THE CONTINUOUS FORMULATION OF SHALLOW NEURAL NETWORKS AS WASSERSTEIN-TYPE GRADIENT FLOWS

XAVIER FERNÁNDEZ-REAL AND ALESSIO FIGALLI

Abstract. It has been recently observed that the training of a single hidden layer artificial neural network can be reinterpreted as a Wasserstein gradient flow for the weights for the error functional. In the limit, as the number of parameters tends to infinity, this gives rise to a family of parabolic equations. This survey aims to discuss this relation, focusing on the associated theoretical aspects appealing to the mathematical community and providing a list of interesting open problems.

1. Introduction

The extensive and successful use of machine learning in recent years has been remarkable. However, from a mathematical viewpoint, an adequate theoretical understanding of its primary governing principles is still missing in many situations. Often, each problem needs to be studied individually, even within the application of the same technique, to obtain the desired visible result.

Recently, a new continuous viewpoint of artificial neural networks has risen, intending to shine some light on this computing system's understanding. This theory has already been developed and shown important results and, roughly speaking, consists in viewing the gradient descent used to optimize parameters in a neural network as a gradient flow in the Wasserstein distance for their own empirical measure.

More precisely, training neural networks can be thought of as discretizations of a gradient flow with the appropriate metric and functional. This observation has opened the door to studying (at a theoretical level) the general convergence properties of such methods deducing properties of the corresponding continuous limit. Most of this study has been conducted from a numerical point of view, and there are still many open questions that are also interesting from a purely theoretical perspective.

2020 *Mathematics Subject Classification.* 35Q49, 49Q22, 68T07.

Key words and phrases. Machine learning, continuous formulation, gradient flow, Wasserstein distance.

This work has received funding from the European Research Council (ERC) under the Grant Agreement No 721675. In addition, X. F. was supported by the SNF grant 200021_182565.

In this framework, new mathematical problems and PDE systems have arisen, which have not yet been fully adopted by the mathematical community. This short survey aims to bridge this gap to present this fascinating problem in the gradient flows community's language.

We refer the interested reader to [CB18, MMN18, JMM19, EMW19, SS20] and references therein for an in-depth introduction to the topic, and also to [E17, EHL19] for an approach more focused on dynamical systems and optimal control problems.

2. Shallow neural network and gradient flows

Given a domain $D \subset \mathbb{R}^n$, and a function $f : D \to \mathbb{R}$, training a single hidden layer artificial neural network (or shallow neural network) consists in approximating f with expressions of the form

$$f_N(x) = f_N(x, w_1, \dots, w_N, \theta_1, \dots, \theta_N) = \frac{1}{N} \sum_{i=1}^{N} w_i h(\theta_i, x), \tag{2.1}$$

where $w_i \in \mathbb{R}$ and $\theta_i \in \Theta \subset \mathbb{R}^d$ are parameters to be optimized (usually taken in pairs (w_i, θ_i)), and $h : \Theta \times D \to \mathbb{R}$ is called the activation function, which is nonlinear. Such construction of approximating functions is often graphically represented as seen in Figure 2.1, and when the number of layer increases, the number of interconnections between the neurons increases as well, very loosely resembling a biological neural network.

In applications, it is usual to assume that

$$d = n + 1 \qquad \text{and} \qquad h(\theta, x) = \sigma(\theta' \cdot x + \theta^{(d)}), \tag{2.2}$$

where $\theta = (\theta', \theta^{(d)}) \in \mathbb{R}^n \times \mathbb{R}$, for a suitable nonlinearity σ.[1] Thus, neural networks try to approximate a given function with linear combinations of nonlinearities. However, for the sake of generality, here we will not consider a specific form of $h(\theta, x)$, and we focus instead on the general formulation where $h(\theta, x)$ can be arbitrary.

The number N of parameters $(w, \theta) \in \mathbb{R}^{d+1}$ used to in (2.1) corresponds to the number of neurons or hidden units. When training a neural network one tries to minimize the expected error, sometimes called *risk* or *generalization error*, obtained from approximating f by f_N. To do so, one needs to define a loss function ℓ, that we consider to be

$$\ell\big(f, f_N\big) = \frac{1}{2} \int_D |f(x) - f_N(x)|^2 dx.$$

[1] A typical nonlinearity is the sigmoid function. Namely, if we denote $\sigma(t) = \frac{1}{1+e^{-t}}$, we consider $h(\theta, x) = \sigma(\theta \cdot x)$. However, nowadays, the most frequently used activation function in applications is not smooth nor bounded: the ReLU function $\sigma(t) = \max\{t, 0\}$.

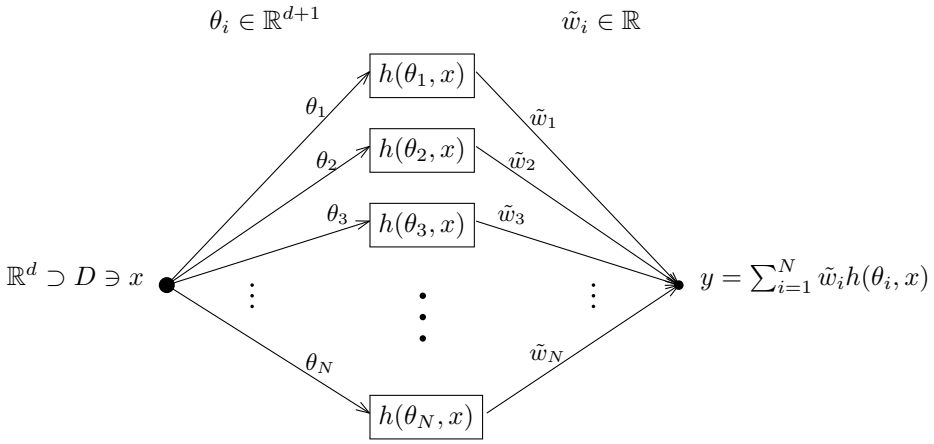

FIGURE 2.1. Graphic representation of the approximating functions given by what is known as a single hidden layer *artificial neural network*. The variables are $\tilde{w}_i = \frac{1}{N} w_i$ according to the notation in (2.1).

Let us denote by $\mathcal{H}_N$ the class of f_N that can be obtained as (2.1). Then, one wants to solve the minimization problem

$$\min_{f_N \in \mathcal{H}_N} \ell\big(f, f_N\big), \tag{2.3}$$

where ℓ is as above. The standard approach nowadays is to start from some choice of weights $\bar{w} = (\bar{w}_1, \dots, \bar{w}_N)$ and $\bar{\theta} = (\bar{\theta}_1, \dots, \bar{\theta}_N)$, and perform gradient descent on these parameters (w, θ) in order to (possibly) achieve the minimizer to (2.3)[2]:

$$\begin{cases} \frac{d}{dt}(w(t), \theta(t)) = -N \nabla_{w,\theta} \ell\big(f, f_N(\cdot, w(t), \theta(t))\big), \\ (w(t), \theta(t)) = (\bar{w}, \bar{\theta}), \end{cases} \tag{2.4}$$

with $w(t) = (w_1(t), \dots, w_N(t))$ and $\theta(t) = (\theta_1(t), \dots, \theta_N(t))$.[3] Unfortunately, given the structure of the approximating functions (2.1), this problem is non-convex, and thus one does not expect to arrive to the minimizer in general.

Because of this degeneracy, a recent approach has been to consider a continuum model where one lets the number N of neurons go to infinity. The general hope is that this limit problem can be studied with PDE techniques, and then one may try to extract informations also on the original problem (with N fixed) provided N is sufficiently large. This latter step has been studied, for

[2]In fact, in reality, one uses stochastic gradient descent, by considering random samples $(x_i, f(x_i))$ of our data or training set.

[3]Actually, to avoid over-fitting, it is usual to add to the loss function ℓ a convex potential on the parameters, see (2.7) or (2.13).

instance, in [CB18], although many questions are still open (see Section 5 for more details).

In this note we shall not discuss the consistency of the approximation as $N \to \infty$, but we instead focus on the analysis of the continuum interpretation. As we shall see, there is more than one way to interpret the limit as $N \to \infty$, and more than one possible formulation exists. In the next sections we first present the continuum energy functionals that one can obtain by taking the limit of $\ell\big(f, f_N\big)$ as $N \to \infty$, and then we shall analyze the possible gradient flows that can arise from this model.

2.1. **The μ formulation.** We start with the most commonly used interpretation of a neural network, when the number of neurons is allowed to go to infinity. In this case, we want to treat the two variables w and θ in the same way. For that, let us slightly reformulate the previous problem.

Set $\xi := (w, \theta) \in \mathbb{R} \times \Theta$, $\Omega := \mathbb{R} \times \Theta \subset \mathbb{R}^{d+1}$, and let us define $\Phi(\xi, x) := w\, h(\theta, x)$, so that we can deal with both parameters simultaneously. Thus, (2.1) can be written as

$$f_N(x) = f_N(x, \xi_1, \dots, \xi_N) = \frac{1}{N} \sum_{i=1}^{N} \Phi(\xi_i, x).$$

Let μ_N denote the empirical distribution of $\{\xi_i\}_{1 \le i \le N}$, namely,

$$\mu_N(\xi) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i}(\xi).$$

Then the function f_N can be expressed in terms of μ_N as

$$f_N(x) = \int_\Omega \Phi(\xi, x) \mu_N(d\xi),$$

and the gradient descent (2.4) can be rewritten only in terms of the empirical measure at time t, that is, $\mu_N(t) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i(t)}(\xi)$ with $\xi_i(t) = (w_i(t), \theta_i(t))$.

