langle D\zeta(X+sh),h\rangle \in \bigcap_{q>1}\mathbb{L}^q(\mathbb{P}^{0,x}_{\sigma,w})$, so it is enough to show

$$\mathbb{E}^{\mathbb{P}^{0,x}_{\sigma,w}}\big[\big|\mathcal{E}^b(X)\big|^p\big]<\infty \quad \text{for some } p>2. \tag{A.4}$$

Further, note that

$$\big|\mathcal{E}^b(X)\big|^p = e^{-p\int_0^{t_0}(\sigma^{-2}D_mF(m_s,X_s)+\nabla U(X_s))\,dX_s-\frac{p}{2}\int_0^{t_0}\frac{|b_s|^2}{\sigma^2}\,ds}.$$

Since D_mF is bounded, in order to prove (A.4), it is enough to show that

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

By Itô formula, we obtain

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

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

Acknowledgements

The authors would like to thank the anonymous referees for their valuable comments which have helped improve the clarity of the paper.

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.

References

[1] L. Ambrosio, N. Gigli and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer, Berlin, 2008. MR2401600

[2] D. Bakry and M. Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84* 177–206, Springer, Berlin, 1985. MR0889476 https://doi.org/10.1007/BFb0075847

[3] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory* **39** (3) (1993) 930–945. MR1237720 https://doi.org/10.1109/18.256500

[4] M. Belkin, A. Rakhlin and A. B. Tsybakov. Does data interpolation contradict statistical optimality? 2018. Available at arXiv:1806.09471.

[5] V. Benes. Existence of optimal stochastic control laws. *SIAM J. Sci. Comput.* (1970) 446–472. MR0300726 https://doi.org/10.1137/0309034

[6] A. Bensoussan. *Stochastic Control of Partially Observable Systems*. Cambridge University Press, Cambridge, 1992. MR1191160 https://doi.org/10.1017/CBO9780511526503

[7] V. I. Bogachev, M. Röckner and S. V. Shaposhnikov. Convergence in variation of solutions of nonlinear Fokker–Planck–Kolmogorov equations to stationary measures. *Journal of Functional Analysis* (2019). MR3957996 https://doi.org/10.1016/j.jfa.2019.03.014

[8] F. Bolley, I. Gentil and A. Guillin. Uniform convergence to equilibrium for granular media. *Arch. Ration. Mech. Anal.* **208** (2) (2013) 429–445. MR3035983 https://doi.org/10.1007/s00205-012-0599-z

[9] M. Bossy, B. Jourdain et al. Rate of convergeance of a particle method for the solution of a 1d viscous scalar conservation law in a bounded interval. *Ann. Probab.* **30** (4) (2002) 1797–1832. MR1944006 https://doi.org/10.1214/aop/1039548372

[10] M. Bossy and D. Talay. A stochastic particle method for the McKean–Vlasov and the Burgers equation. *Math. Comp.* **66** (217) (1997) 157–192. MR1370849 https://doi.org/10.1090/S0025-5718-97-00776-X

[11] O. Butkovsky. On ergodic properties of nonlinear Markov chains and stochastic McKean–Vlasov equations. *Theory Probab. Appl.* **58** (4) (2014) 661–674. MR3403022 https://doi.org/10.1137/S0040585X97986825

[12] P. Cardaliaguet, F. Delarue, J. M. Lasry and P. L. Lions. The master equation and the convergence problem in mean field games, 2015. Available at arXiv:1509.02505. MR3616319

[13] R. Carmona and F. Delarue. *Probabilistic Theory of Mean Field Games with Applications II*. Springer, Berlin, 2018. MR3753660

[14] J. A. Carrillo, R. J. McCann and C. Villani. Kinetic equilibration rates for granular media and related equations: Entropy dissipation and mass transportation estimates. *Rev. Mat. Iberoam.* **19** (3) (2003) 971–1018. MR2053570 https://doi.org/10.4171/RMI/376

[15] P. Cattiaux, A. Guillin and F. Malrieu. Probabilistic approach for granular media equations in the non-uniformly convex case. *Probab. Theory Related Fields* **140** (1–2) (2008) 19–40. MR2357669 https://doi.org/10.1007/s00440-007-0056-3

[16] J. F. Chassagneux, D. Crisan and F. Delarue. A probabilistic approach to classical solutions of the master equation for large population equilibria, 2014. Available at arXiv:1411.3009. MR3332710 https://doi.org/10.1007/978-3-319-11292-3_4

[17] J. F. Chassagneux, L. Szpruch and A. Tse. Weak quantitative propagation of chaos via differential calculus on the space of measures, 2019. Available at arXiv:1901.02556.

[18] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems* 3040–3050, 2018.

[19] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** (4) (1989) 303–314. MR1015670 https://doi.org/10.1007/BF02551274

[20] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley, New York, 1997. MR1431744 https://doi.org/10.1002/9781118165904

[21] A. Durmus, A. Eberle, A. Guillin and R. Zimmer. An elementary approach to uniform in time propagation of chaos, 2018. Available at arX:1805.11387. MR4163850 https://doi.org/10.1090/proc/14612

[22] A. Eberle, A. Guillin and R. Zimmer. Quantitative Harris-type theorems for diffusions and McKean–Vlasov processes. *Trans. Amer. Math. Soc.* **371** (10) (2019) 7135–7173. MR3939573 https://doi.org/10.1090/tran/7576

[23] H. Föllmer. Time reversal on Wiener space. In *Stochastic Processes—Mathematics and Physics* 119–129. S. A. Albeverio, P. Blanchard and L. Streit (Eds). Springer, Berlin, 1986. MR0838561 https://doi.org/10.1007/BFb0080212

[24] T. Hastie, A. Montanari, S. Rosset and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation, 2019. Available at arXiv:1903.08560.