Letting $N \to \infty$, the space of empirical measures can approximate any probability measure $\mu \in \mathscr{P}(\Omega)$. Hence, this suggests the study of approximating functions defined as

$$f_\mu(x) := \int_\Omega \Phi(\xi, x) \mu(d\xi) \qquad \forall\, \mu \in \mathscr{P}(\Omega). \tag{2.5}$$

Then, our minimization problem consists in minimizing

$$F(\mu) := \frac{1}{2} \int_D (f - f_\mu)^2 dx$$

among probability measures $\mu \in \mathscr{P}(\Omega)$. That is,

$$\min_{\mu \in \mathscr{P}(\Omega)} F(\mu) = \min_{\mu \in \mathscr{P}(\Omega)} \frac{1}{2} \int_D \left(f - \int_\Omega \Phi(\xi, x) \mu(d\xi) \right)^2 dx. \tag{2.6}$$

In other words, we are looking at the best way of approximating f in $L^2(D)$ using functions of the form (2.5).

Note that, for many choices of Φ, the set of functions of the form (2.5) may be dense in $L^2(D)$, so that the minimum may be zero (and we want to study ways to attain or approximate it). Moreover, oftentimes, to avoid overfitting in the training space, it is common to add a potential term used as a renormalization in the optimization of the neural networks. Therefore, the energy that we want to minimize over $\mu \in \mathscr{P}(\Omega)$ becomes

$$F(\mu) = \frac{1}{2}\int_D \left(f - \int_\Omega \Phi(\xi, x)\mu(d\xi)\right)^2 + \int_\Omega V(\xi)\mu(d\xi), \tag{2.7}$$

for some fixed function $V : \Omega \to \mathbb{R}$. A natural choice of V is given by the quadratic potential

$$V(\xi) = \frac{\lambda}{2}|\xi|^2, \qquad \text{with } \lambda > 0. \tag{2.8}$$

Notice that, with this additional term, the minimum of our functional will not be zero anymore.

We remark that, by considering probability measures instead of discrete parameters, we are not losing information. Indeed, if we restrict our problem to the set of atomic measures with N atoms, then we go back to formulation (2.1).

Remark 2.1. One might benefit from the convex structure of the functional F with respect to the classical linear structure of $\mathscr{P}(\Omega)$, namely,

$$F\left(\alpha\mu_1 + (1-\alpha)\mu_2\right) \le \alpha F(\mu_1) + (1-\alpha)F(\mu_2) \qquad \forall\, \alpha \in [0,1].$$

In particular, from here one can show that if μ_1 and μ_2 are two local minimizers then $\int_D \Phi(\xi, x)\mu_1(d\xi) = \int_D \Phi(\xi, x)\mu_2(d\xi)$ for all $x \in D$ and their potential energy is the same, i.e., $\int_\Omega V(\xi)\mu_1(d\xi) = \int_\Omega V(\xi)\mu_2(d\xi)$.[4] In particular, local minimizers are unique under Φ.

An advantage of the continuous formulation is that the invariance with respect to permutations of neurons is included in the model. Also, assuming that one already knows symmetries for the objective function (for example,

[4]Indeed, suppose that μ_1 and μ_2 are two local minimizers, and for $\alpha \in [0,1]$ consider $\mu_\alpha := (1-\alpha)\mu_0 + \alpha\mu_1$. Then, we can compute $\frac{d}{d\alpha}F(\mu_\alpha)$, which equals

$$\frac{d}{d\alpha}F(\mu_\alpha) = \alpha\int_D |f-f_1|^2 - (1-\alpha)\int_D |f-f_0|^2 + (1-2\alpha)\int_D (f-f_0)(f-f_1) - \int V(\mu_0-\mu_1).$$

Since μ_0 and μ_1 are local minimizers we have $\frac{d}{d\alpha}\big|_{\alpha=0}F(\mu_\alpha) \ge 0$ and $\frac{d}{d\alpha}\big|_{\alpha=1}F(\mu_\alpha) \le 0$, and therefore

$$0 \ge \frac{d}{d\alpha}\big|_{\alpha=1}F(\mu_\alpha) - \frac{d}{d\alpha}\big|_{\alpha=0}F(\mu_\alpha) = \int_D |f_0 - f_1|^2$$

thus $f_0 = f_1$. This implies that $\frac{d}{d\alpha}F(\mu_\alpha) = -\int V(\mu_0-\mu_1)$, so it follows from $\frac{d}{d\alpha}\big|_{\alpha=0}F(\mu_\alpha) \ge 0$ and $\frac{d}{d\alpha}\big|_{\alpha=1}F(\mu_\alpha) \le 0$ that $\int V\mu_0 = \int V\mu_1$.

rotational symmetry to identify certain images), they can be incorporated directly into the minimization problem, much more easily than in the discrete case.

2.2. **Comparison between the continuous and discrete model.** At the discrete level, adding a potential term corresponds to considering the minimization of the loss functional

$$F_N(f_N) = \frac{1}{2}\int_D |f(x) - f_N(x)|^2\,dx + \sum_{i=1}^N V(\xi_i), \tag{2.9}$$

for some convex function V. Equivalently, we are considering the discrete minimization problem

$$\min_{\xi_i\in\Omega} F_N(f_N) \qquad \text{for} \qquad f_N(x) = \frac{1}{N}\sum_{i=1}^N \Phi(\xi_i, x). \tag{2.10}$$

We have seen that this minimization problem can be interpreted as a particular case of the more general problem for probability measures. Namely, if we consider F given by (2.7), then Problem (2.10) generalizes to

$$\min_{\mu\in\mathscr{P}(\Omega)} F(\mu). \tag{2.11}$$

Notice also that, while Problem (2.10) is heavily non-convex, Problem (2.11) has a convex structure (see Remark 2.1).

2.2.1. *Consistency.* The consistency between Problems (2.10) and (2.11) has generated some research in the recent years. These are some results:

(i) If μ_N is the empirical distribution of a minimizer of F_N, and μ is a minimizer of F, then $F_N(\mu_N) = F(\mu) + O(N^{-1})$. In addition, if V is coercive, then μ_N converges weakly* to a minimizer μ of F (up to subsequences).

(ii) The Wasserstein gradient flow of F with initialization μ_N is the same as the corresponding gradient descent of the discretized problem, cf. (2.4) (see [CB18]).

(iii) As shown in [MMN18] (see also [RV18]), the stochastic gradient descent for (2.10) (cf. (2.4)) converges to the gradient flow of (2.11) with its own initialization. More precisely, if one denotes by $\mu_N^{(k)}$ the empirical distribution of the parameters $(\xi_i^k)_{1\le i\le N}$ in the stochastic gradient descent for F_N at step k, then one can prove quantitative convergence of $\mu_N^{(t/\varepsilon)}$ to μ_t as $N\to\infty$ and $\varepsilon\downarrow 0$, where μ_t is the gradient flow in the Wasserstein metric for the functional F.

(iv) In [CB18] the authors proved that if one approximates an initial measure μ_0 by N atoms, the corresponding gradient descents converge, under some conditions on the initial measure, to the gradient flow for F with initial measure μ_0, also as $t\to\infty$. Thus, they showed that

one does actually benefit from the convex structure in (2.11): given a nice enough initial measure (initial configuration of weights, with enough neurons), its gradient descent will converge to a configuration of parameters very close to a minimizer for F. This is currently a non-quantitative result that checks the consistency of the problem posed.

All these facts show that studying the minimization problem (2.11) could be very useful in trying to derive properties for the discrete problem (2.10).

On the other hand one should keep in mind that, in general, the Wasserstein gradient flow of (2.6) may not converge to a global minimizer, but just to a stationary point. For example, given an initial configuration with a fixed number of deltas, the corresponding gradient flow never increases the amount of deltas, and thus it converges to some measure that has at most the same number of deltas as the initial configuration. In particular, this limit will not generally be a global minimizer of F. Still, the result in [CB18] says that such limiting configuration will approximate a minimizer, under suitable conditions.

2.3. **The (ρ, H) formulation.** An alternative approach to the previous generalization (what we called "the μ formulation") consists in taking advantage of the structure of Φ, where the weights w and positions θ have asymmetric roles. One can think of this approach as a charged particles system, where we can discretize in θ (positions of the particle) assigning a coefficient w to each atomic measure of the discretization (charge of the particle). We refer to some examples in [EMW19].

While these continuous methods a priori do not necessarily arise from a discrete gradient descent, they yield other evolution equations whose discretization could benefit from additional properties. As we will see, some of these associated PDE systems also dissipate energy, suggesting that alternative gradient flow formulations are possible and interesting.

Recall that we have $\theta \in \Theta \subset \mathbb{R}^d$ and $w \in \mathbb{R}$. Consider the measure in θ given by

$$\rho_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} w_i \delta_{\theta_i}(\theta)$$

(observe that now ρ_N is not necessarily a probability measure, and not even a positive measure, since the weights w_i may be negative). Then the function f_N in (2.1) can be expressed as

$$f_N(x) = \int_\Theta h(\theta, x) \rho_N(d\theta).$$

This suggests considering functions of the form

$$f_m(x) = \int_\Theta h(\theta, x) m(d\theta),$$

where now $m \in \mathcal{M}$ is a finite (signed) measure.

Notice that this is related to what we were doing before with probability measures μ defined on $\Omega = \mathbb{R} \times \Theta$. Indeed, to see the relation between the two formulations, given $\mu \in \mathcal{P}(\Omega)$ consider its disintegration with respect to θ. Namely, one can write

$$\mu(d\xi) = \nu_\theta(dw) \otimes \rho(d\theta),$$

where ρ and ν are formally defined as[5]

$$\rho(\theta) = \int_{\mathbb{R}} \mu(dw, \theta), \qquad \nu_\theta(w) = \frac{1}{\rho(\theta)} \mu(w, \theta).$$

Then, given $\Phi(\xi, x) = w\, h(\theta, x)$, we have

$$\int_\Omega \Phi(\xi, x) \mu(d\xi) = \int_\Theta \left(\int_{\mathbb{R}} w\, \nu_\theta(dw) \right) h(\theta, x) \rho(d\theta) = \int_\Theta h(\theta, x) m(d\theta)$$

where

$$m(d\theta) = \left(\int_{\mathbb{R}} w\, \nu_\theta(dw) \right) \rho(d\theta).$$

In other words, (2.6) is equivalent to the minimization problem

$$\min_{m \in \mathcal{M}(\Theta)} \frac{1}{2} \int_D \left(f - \int_\Theta h(\theta, x) m(d\theta) \right)^2 dx,$$

where $\mathcal{M}(\Theta)$ denotes the set of (finite) signed measures on Θ. Equivalently, if we define

$$H(\theta) = \int_{\mathbb{R}} w\, \nu_\theta(dw),$$

then our problem consists in finding the best approximation of f in $L^2(D)$ with functions of the form

$$f_{\rho, H}(x) = \int_\Theta H(\theta) h(\theta, x) \rho(d\theta). \tag{2.12}$$

In addition, keeping the same notation, and assuming to introduce a potential term of the form $V(\xi) = \frac{\lambda}{2}|\xi|^2$ in the μ formulation, then by Jensen's inequality we have

$$\int_\Omega |\xi|^2 \mu(d\xi) = \int_\Theta \int_{\mathbb{R}} w^2 \nu_\theta(dw) \rho(\theta) + \int_\Theta |\theta|^2 \rho(d\theta) \geq \int_\Theta \left(H(\theta)^2 + |\theta|^2 \right) \rho(d\theta).$$

In particular, if we were assuming $\int |\xi|^2 \mu(d\xi) < +\infty$ (i.e., μ has bounded second moments) in the previous formulation, then it is natural to assume ρ to have bounded second moments as well, and $H \in L^2(\Theta, \rho)$[6].

[5]This definition of the disintegration is correct if μ is absolutely continuous, and therefore can be identified as a function. Otherwise, the existence and uniqueness of such representation is provided by the disintegration theorem (see for instance [FG20, Theorem 1.4.10 and Appendix B]).

[6]Similarly, if our potential term was given by the p-moments instead, i.e., $V(\xi) = \lambda|\xi|^p$ for some $p \geq 1$ and $\lambda > 0$, then it would be natural to assume $H \in L^p(\Theta, \rho)$.

Notice that the expression (2.12) is similar to (2.5), the one appearing in the μ formulation. There are however, two main differences: on the one hand, the number of parameters has been reduced (from $\xi = (w, \theta)$ to simply θ); on the other hand, we are optimizing not only over probability measures ρ but also over functions $H \in L^2(\Theta, \rho)$. Thus, by looking at the explicit expression $\Phi(\xi, x)$ had in the previous formulation, we are trading off the amount of parameters of our problem with a new variable to optimize. We can do so because, in reality, the freedom given to the measure $\nu_\theta(dw)$ was limited: since $\Phi(\xi, x) = w\, h(\theta, x)$, we only see it through its first moment. In particular, given $\mu = \nu_\theta(dw) \otimes \rho(d\theta)$, one can replace it with $\delta_{H(\theta)}(dw) \otimes \rho(dw)$ and the problem remains the same. This is the idea behind what we call "the (ρ, H) formulation".

In conclusion, in the (ρ, H) formulation we are considering the functional

$$G(\rho, H) = \frac{1}{2} \int_D \left(f - \int_\Theta H(\theta) h(\theta, x) \rho(d\theta) \right)^2 dx + \int_\Theta \bar{V}(H, \theta) \rho(d\theta), \quad (2.13)$$

where now we have removed the dependence on the variable w, and we have added a potential term $\bar{V} : \mathbb{R} \times \Theta \to \mathbb{R}$. Note that, in this case, the L^2 regularization induced by (2.8) corresponds to $\bar{V}(H, \theta) = \frac{\lambda}{2}(H^2 + |\theta|^2)$.

3. PDE Formulations

In this section we first compute the Wasserstein gradient flow in the μ formulation (see Section 3.1). Then we discuss some evolution equations in the (ρ, H) formulation, as introduced in [EMW19] (Section 3.2). Finally, in Section 3.3, we present a new original approach to the problem of defining a gradient flow (ρ, H) formulation, based on propagation of chaos.

3.1. Gradient flow in the μ formulation.

Recall that $\Omega = \mathbb{R} \times \Theta \subset \mathbb{R}^{d+1}$, and $\xi = (w, \theta) \in \Omega$ denotes the parameters in this setting. Let $D \subset \mathbb{R}^n$, and let $h(\theta, x) : \Theta \times D \to \mathbb{R}$ be a given function, and let $\Phi(\xi, x) = w\, h(\theta, x)$.

We consider the minimization problem

$$\min_{\mu \in \mathscr{P}(\Omega)} F(\mu), \quad (3.1)$$

where

$$F(\mu) = \frac{1}{2} \int_D \left(\int_\Omega \Phi(\xi, x) \mu(d\xi) - f(x) \right)^2 dx + \int_\Omega V(\xi)\, \mu(d\xi). \quad (3.2)$$

Note that this expression can be rewritten as

$$F(\mu) = \bar{F} + \int_{\Omega \times \Omega} K(\xi, \bar{\xi}) \mu(d\xi) \mu(d\bar{\xi}) + \int_\Omega \mathcal{S}(\xi) \mu(d\xi) + \int_\Omega V(\xi) \mu(d\xi), \quad (3.3)$$

where

$$K(\xi,\bar\xi)=\frac12\int_D \Phi(\xi,x)\Phi(\bar\xi,x)dx, \qquad \mathcal{S}(\xi)=-\int_D \Phi(\xi,x)f(x)\,dx, \tag{3.4}$$

and $\bar F=\frac12\|f\|^2_{L^2(D)}$ is a constant. We remark that the smoothness of $\Phi(\xi,x)$ is related to the smoothness of $\mathcal{S}$ (in particular, if Φ is smooth, then $\mathcal{S}$ is a smooth).

The first variation of F with respect to μ at fixed measure $\mu_*\in\mathscr{P}(\Omega)$ is given by[7]

$$\begin{aligned}\frac{\delta F}{\delta\mu}(\mu_*)&=\int_D \Phi(\cdot,x)\left[\int_\Omega \Phi(\bar\xi,x)\,d\mu_*(\bar\xi)-f(x)\right]dx+V\\&=2\int_\Omega K(\cdot,\bar\xi)\mu_*(d\bar\xi)+\mathcal{S}+V,\end{aligned} \tag{3.5}$$

so that the Wasserstein subdifferential on the support of μ_* is

$$\begin{aligned}\nabla\frac{\delta F}{\delta\mu}(\mu_*)&=\int_D \nabla_\xi\Phi(\cdot,x)\left[\int_\Omega \Phi(\bar\xi,x)\,d\mu_*(\bar\xi)-f(x)\right]dx+\nabla V\\&=2\int_\Omega \nabla_\xi K(\cdot,\bar\xi)\mu_*(d\bar\xi)+\nabla\mathcal{S}+\nabla V\end{aligned}$$

(see for instance [AGS08, Chapter 10] or [FG20, Chapter 4.2]). Also, the Wasserstein gradient flow of F is by definition (see [FG20, Chapter 4.2]

$$\partial_t\mu_t=\operatorname{div}\left(\mu_t\nabla\frac{\delta F}{\delta\mu}(\mu_t)\right), \tag{3.6}$$

therefore the formulas above give us the following PDE:

$$\boxed{\partial_t\mu_t=\operatorname{div}\left(\mu_t\nabla\mathcal{L}(\mu_t)\right)+\operatorname{div}\left(\mu_t\nabla\mathcal{S}\right)+\operatorname{div}(\mu_t\nabla V),} \tag{3.7}$$

with

$$\mathcal{L}(\mu_t)(\xi)=2\int_\Omega K(\xi,\bar\xi)\mu_t(d\bar\xi). \tag{3.8}$$

Notice that $\mathcal{L}$ is an integral operator that is positive semi-definite.[8] Also, it can be checked by a direct computation that a solution $\mu_t(\xi)$ of the PDE satisfies

[7]By definition, $\frac{\delta F}{\delta\mu}(\mu_*)$ is defined as the unique element such that

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}F(\mu_*+\varepsilon\varphi)=\int_\Omega\frac{\delta F}{\delta\mu}(\mu_*)\,\varphi\,d\xi \qquad \forall\,\varphi\in C_c^\infty(\Omega).$$

[8]Indeed, it follows by (3.4) that

$$\int_\Omega\mathcal{L}(\mu)(\xi)\,\mu(d\xi)=\frac12\int_D\left(\int_\Omega\Phi(\xi,x)\mu(d\xi)\right)^2dx\ge0.$$

an energy dissipation from the gradient flow structure, that is, the energy is monotone non-increasing along trajectories:

$$\frac{d}{dt}F(\mu_t) = -\int_\Omega \left|\nabla \frac{\delta F}{\delta \mu}(\mu_t)\right|^2 \mu_t(d\xi). \tag{3.9}$$

In particular, stationary points corresponds to measures for which the derivative of the energy is zero. This motives the following:

Definition 3.1. We say that measure μ_* is a stationary point of our functional F (in the Wasserstein sense) if

$$\nabla \frac{\delta F}{\delta \mu}(\mu_*) = 0 \qquad \text{on} \quad \text{supp}(\mu_*). \tag{3.10}$$

Notice that, if we consider the natural potential term $V(\xi) = \frac{\lambda}{2}|\xi|^2$, then our PDE (3.7) becomes

$$\partial_t \mu_t = \operatorname{div}\left(\mu_t \nabla \mathcal{L}(\mu_t)\right) + \operatorname{div}\left(\mu_t \nabla \mathcal{S}\right) + \lambda \operatorname{div}(\mu_t\, \xi).$$

Let us conclude this subsection by observing that, in general, the previous PDEs are posed when the domain $\Omega = \mathbb{R}^{d+1}$. If, instead, one considers Θ a bounded smooth domain, an extra zero Neumann-boundary condition (so that the mass cannot escape) needs to be imposed:

$$\boxed{\boldsymbol{\nu} \cdot \nabla \left(\mathcal{L}(\mu_t) + \mathcal{S} + V\right)\, \mu_t = 0 \quad \text{on } \partial\Omega,} \tag{3.11}$$

where $\boldsymbol{\nu}$ denotes the unit outer normal vector to $\partial\Omega$.