[25] U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *Ann. Probab.* **14** (4) (1986) 1188–1205. MR0866342

[26] D. Henry. *Geometric Theory of Semilinear Parabolic Equations*. Springer, Berlin, 1981. MR0610244

[27] R. Holley and D. Stroock. Simulated annealing via Sobolev inequalities. *Comm. Math. Phys.* **115** (4) (1988) 553–569. MR0933455

[28] R. A. Holley, S. Kusuoka and D. W. Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *J. Funct. Anal.* **83** (2) (1989) 333–347. MR0995752 https://doi.org/10.1016/0022-1236(89)90023-2

[29] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4** (2) (1991) 251–257.

[30] C. R. Hwang et al. Laplace's method revisited: Weak convergence of probability measures. *Ann. Probab.* **8** (6) (1980) 1177–1182. MR0602391

[31] A. Javanmard, M. Mondelli and A. Montanari. Analysis of a two-layer neural network via displacement convexity, 2019. Available at arXiv:1901.01375. MR4185822 https://doi.org/10.1214/20-AOS1945

[32] R. Jordan and D. Kinderlehrer. An extended variational principle. In *Partial Differential Equations and Applications: Collected Papers in Honor of Carlo Pucci*. CRC, 1996. MR1371591 https://doi.org/10.5006/1.3292113

[33] R. Jordan, D. Kinderlehrer and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.* **29** (1) (1998) 1–17. MR1617171 https://doi.org/10.1137/S0036141096303359

[34] I. Karatzas, W. Schachermayer and B. Tschiderer, Pathwise Otto calculus, 2019. Available at arXiv:1811.08686.

[35] O. A. Ladyzenskaja, V. A. Solonnikov and N. N. Ural'ceva. *Linear and Quasi-Linear Equations of Parabolic Type*. *Translations of Mathematical Monographs*. AMS, Providence, 1968.

[36] Y. LeCun, Y. Bengio and G. Hinton. Deep learning. *Nature* **521** (7553) (2015) 436.

[37] F. Malrieu et al. Convergence to equilibrium for granular media equations and their Euler schemes. *Ann. Appl. Probab.* **13** (2) (2003) 540–560. MR1970276 https://doi.org/10.1214/aoap/1050689593

[38] S. Mei, T. Misiakiewicz and A. Montanari. Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit, 2019. Available at arXiv:1902.06015.

[39] S. Mei, A. Montanari and P. M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA* **115** (33) (2018) E7665–E7671. MR3845070 https://doi.org/10.1073/pnas.1806579115

[40] S. Mischler and C. Mouhot. Kac's program in kinetic theory. *Invent. Math.* **193** (1) (2013) 1–147. MR3069113 https://doi.org/10.1007/s00222-012-0422-3

[41] D. Nualart. *The Malliavin Calculus and Related Topics*. Springer, Berlin, 2006. MR2200233

[42] F. Otto. *The Geometry of Dissipative Evolution Equations: The Porous Medium Equation*, 2001. MR1842429 https://doi.org/10.1081/PDE-100002243

[43] F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.* **173** (2000) 361–400. MR1760620 https://doi.org/10.1006/jfan.1999.3557

[44] G. M. Rotskoff and E. Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error, 2018. Available at arXiv:1805.00915.

[45] J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks, 2018. Available at arXiv:1805.01053. MR4074020 https://doi.org/10.1137/18M1192184

[46] A. S. Sznitman. *Topics in Propagation of Chaos*. Springer, Berlin, 1991. MR1108185 https://doi.org/10.1007/BFb0085169

[47] L. Szpruch, S. Tan and A. Tse. Iterative particle approximation for McKean–Vlasov sdes with application to multilevel Monte Carlo estimation. *Ann. Appl. Probab.* **29** (2019) 2230–2265. MR3983338 https://doi.org/10.1214/18-AAP1452

[48] L. Szpruch and A. Tse. Antithetic multilevel particle system sampling method for McKean–Vlasov SDEs, 2019. Available at arXiv:1903.07063.

[49] J. Tugaut. Convergence to the equilibria for self-stabilizing processes in double-well landscape. *Ann. Probab.* **41** (3A) (2013) 1427–1460. MR3098681 https://doi.org/10.1214/12-AOP749

[50] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, Berlin, 2013. MR1367965 https://doi.org/10.1007/978-1-4757-2440-0

[51] A. Y. Veretennikov. On ergodic measures for McKean–Vlasov stochastic equations. In *Monte Carlo and Quasi-Monte Carlo Methods 2004* 471–486. Springer, Berlin, 2006. MR2208726 https://doi.org/10.1007/3-540-31186-6_29

[52] C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals. Understanding deep learning requires rethinking generalization, 2016. Available at arXiv:1611.03530.