3.2. **A first PDE approach in the (ρ, H) formulation.** As discussed before, an alternative approach is based on the (ρ, H) formulation described in subsection 2.3. So, it makes sense to design an appropriate evolution system of PDEs with good convergence properties, which could potentially lead to a nice particle method in the discrete case.

Let $\Theta \subset \mathbb{R}^d$ be the parameter space in this setting. Let $D \subset \mathbb{R}^n$ and let $h(\theta, x) : \Theta \times D \to \mathbb{R}$ be a fixed activation function.

We consider now the functional

$$G(\rho, H) = \frac{1}{2}\int_D \left(\int_\Theta H(\theta) h(\theta, x)\rho(d\theta) - f(x)\right)^2 dx + \int_\Theta \bar{V}(H, \theta)\rho(d\theta), \tag{3.12}$$

where, as before, $f \in L^2(D)$ is a given function.

As in (3.3), we can write

$$\begin{aligned} G(\rho, H) = \bar{G} &+ \int_{\Theta\times\Theta} \bar{K}(\theta, \bar{\theta}) H(\theta)\rho(d\theta) H(\bar{\theta})\rho(d\bar{\theta}) \\ &+ \int_\Theta \overline{\mathcal{S}}(\theta) H(\theta)\rho(d\theta) + \int_\Theta \bar{V}(H, \theta)\rho(d\theta), \end{aligned} \tag{3.13}$$

where

$$\bar{K}(\theta,\bar{\theta}) = \frac{1}{2}\int_D h(\theta,x)h(\bar{\theta},x)dx, \qquad \overline{\mathcal{S}}(\theta) = -\int_D h(\theta,x)f(x)\,dx, \tag{3.14}$$

and $\bar{G} = \frac{1}{2}\|f\|^2_{L^2(D)}$ is a constant. Notice that, also as before, the function $\overline{\mathcal{S}}$ is smooth if h is smooth with respect to θ.

We shall directly focus on the quadratic potential

$$\bar{V}(H,\theta) = \bar{V}_\lambda(H,\theta) := \frac{\lambda}{2}\left(H^2 + |\theta|^2\right), \tag{3.15}$$

so that, as discussed in Subsection 2.3, the natural space for H is given by $L^2(\Theta,\rho)$, and our minimization problem is given by

$$\min_{\substack{\rho\in\mathscr{P}(\Theta)\\ H\in L^2(\Theta,\rho)}} G(\rho,H). \tag{3.16}$$

We now want to obtain an evolution system of PDEs for (ρ,H) with nice properties. As we shall see, this can be performed in more than one way.

We first start with the evolution of $\rho\in\mathscr{P}(\Theta)$. As before, it makes sense to make it evolve according to the Wasserstein gradient flow of G. Namely, if we denote (ρ_t,H_t) our evolution variables, we have

$$\partial_t\rho_t = \operatorname{div}\left(\rho_t\nabla\frac{\delta G}{\delta\rho}(\rho_t,H_t)\right),$$

where the first variation density of G with respect to ρ at $(\rho_*,H_*)\in\mathscr{P}(\Theta)\times L^2(\Theta,\rho_*)$ is given by

$$\frac{\delta G}{\delta\rho}(\rho_*,H_*) = 2H_*(\cdot)\int_\Theta \bar{K}(\cdot,\bar{\theta})H_*(\bar{\theta})\rho_*(d\bar{\theta}) + \overline{\mathcal{S}}H_* + \bar{V}_\lambda(\theta,H_*),$$

so that

$$\begin{aligned}\nabla\frac{\delta G}{\delta\rho}(\rho_*,H_*) = 2\nabla\left[H_*(\cdot)\int_\Theta \bar{K}(\cdot,\bar{\theta})H_*(\bar{\theta})\rho(d\bar{\theta})\right]\\ + \nabla\left(\overline{\mathcal{S}}H_*\right) + \partial_H\bar{V}_\lambda(\cdot,H_*)\nabla H_* + (\nabla_\theta\bar{V}_\lambda)(\cdot,H_*).\end{aligned}$$

Thus, recalling (3.15), the evolution of ρ_t is given by

$$\boxed{\partial_t\rho_t = \operatorname{div}\left[\rho_t\nabla(H_t\overline{\mathcal{L}}(\rho_t,H_t))\right] + \operatorname{div}\left(\rho_t\nabla(\overline{\mathcal{S}}H_t)\right) + \lambda\operatorname{div}\left[\rho_t\left(H_t\nabla H_t+\theta\right)\right],} \tag{3.17}$$

where

$$\overline{\mathcal{L}}(\rho_t,H_t)(\theta) := 2\int_\Theta \bar{K}(\theta,\bar{\theta})H_t(\bar{\theta})\rho_t(d\bar{\theta}) \tag{3.18}$$

is a positive semi-definite integral operator (cf. (3.8)).

This gives the evolution of ρ_t, and one needs to couple it with an evolution for H_t. We shall present now two possible approaches.

3.2.1. *Separating variables.* The first way to obtain an evolution for the non-conserved variable H_t is to disregard partially the interaction between H and ρ: one performs the Wasserstein gradient flow of ρ on the one hand, and the $L^2(\Theta,\rho)$ gradient flow of H on the other (see [EMW19, Examples 1 and 2]).

Namely, one considers

$$\partial_t H_t = -\frac{\delta G}{\delta H}(\rho_t, H_t)$$

where, for a fixed ρ_t, $\frac{\delta G}{\delta H}(\rho_t,\cdot)$ denotes the variation of $G(\rho_t,\cdot)$ with respect to H in $L^2(\Theta,\rho_t)$.[9] This is

$$\frac{\delta G}{\delta H}(\rho_*, H_*) = 2\int_\Theta \bar K(\cdot,\bar\theta)H_*(\bar\theta)\rho(d\bar\theta) + \overline{\mathcal{S}} + \partial_H \bar V_\lambda(\cdot, H_*)$$

on $\operatorname{supp}\rho_*$, and therefore the evolution of (ρ_t, H_t) is given by

$$\boxed{\begin{array}{c}\partial_t\rho_t = \operatorname{div}\left[\rho_t\nabla(H_t\overline{\mathcal{L}}(\rho_t,H_t))\right] + \operatorname{div}\left(\rho_t\nabla(\overline{\mathcal{S}}H_t)\right) + \lambda\operatorname{div}\left[\rho_t\big(H_t\nabla H_t+\theta\big)\right]\\[2ex] \partial_t H_t = -\overline{\mathcal{L}}(\rho_t,H_t) - \overline{\mathcal{S}} - \lambda H_t\end{array}} \tag{3.19}$$

(this is coupled with a Neumann boundary condition for ρ_t, analogous to (3.11), whenever the domain Θ is not $\mathbb{R}^d$).

Note that this evolution has some difficulties, since one needs to make sure that all the terms appearing in the above PDE are well defined, at least in a weak sense. For instance, one needs to ensure that $\rho_t H_t \nabla H_t$ is well defined. Giving a meaning to this expression may be delicate if ρ is a singular measure. However, at least in the smooth case, this PDE makes sense. In addition, there is dissipation of the energy G along the path (ρ_t, H_t), namely

$$\frac{d}{dt}G(\rho_t,H_t) = -\int_\Theta \left|\nabla\frac{\delta G}{\delta\rho}(\rho_t,H_t)\right|^2\rho_t(d\theta) - \int_\Theta\left|\frac{\delta G}{\delta H}(\rho_t,H_t)\right|^2\rho_t(d\theta).$$

In particular, since $G(\rho_t,H_t)$ controls the $L^2(\Theta,\rho_t)$ norm of H_t, if one starts from a pair (ρ_0,H_0) with $H_0\in L^2(\Theta,\rho_0)$, then $H_t\in L^2(\Theta,\rho_t)$ (whenever the evolution is well-defined). Also, integrating the dissipation inequality above over any time interval implies that

$$\int_0^\infty\left[\int_\Theta\left|\nabla\frac{\delta G}{\delta\rho}(\rho_t,H_t)\right|^2\rho_t(d\theta) + \int_\Theta\left|\frac{\delta G}{\delta H}(\rho_t,H_t)\right|^2\rho_t(d\theta)\right]dt \le G(\rho_0,H_0),$$

[9] Namely, for a fixed $\rho_*\in\mathscr{P}(\Theta)$, $\frac{\delta G}{\delta H}(\rho_*,H_*)$ is the unique function in $L^2(\Theta,\rho_*)$ such that

$$\left\langle\frac{\delta G}{\delta H}(\rho_*,H_*),\varphi\right\rangle_{L^2(\Theta,\rho_*)} = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} G(H_*+\varepsilon\varphi)\qquad\forall\,\varphi\in L^2(\Theta,\rho_*).$$

which implies in particular that $\nabla \frac{\delta G}{\delta \rho}(\rho_t, H_t)$ and $\frac{\delta G}{\delta H}(\rho_t, H_t)$ belong to $L^2(\Theta, \rho_t)$ for a.e. t.

3.2.2. *Transporting along the flow of ρ_t.* Another way to describe the evolution of H_t is by incorporating the information that it is transported along the flow in the corresponding variable to be studied.

More precisely, note that the evolution of ρ_t in (3.19) can be written as a continuity equation (see [FG20, Eq. (4.6)]):

$$\partial_t \rho_t + \operatorname{div}(\rho_t \boldsymbol{v}_t) = 0, \qquad \text{where} \quad \boldsymbol{v}_t = -\nabla \frac{\delta G}{\delta \rho}(\rho_t, H_t).$$

Hence, if we define $X_t : \Theta \to \Theta$ as the flow of $\boldsymbol{v}_t$, namely

$$\begin{cases} \dot{X}_t &= \boldsymbol{v}_t \circ X_t \\ X_0 &= \mathrm{Id}, \end{cases} \tag{3.20}$$

then $\rho_t = (X_t)_{\#}\rho_0$, where $(X_t)_{\#}\rho_0$ denotes the push-forward measure of ρ_0 through the map X_t[10].

Thus, instead of considering simply H_t (which does not see the flow for ρ_t), an alternative option consists in rewriting the functional in terms of the variable $H_t \circ X_t$, which corresponds to transporting H_t along the flow of ρ_t. Hence, recalling that we are considering the potential $\bar{V}_\lambda$ from (3.15), one considers the evolution of $H_t \circ X_t$ given by

$$\partial_t (H_t \circ X_t) = -(\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}} - \lambda H_t) \circ X_t.$$

Noticing that $\partial_t (H_t \circ X_t) = [\partial_t H_t + \boldsymbol{v}_t \cdot \nabla H_t] \circ X_t$ (as a consequence of (3.20)), one obtains

$$\partial_t H_t + \boldsymbol{v}_t \cdot \nabla H_t = -\overline{\mathcal{L}}(\rho_t, H_t) - \overline{\mathcal{S}} - \lambda H_t.$$

Hence, the evolution system now becomes

$$\boxed{\begin{array}{c} \partial_t \rho_t + \operatorname{div}(\rho_t \boldsymbol{v}_t) = 0 \\ \\ \partial_t H_t + \boldsymbol{v}_t \cdot \nabla H_t = -\overline{\mathcal{L}}(\rho_t, H_t) - \overline{\mathcal{S}} - \lambda H_t, \end{array}} \tag{3.21}$$

with

$$\boxed{\boldsymbol{v}_t = -\nabla (H_t \overline{\mathcal{L}}(\rho_t, H_t) + H_t \overline{\mathcal{S}}) - \lambda H_t \nabla H_t - \lambda \theta,} \tag{3.22}$$

and again there is a zero Neumann boundary condition for ρ_t whenever Θ is not $\mathbb{R}^d$ (see (3.11)).

This corresponds the system introduced in [EMW19, Section 5.4] in the zero potential case ($\lambda = 0$), where they also design a particle method for this "modified gradient flow". This is definitely a very interesting model.

[10] That is,

$$\int_\Theta \varphi(\theta)[(X_t)_{\#}\rho_0](d\theta) = \int_\Theta \varphi(X_t(\theta))\rho_0(d\theta))$$

for any Borel function $\varphi : \Theta \to \mathbb{R}$.

However, since this system does not seem to dissipate energy in general, the mathematical analysis becomes more complicated.

3.3. **A gradient flow in the (ρ, H) formulation via propagation of chaos.** Let us give yet another possible evolution for (ρ_t, H_t) that produces a dissipative flow and does not rely on the smoothness of the measure. In this case, we do so by expressing the evolution in the μ formulation in the (ρ, H) variables, under a propagation of chaos assumption. As we shall explain below, the resulting system in this case is given by

$$\boxed{\begin{aligned} &\partial_t \rho_t + \operatorname{div}(\rho_t \boldsymbol{w}_t) = 0 \\ &\partial_t H_t + \boldsymbol{w}_t \cdot \nabla H_t = -\overline{\mathcal{L}}(\rho_t, H_t) - \overline{\mathcal{S}} - \partial_H \bar{V}(\cdot, H_t), \end{aligned}} \tag{3.23}$$

where now the vector $\boldsymbol{w}_t$ is (cf. (3.22))

$$\boxed{\boldsymbol{w}_t = -H_t \nabla(\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}) - \nabla_\theta \bar{V}(\cdot, H_t),} \tag{3.24}$$

and with zero Neumann boundary conditions

$$\boxed{\boldsymbol{\nu} \cdot \boldsymbol{w}_t \, \rho_t = 0 \quad \text{on } \partial\Theta} \tag{3.25}$$

whenever Θ is not $\mathbb{R}^d$. In particular, the evolution of H_t is still given by the corresponding evolution along the flow transporting ρ_t, but differently from before, ρ_t is not the standard Wasserstein flow. As proved below, this system has the main advantage that it dissipates energy, see Proposition 3.3.

In order to motivate the previous evolution system for the pair (ρ_t, H_t), we start by rewriting the PDE (3.7) as a hierarchy system in the (w, θ) variables. This is an infinite non-closed system of PDEs that depends on higher moments for the disintegration ν_θ and for the first derivatives of the potential.

Lemma 3.2. *Let $\Phi(\xi, x) = w\, h(\theta, x)$, and $\Omega = \mathbb{R} \times \Theta$ with Θ a smooth domain.*

Consider μ_t a (smooth and fast decaying) solution to (3.7)*, and define the disintegration into probability measures*

$$\mu_t(\xi) = \nu_{\theta,t}(w) \otimes \rho_t(\theta).$$

Define $H_{t,i}(\theta) := \int_{\mathbb{R}} w^i \nu_{\theta,t}(dw)$,

$$V^w_{t,i}(\theta) := \int_{\mathbb{R}} w^i \partial_w V(w, \theta) \nu_{\theta,t}(dw), \quad \boldsymbol{V}^\theta_{t,i}(\theta) := \int_{\mathbb{R}} w^i \nabla_\theta V(w, \theta) \nu_{\theta,t}(dw),$$

and consider

$$\overline{\mathcal{L}}(\rho_t, H_t)(\theta) := 2 \int_\Theta \bar{K}(\theta, \bar{\theta}) H_t(\bar{\theta}) \rho_t(d\bar{\theta}), \tag{3.26}$$

and $\bar K$ *and* $\overline{\mathcal{S}}$ *be given by* (3.14)*. Then, we have*

$$\begin{cases} \partial_t \rho_t = \operatorname{div}_\theta \left(\rho_t H_{t,1} \nabla_\theta \left[\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}} \right] \right) + \operatorname{div}_\theta(\rho_t \boldsymbol{V}^\theta_{t,0}) \\ \partial_t (H_{t,i} \rho_t) = \operatorname{div}_\theta \left(\rho_t H_{t,i+1} \nabla_\theta \left[\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}} \right] \right) + \operatorname{div}_\theta(\rho_t \boldsymbol{V}^\theta_{t,i}) \\ \qquad\qquad -i H_{t,i-1} \rho_t (\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}}) - i V^w_{t,i-1} \rho_t \qquad \forall\, i \geq 1, \end{cases} \tag{3.27}$$

with boundary conditions (whenever $\Theta \neq \mathbb{R}^d$*)*

$$\boldsymbol{\nu} \cdot \left\{ \rho_t H_{t,i} \nabla_\theta \left[\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}} \right] + \rho_t \boldsymbol{V}^\theta_{t,i-1} \right\} = 0 \quad \text{on } \partial\Theta. \tag{3.28}$$

Proof. Notice that

$$\mathcal{L}(\mu_t)(\xi) = w \overline{\mathcal{L}}(\rho_t, H_{t,1})(\theta), \qquad \mathcal{S} = w \overline{\mathcal{S}} \tag{3.29}$$

(recall (3.4), (3.8), (3.14), and (3.18)). Integrating (3.7) with respect to w and recalling (3.29), we obtain the first equation

$$\partial_t \rho_t = \operatorname{div}_\theta(\rho_t H_{t,1} \nabla_\theta \overline{\mathcal{L}}(\rho_t, H_{t,1})) + \operatorname{div}_\theta(\rho_t H_{t,1} \nabla_\theta \overline{\mathcal{S}}) + \operatorname{div}_\theta(\rho_t \boldsymbol{V}^\theta_{t,0})$$

using that

$$\int_{\mathbb{R}} w \mu_t(dw, \theta) = \rho_t H_{t,1}, \qquad \int_{\mathbb{R}} \nabla_\theta V(w, \theta) \mu_t(dw, \theta) = \rho_t \boldsymbol{V}^\theta_{t,0},$$

and that $\mu_t w$ has sufficient decay in w so that the terms in ∂_w in the divergence disappear when integrating by parts.

Similarly, given $i \geq 1$, we multiply (3.7) by w^i and then integrate with respect to w, to obtain

$$\begin{aligned} \partial_t (H_{t,i} \rho_t) = \operatorname{div}_\theta \left(\rho_t H_{t,i+1} \nabla_\theta \left[\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}} \right] \right) + \operatorname{div}_\theta(\rho_t \boldsymbol{V}^\theta_{t,i}) \\ + \int_{\mathbb{R}} w^i \partial_w \left[\left(\overline{\mathcal{L}}(\rho_t, H_{t,1}) + \overline{\mathcal{S}} + \partial_w V \right) \nu_{\theta,t}(dw) \otimes \rho_t \right]. \end{aligned}$$

Integrating by parts, we obtain the desired result.

The Neumann boundary conditions follow with the same procedure. □

As noticed above, the previous system (3.27) is not closed, as the i-th equation depends on H_{i+1}. Note however that the system could be closed if one knew that $\nu_{\theta,t}(w) = \delta_{H_t(\theta)}(w)$, since in that case

$$H_{t,i}(\theta) = H_t(\theta)^i = H^i_{t,1}(\theta) \qquad \forall\, i \geq 1.$$

This suggests a propagation of chaos assumption on the w variable in the previous expressions: by *assuming* that μ preserves being a delta in the w-variable (namely, $\nu_{\theta,t}(dw) = \delta_{H_t(\theta)}$ for some $H_t(\theta)$ for all $t \geq 0$), one gets a well-defined system of equations that now depends only on (ρ, H) and no longer sees the μ structure from before. In this case, if we denote $H_t = H_{t,1}$, we have that $H_{t,2} = H_t^2$, $V^w_{t,0} = \partial_w V(H_t, \theta)$, and $\boldsymbol{V}^\theta_{t,i} = H_t^i \nabla_\theta V(H_t, \theta)$. Also, since the equation for $H_{t,1}$ is already closed, one does not need to look at the other equations for $i \geq 2$.

Based on this discussion, our proposed new system is given by the following evolution equations:

$$\begin{cases} \partial_t \rho_t = \operatorname{div}\left(\rho_t H_t \nabla\left[\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}\right]\right) + \operatorname{div}(\rho_t \nabla_\theta \bar{V}(\theta, H_t)) \\ \partial_t (H_t \rho_t) = \operatorname{div}\left(\rho_t H_t^2 \nabla\left[\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}\right]\right) - \rho_t(\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}) \\ \qquad\qquad +\operatorname{div}(\rho_t H_t \nabla_\theta \bar{V}(\theta, H_t)) - \rho_t \partial_H \bar{V}(\theta, H_t), \end{cases} \tag{3.30}$$

where $\overline{\mathcal{L}}$ is given by (3.26), and it is combined with the zero Neumann boundary condition

$$\boldsymbol{\nu} \cdot \left\{\rho_t H_t \nabla\left[\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}\right] + \rho_t \nabla_\theta \bar{V}(\theta, H_t)\right\} = 0 \quad \text{on } \partial\Theta \tag{3.31}$$

whenever Θ is not $\mathbb{R}^d$ (cf. (3.27)-(3.28)).

Note that (3.30)-(3.31) is exactly our proposed model (3.23)-(3.25). In order to see that this new system is a reasonable candidate for the minimization of the energy (3.13), we prove now that the energy decreases along this evolution.

Proposition 3.3. *Let (ρ_t, H_t) solve* (3.30)-(3.31)*, and let G be given by* (3.13)*. Then*

$$\begin{aligned} \frac{d}{dt} G(\rho_t, H_t) = &- \int_\Theta \left| H_t \nabla(\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}) + \nabla_\theta \bar{V}(\theta, H_t)\right|^2 \rho_t \\ &- \int_\Theta \left(\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}} + \partial_H \bar{V}(\theta, H_t)\right)^2 \rho_t. \end{aligned}$$

In particular, the energy G is decreasing along (ρ_t, H_t).

Proof. We compute the derivative of $G(\rho_t, H_t)$ starting from (3.13). We have

$$\begin{aligned} \frac{d}{dt} G(\rho_t, H_t) &= \int_\Theta (\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}})\, \partial_t (H_t \rho_t) \\ &\quad + \int_\Theta \partial_H \bar{V}(\theta, H_t) \rho_t \partial_t H_t + \int_\Theta \bar{V}(\theta, H_t) \partial_t \rho_t \\ &= I + II + III, \end{aligned}$$

so that we can use (3.30) to substitute the time derivatives by the corresponding expressions. In particular, using that $\partial_t(H_t\rho_t) = H_t \partial_t \rho_t + \rho_t \partial_t H_t$ and (3.30), we deduce that

$$\begin{aligned} \rho_t \partial_t H_t = &\rho_t H_t \nabla H_t \cdot \nabla\left[\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}\right] - \rho_t(\overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}) \\ &+ \rho_t \nabla H_t \cdot \nabla_\theta \bar{V}(\theta, H_t) - \rho_t \partial_H \bar{V}(\theta.H_t). \end{aligned}$$

For the sake of readability, let us denote

$$\begin{aligned} \overline{\mathcal{N}}_t &:= \overline{\mathcal{L}}(\rho_t, H_t) + \overline{\mathcal{S}}, & \bar{V}_t &:= \bar{V}(\theta, H_t), \\ \partial_H \bar{V}_t &:= (\partial_H \bar{V})(\theta, H_t), & \nabla_\theta \bar{V}_t &:= (\nabla_\theta \bar{V})(\theta, H_t). \end{aligned}$$

Using these formulas, and integrating by parts using (3.31), we get

$$I = \int_\Theta \overline{\mathcal{N}}_t\, \partial_t(H_t\rho_t) = -\int_\Theta |\nabla\overline{\mathcal{N}}_t|^2 H_t^2\rho_t - \int_\Theta \overline{\mathcal{N}}_t^2\rho_t \\ - \int_\Theta \nabla\overline{\mathcal{N}}_t\cdot\nabla_\theta\bar V_t H_t\rho_t - \int_\Theta \partial_H\bar V_t\overline{\mathcal{N}}_t\rho_t,$$

$$II = \int_\Theta \partial_H\bar V_t\nabla H_t\cdot\nabla\overline{\mathcal{N}}_t H_t\rho_t - \int_\Theta \partial_H\bar V_t\overline{\mathcal{N}}_t\rho_t \\ + \int_\Theta \partial_H\bar V_t\nabla H_t\cdot\nabla_\theta\bar V_t\rho_t - \int_\Theta |\partial_H\bar V_t|^2\rho_t,$$

and

$$III = -\int_\Theta \nabla_\theta\bar V_t\cdot\nabla\overline{\mathcal{N}}_t H_t\rho_t - \int_\Theta \partial_H\bar V_t\nabla H_t\cdot\nabla\overline{\mathcal{N}}_t H_t\rho_t \\ - \int_\Theta |\nabla_\theta\bar V_t|^2\rho_t - \int_\Theta \partial_H\bar V_t\nabla H_t\cdot\nabla_\theta\bar V_t\rho_t.$$

Adding these identities, one finally gets

$$I+II+III = -\int_\Theta |\nabla\overline{\mathcal{N}}_t|^2 H_t^2\rho_t - \int_\Theta |\nabla_\theta\bar V_t|^2\rho_t - 2\int_\Theta \nabla\overline{\mathcal{N}}_t\cdot\nabla_\theta\bar V_t H_t\rho_t \\ - \int_\Theta \overline{\mathcal{N}}_t^2\rho_t - \int_\Theta |\partial_H\bar V_t|^2\rho_t - 2\int_\Theta \partial_H\bar V_t\overline{\mathcal{N}}_t\rho_t,$$

from which we obtain the desired result. □

Remark 3.4. The fact that the system (3.23) dissipates energy suggests that there might be a gradient flow structure associated to it. We claim that this is the case.

Indeed, denote by $\Gamma(\rho^{(1)},\rho^{(2)})$ the set of transport plans between $\rho^{(1)}$ and $\rho^{(2)}$, namely

$$\Gamma(\rho^{(1)},\rho^{(2)}) := \left\{\gamma\in\mathscr{P}(\Theta\times\Theta) : \pi^j_\#\gamma = \rho^{(j)}\right\},$$

where $\pi^j:\Theta\times\Theta\to\Theta$, $j=1,2$, are the canonical projection onto the first and second factor respectively. Then, we consider the distance between $(\rho^{(1)},H^{(1)})$ and $(\rho^{(2)},H^{(2)})$ given by

$$\mathcal{D}^2((\rho^{(1)},H^{(1)}),(\rho^{(2)},H^{(2)})) := \\ := \inf_{\gamma\in\Gamma(\rho^{(1)},\rho^{(2)})}\int_{\Theta\times\Theta}\left(|\theta_1-\theta_2|^2 + |H^{(1)}(\theta_1)-H^{(2)}(\theta_2)|^2\right)d\gamma(\theta_1,\theta_2).$$

Alternatively, one can define it dynamically (*à la* Benamou-Brenier, see for instance [FG20, Chapter 4.2]) as

$$\begin{aligned}\mathcal{D}^2((\rho^{(1)},H^{(1)}),(\rho^{(2)},H^{(2)})) :=\\ := \inf\Bigg\{ \int_0^1 \int_\Theta |\boldsymbol{v}_s|^2 (1+|\dot{H}_s|^2)\rho_s \, ds : \partial_s\rho_s + \operatorname{div}(\boldsymbol{v}_s\rho_s) = 0,\\ \rho_0 = \rho^{(1)}, \rho_1 = \rho^{(2)}, H_0 = H^{(1)}, H_1 = H^{(2)}\Bigg\}.\end{aligned}$$

Then, a classical but tedious computation shows that, at least formally, the gradient flow of G with the distance $\mathcal{D}$ is given by (3.23).

It would be interesting to make this argument rigorous (perhaps using a scheme *à la* JKO [JKO98, AGS08]), and to use this gradient flow interpretation to better study (3.23).

4. Regularized problems

In order to study the behavior of solutions to the PDEs constructed in the previous sections (namely (3.7), (3.19), (3.21)-(3.22), or (3.23)-(3.24)), it is sometimes convenient to regularize them by adding a small perturbation to the energy functional (or the PDE) that regularizes it.

For simplicity, we focus here on (3.7), although similar discussions could be done to the other PDEs. We present here two possible of such strategies, by converting the original PDE into a heat-type equation or a porous medium-type equation.

4.1. **Heat regularization.**

A natural way to control the degeneracy of critical points of our functional in (2.7) or (3.2), is to add a small entropy term in the minimization procedure. That is, for $\tau > 0$, consider the functional

$$F_\tau(\mu) = \frac{1}{2}\int_D \left(\int_\Omega \Phi(\xi,x)\mu(d\xi) - f(x)\right)^2 dx + \int_\Omega V(\xi)\,\mu(d\xi) + \tau \operatorname{Ent}(\mu) \quad (4.1)$$

where

$$\operatorname{Ent}(\mu) := \begin{cases} \int_\Omega \rho(\xi)\log(\rho(\xi))d\xi & \text{if } \mu(d\xi) = \rho(\xi)\,d\xi,\\ +\infty & \text{if } \mu \not\ll d\xi.\end{cases}$$

Adding this entropy term corresponds to a variation in the stochastic gradient descent in which, when performing discrete in time approximations of (2.4), one adds a noisy diffusion term. Alternatively, in terms of the PDE describing the evolution of the gradient flow in the Wasserstein metric of F_τ, the addition of the entropy corresponds to adding a small diffusive term in the right-hand side of (3.7) (see for instance [JKO98]). Thus, if μ_t is the Wasserstein gradient flow of F_τ, then

$$\partial_t\mu_t = \operatorname{div}\left(\mu_t\nabla\mathcal{L}(\mu_t)\right) + \operatorname{div}\left(\mu_t\nabla\mathcal{S}\right) + \operatorname{div}(\mu_t\nabla V) + \tau\Delta\mu_t, \quad (4.2)$$

where we are using the notation in (3.4) and (3.8).

As before, when Ω is not $\mathbb{R}^{d+1}$, we add zero Neumann boundary conditions:

$$\boldsymbol{\nu} \cdot \{\nabla \left(\mathcal{L}(\mu_t) + \mathcal{S} + V\right) \mu_t + \tau \nabla \mu_t\} = 0 \quad \text{on } \partial\Omega. \tag{4.3}$$

The PDE (4.2)-(4.3) presents a nicer structure than the original (3.7), and it has been studied in the context of training shallow neural networks. In particular, in [MMN18, JMM19], this equation appears when approximating functions f by an increasing number of "bumps". There, the authors prove existence and uniqueness of solutions (even in domains with Neumann boundary conditions (4.3)), and they provide some regularity and convergence estimates for solutions. Observe that, in this case, one gets immediate smoothing (and also immediate full support) for μ_t.

It is interesting to rewrite (4.1) in a different way, in terms of stationary solutions. Indeed, let us denote by $\mu_* \in \mathscr{P}(\Omega)$ a stationary solution (namely, such that the corresponding dissipation vanishes, see Definition 3.1). Then, we can write

$$\begin{aligned} F_\tau(\mu) - F_\tau(\mu_*) = &\int_{\Omega\times\Omega} K(\xi,\bar\xi)\mu(d\xi)\mu(d\bar\xi) - \int_{\Omega\times\Omega} K(\xi,\bar\xi)\mu_*(d\xi)\mu_*(d\bar\xi) \\ &+ \int_\Omega \mathcal{S}(\xi)(\mu-\mu_*)(d\xi) + \int_\Omega V(\xi)(\mu-\mu_*)(d\xi) \\ &+ \tau\left(\mathrm{Ent}(\mu) - \mathrm{Ent}(\mu_*)\right). \end{aligned} \tag{4.4}$$

On the other hand, since μ_* is a stationary solution, it has full support and the first variation density of F_τ must be constant everywhere. That is,

$$\frac{\delta F_\tau}{\delta \mu}(\mu_*) = 2\int_\Omega K(\xi,\bar\xi)\mu_*(d\bar\xi) + \mathcal{S}(\xi) + V(\xi) + \tau\log(\mu_*) \equiv \lambda \quad \text{in } \Omega$$

for some $\lambda \in \mathbb{R}$. In particular, integrating with respect to both μ_* and μ, we get

$$\begin{aligned} 2\int_{\Omega\times\Omega} K(\xi,\bar\xi)\mu_*(d\xi)\mu_*(d\bar\xi) + \int_\Omega \left(\mathcal{S}(\xi)+V(\xi)\right)\mu_*(d\xi) + \tau\int_\Omega \log(\mu_*)\mu_* = \lambda, \\ 2\int_{\Omega\times\Omega} K(\xi,\bar\xi)\mu(d\xi)\mu_*(d\bar\xi) + \int_\Omega \left(\mathcal{S}(\xi)+V(\xi)\right)\mu(d\xi) + \tau\int_\Omega \log(\mu_*)\mu = \lambda. \end{aligned}$$

We can now subtract the previous two expressions and substitute in (4.4) to obtain

$$F_\tau(\mu) = F_\tau(\mu_*) + \int_{\Omega\times\Omega} K(\xi,\bar\xi)(\mu_t(d\xi)-\mu_*(d\xi))(\mu_t(d\bar\xi)-\mu_*(d\bar\xi)) + \tau D_{KL}(\mu\|\mu_*), \tag{4.5}$$

where

$$D_{KL}(\mu\|\mu_*) := \begin{cases} \int_\Omega \mu \log\left(\frac{\mu}{\mu_*}\right) & \text{if } \mu \ll \mu_*, \\ \\ +\infty & \text{if } \mu \not\ll \mu_*, \end{cases}$$

is the relative entropy (also called Kullback-Leibler divergence) of μ with respect to μ_*. Note that the middle term in (4.5) is always non-negative (since K is positive semi-definite), and that $D_{KL}(\mu\|\nu) \geq 0$ with equality if and only if $\mu = \nu$. In particular, $F(\mu) \geq F(\mu_*)$ with equality if and only if $\mu = \mu_*$.

Hence, besides obtaining a nice expression for F_τ in terms of a stationary solution, (4.5) also shows that stationary solutions are unique and they coincide with the unique minimizer of the functional (4.1).

4.2. **The porous medium regularization.** Another possible regularization, that has been much less studied in this context, is the one arising from the porous medium equation.

In this case, we consider the functional

$$F_\tau(\mu) = \frac{1}{2}\int_D \left(\int_\Omega \Phi(\xi, x)\mu(d\xi) - f(x)\right)^2 dx + \int_\Omega V(\xi)\,\mu(d\xi) + \frac{\tau}{2}\int_\Omega \mu^2 \quad (4.6)$$

for some small parameter $\tau > 0$ (again, $\int_\Omega \mu^2 = +\infty$ by definition if μ is not absolutely continuous). Then, the Wasserstein gradient flow is given by

$$\partial_t \mu_t = \operatorname{div}\left(\mu_t \nabla \mathcal{L}(\mu_t)\right) + \operatorname{div}\left(\mu_t \nabla \mathcal{S}\right) + \operatorname{div}(\mu_t \nabla V) + \tau \operatorname{div}(\mu_t \nabla \mu_t), \qquad (4.7)$$

(with the analogous Neumann boundary condition when Ω is not $\mathbb{R}^{d+1}$, cf. (3.11) and (4.3)).

In this context, one still expects nice properties of the corresponding evolution of the gradient flow, consistent with those in the porous medium equation [Vaz06]. In particular, any stationary solution should have full support (since the support increases with time, up until covering the whole domain). and the same reasoning as in the case of the heat regularization (which was based on the full support of a stationary solution μ_*) applies, and we get

$$F_\tau(\mu) = F(\mu_*) + \int K(\xi, \bar{\xi})(\mu_t(d\xi) - \mu_*(d\xi))(\mu_t(d\bar{\xi}) - \mu_*(d\bar{\xi})) + \frac{\tau}{2}\int_\Omega (\mu - \mu_*)^2, \qquad (4.8)$$

which is similar to (4.5), where the relative entropy is substituted by the L^2 distance. Thus, from (4.8) we also get the uniqueness of stationary solutions (and hence, they coincide with the unique minimizer).

Remark 4.1. The two previous regularizations also make sense in the (ρ, H) formulation setting. In particular, one could also add a Laplacian or porous medium term to the PDE transporting ρ in (3.19), (3.21), or (3.23), in order to obtain improved convergence properties.

4.3. **An observation without regularization.** We can also rewrite the functional F in (3.3) in terms of a local minimizer (thus removing the explicit dependence on f in its expression), even in the case without regularization.

That is, let μ_* be a local minimizer for F. In particular, it is a stationary point, and it satisfies[11]

$$\frac{\delta F}{\delta \mu}(\mu_*) \equiv \lambda \qquad \text{in} \quad \text{supp}(\mu_*). \tag{4.9}$$

Moreover, from the local minimality condition, we also have [12]

$$\frac{\delta F}{\delta \mu}(\mu_*) \geq \lambda \qquad \text{in} \quad \Omega. \tag{4.10}$$

So, combining (4.9)-(4.10), and proceeding as in the regularized cases, we obtain

$$F(\mu_t) - F(\mu_*) \geq \int \left(\int \Phi(\xi, x)(\mu_t(d\xi) - \mu_*(d\xi)) \right)^2 dx.$$

In particular we recover the uniqueness of local minimizers under Φ, that we already knew by Remark 2.1.

5. Open questions

We conclude this manuscript by discussing some open questions that we believe to have a mathematical interest.

5.1. **Regularity and convergence.** One of the main open questions is concerned the convergence properties of our gradient flows, and its relation to the discrete version of the gradient descent. The main currently known results in this direction can be found (and referenced) in [CB18], where the authors are able to prove the consistency between the many neurons limits and the Wasserstein gradient flow as time goes to infinity, whenever such limits exist. Nonetheless, many questions remain open in this setting, starting from a quantitative (uniform) convergence to the Wasserstein gradient flow, in the limits as $N \to \infty$ and $t \to \infty$. Furthermore, the results in [CB18] use the specific (homogeneous) structure of the activation function. Thus, the results

[11] To see this, take $\varphi \in C_c^\infty(\Omega)$ with $\int_\Omega \varphi(\xi)\mu_*(d\xi) = 0$, and for $|\varepsilon| \ll 1$ we consider the variation $\mu_\varepsilon := (1 + \varepsilon\varphi)\mu_* \in \mathscr{P}(\Omega)$. Then, by local minimality, we get

$$0 = \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} F(\mu_\varepsilon) = \int_\Omega \frac{\delta F}{\delta \mu}(\mu_*)(\xi)\, \varphi(\xi)\, \mu_*(d\xi) \qquad \forall\, \varphi \in C_c^\infty(\Omega) \text{ s.t. } \int_\Omega \varphi(\xi)\mu_*(d\xi) = 0.$$

By the arbitrariness of φ, this implies that $\frac{\delta F}{\delta \mu}(\mu_*)$ is constant on $\text{supp}(\mu_*)$.

[12] To see this, given $\nu \in \mathscr{P}(\Omega)$, for $\varepsilon \in [0,1]$ we consider the variation $\mu_\varepsilon := (1-\varepsilon)\mu_* + \varepsilon\nu \in \mathscr{P}(\Omega)$. Then, by local minimality, we get

$$\begin{aligned} 0 \leq \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} F(\mu_\varepsilon) &= \int_\Omega \frac{\delta F}{\delta \mu}(\mu_*)(\xi)\nu(d\xi) - \int_\Omega \frac{\delta F}{\delta \mu}(\mu_*)(\xi)\mu_*(d\xi) \\ &= \int_\Omega \frac{\delta F}{\delta \mu}(\mu_*)(\xi)\nu(d\xi) - \lambda = \int_\Omega \left[\frac{\delta F}{\delta \mu}(\mu_*)(\xi) - \lambda \right] \nu(d\xi), \end{aligned}$$

where the second equality follows from (4.9). By the arbitrariness of $\nu \in \mathscr{P}(\Omega)$, this implies that $\frac{\delta F}{\delta \mu}(\mu_*)$ is everywhere greater than or equal to λ.

included in [CB18] in more general settings remain open, even if one assumes discriminating smooth kernels.

Concerning the continuous formulation (3.7), this PDE pose a series of interesting challenges. For example:

(i) What are reasonable assumptions on μ_0 and the data, to expect a conservation of its smoothness over time? (That is, to avoid convergence in finite and/or infinite time to a singular measure.)
(ii) It looks likely to us that one can prove a qualitative rate of convergence, using for instance the approach in [CGW20]. More challenging and relevant in this setting is to obtain quantitative convergence rates. Such quantification seems far from being easy in the μ formulation case, where one would need to find an "entropy-entropy dissipation inequality", showing that the dissipation (3.9) controls $F(\mu_t) - F(\mu_*)$, at least when the μ_t is close to the minimizer μ_*.
(iii) Even in the regularized cases (4.2) or (4.7), finding quantitative rates of convergence is an interesting open problem.[13]

5.2. **Multi-layer neural networks.** Training a multi-layer neural network corresponds to the approximation problem of a given function $f \in L^2(D)$, where the approximating functions are obtained by iterations of the construction in (2.1).

Assume for simplicity that $h(\theta, x) = \sigma(\theta \cdot x)$ and ignore the independent term (i.e., $n = d$ and $\theta^{(d)} = 0$, see (2.2)). Then, in the two-layer case, given an input $x \in \mathbb{R}^n$, we want to approximate a given output $f(x)$ through a neural network with two hidden layers, consisting of N_1 and N_2 neurons each. Let us denote the parameters in this case as $\{w_j\}_{1 \le j \le N_2}$ with $w_i \in \mathbb{R}$, $\{\theta_j\}_{1 \le j \le N_1}$ with $\theta_j \in \mathbb{R}^n$, and $\{b_{ji}\}_{1 \le j \le N_2, 1 \le i \le N_1}$ with $b_{ji} \in \mathbb{R}$. The corresponding approximating function is then given by

$$\sum_{j=1}^{N_2} w_j \sigma\bigg(\sum_{i=1}^{N_1} b_{ji} \sigma(\theta_i \cdot x) \bigg), \tag{5.1}$$

for some activation function $\sigma : \mathbb{R} \to \mathbb{R}$ (see Figure 5.2 and compare with Figure 2.1). Thus, we want to optimize the parameters in order to minimize a functional of the form

[13]Consider for simplicity the PDE (4.2) with $V \equiv 0$. Then, assuming that for t large μ_t is close in some strong sense to the stationary state μ_*, and that μ_* is smooth and has full support, then one can get an inequality of the form

$$\frac{d}{dt} F_\tau(\mu_t) \le -c \left(F_\tau(\mu_t) - F_\tau(\mu_*) + \tau \mathcal{F}(\mu_t, \mu_*) \right)^2$$

for some suitable function $\mathcal{F}(\mu_t, \mu_*)$ such that $\mathcal{F}(\mu_t, \mu_*) \to 0$ as $t \to \infty$. This suggests a rate of convergence of the form $F_\tau(\mu_t) - F_\tau(\mu_*) \sim \frac{1}{t}$, at least in the regularized case.

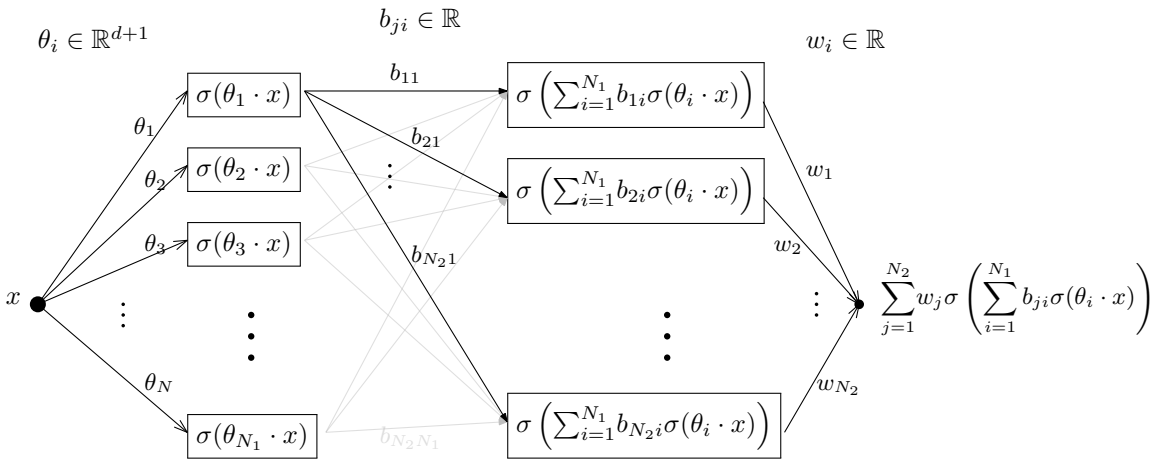

FIGURE 5.2. Graphic representation of the approximating functions given by what is known as a *2-layer neural network*, (5.1).

$$\frac{1}{2}\int_D \left(\sum_{j=1}^{N_2} w_j \sigma\left(\sum_{i=1}^{N_1} b_{ji}\sigma(\theta_i \cdot x)\right) - f(x)\right)^2 dx.$$

The corresponding expression of the previous functional in the (appropriate) limit $N_1, N_2 \to \infty$ is an interesting open problem, and some possible interpretations have recently been suggested in [AOY19, Ngu19, NP20, SS20b]. However, a simple unified connection between multiple layers neural networks and Wasserstein gradient flows, as the one presented in this paper, seems to be missing.

In this direction, it might be worth mentioning that the (ρ, H)-approach seems more adequate when dealing with systems in which one needs to consider separately each of the layers: already in the single layer case, the (ρ, H)-formulation is the one that takes advantage of the structure of the activation functions $w\, h(\theta, x)$. Even there, however, one does not fully take advantage of the linear structure of $h(\theta, x) = \sigma(\theta \cdot x)$ inside the function σ.

REFERENCES

[AGS08] L. Ambrosio, N. Gigli, G. Savare, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lectures in Mathematics, Springer, 2008.

[AOY19] D. Araújo, R. I. Oliveira, D. Yukimura, *A mean-field limit for certain deep neural networks*, Preprint arXiv `https://arxiv.org/abs/1906.00193`

[CGW20] J. Carrillo, R. Gvalani, J. Wu, *An invariance principle for gradient flows in the space of probability measures*, Preprint arXiv `https://arxiv.org/abs/2010.00424`

[CB18] L. Chizat, F. Bach, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, Advances in Neural Information Processing Systems (NeurIPS), 2018.

[E17] W. E, *A Proposal on Machine Learning via Dynamical Systems*, Commun. Math. Stat. 5 (2017), 1-11.

[EHL19] W. E, J. Han, Q. Li, *A mean-field optimal control formulation of deep learning*, Res Math Sci 6 (2019), 10.

[EMW19] W. E, C. Ma, L. Wu, *Machine learning from a continuous viewpoint*, Preprint arXiv `https://arxiv.org/abs/1912.12777`

[FG20] A. Figalli, G. Glaudo, *An invitation to Optimal Transport, Wasserstein Distances and Gradient Flows*, Preprint, 2020

[JMM19] A. Javanmard, M. Mondelli, A. Montanari, *Machine learning from a continuous viewpoint*, Preprint arXiv `https://arxiv.org/abs/1912.12777`

[JKO98] R. Jordan, D. Kinderlehrer, F. Otto, *The variational formulation of the Fokker-Planck equation*, Siam. J. Math. Anal. 29 (1998), 1-17.

[MMN18] S. Mei, A. Montanari, P. Nguyen, *A mean field view of the landscape of two-layer neural networks*, PNAS, 115 (2018), 7665-7671. ArXiv version: `https://arxiv.org/abs/1804.06561`

[Ngu19] P.-M. Nguyen, *Mean field limit of the learning dynamics of multilayer neural networks*, Preprint arXiv `https://arxiv.org/abs/1902.02880`

[NP20] P.-M. Nguyen, H.-T. Pham, *A rigorous framework for the mean field limit of multilayer neural networks*, Preprint arXiv `https://arxiv.org/abs/2001.11443`

[RV18] G. Rotskoff, E. Vanden-Eijnden, *Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error*, preprint arXiv (2018).

[SS20] J. Sirignano, K. Spiliopoulos, *Mean field analysis of neural networks: a law of large numbers*, SIAM J. Appl. Math., 80(2), 725-752.

[SS20b] J. Sirignano, K. Spiliopoulos, *Mean field analysis of deep neural networks*, Math. Oper. Res. (to appear).

[Vaz06] J. L. Vázquez, *The Porous Medium Equation: Mathematical Theory*, Oxford Lecture Series in Mathematics and its Applications, 33. Oxford University Press, Oxford, 2006.

EPFL, SB Station 8, CH-1015 Lausanne, Switzerland
Email address: `xavier.fernandez-real@epfl.ch`

ETH Zürich, Department of Mathematics, Rämistrasse 101, 8092 Zürich, Switzerland
Email address: `alessio.figalli@math.ethz.ch`