Università degli studi di Pavia
Dipartimento di Matematica
Corso di Laurea Magistrale in Matematica

UNIVERSITÀ
DI PAVIA

Gradient Flows and Optimal Transport in Neural Network Training

Tesi di Laurea Magistrale in Matematica

Relatore (Supervisor):
Matteo Negri

Tesi di Laurea di:
Edward Francisco Buccieri
Matricola 547273

Anno Accademico 2025/2026

Contents

Notation

Domains	
D	Domain of the input data, typically an open bounded subset of $\mathbb{R}^d$.
Ω	Parameter space for the neuronal map, typically a closed convex set in $\mathbb{R}^{d+2}$.

Optimal Transport and Measure Theory	
$T_{\#}\mu$	Push-forward of the measure μ through the measurable map T.
$\Gamma(\mu, \nu)$	Set of transport plans (or couplings) between measures μ and ν.
$W_2(\mu, \nu)$	2-Wasserstein distance between probability measures μ and ν.

Neural Networks and Optimization	
$u = (w, \tilde{\theta})$	Parameter vector of a single neuron, with $\tilde{\theta} = (\theta, b)$.
φ_u	Neuronal map evaluated at $u \in \Omega$.
Φ_N	Shallow neural network with N neurons.
Φ_μ	Shallow neural network parameterized with a measure μ
$R(\Phi_\mu)$	Loss function evaluated on the network Φ_μ.
$V(u)$	Regularization potential term evaluated at $u \in \Omega$.
$\tilde{F}(\mu)$	Regularized loss functional evaluated at μ.
$\sigma(s)$	Activation function, such as the ReLU or the sigmoid.

Introduction

In this thesis, we will see how the theory of gradient flows and optimal transport can be used to describe neural network training. In recent years, artificial neural networks have become a very useful tool in supervised machine learning, where the goal is to write an algorithm that can learn from a set of labeled data (training set) and generalize to new unseen data. A neural network is essentially an algorithm that emulates biological neural processes and, from a theoretical point of view, it can be understood as a parameterized function whose parameters can be adjusted to learn an objective function which represents the training set. This process of adjustment is called training and its dynamics can be described as a non-convex minimization problem, where the target is a functional representing the total error in the approximation, called loss function. Here, we consider the case of shallow neural networks, i.e. neural networks with only one hidden layer. The classical formulation of training can be seen as a gradient flow with respect to the Euclidean norm. While regularity assumptions on the loss function guarantee the local existence and uniqueness of this flow, the non-convexity of the optimization landscape complicates the analysis of convergence of the flow to global minimizers. This non-convexity drives us to a reformulation of training in terms of probability measures, which also allows us to study the limit of the network as the number of neurons goes to infinity. To explore this formulation, we fix a closed convex subset $\Omega \subset \mathbb{R}^{d+2}$ and interpret the set of parameters of N neurons $U = (u_1, \dots, u_N) \in \Omega^N$ as an interactive particle system, considering the empirical measure $\mu_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{u_n}$ associated with them. Starting from an initial empirical measure $\mu_{N,0}$, we prove that there exists a unique particle gradient flow $U : [0, +\infty) \to \Omega^N$ for the loss functional $\tilde{F}_N : \Omega^N \to [0, +\infty)$, defined as

$$\Omega^N \ni U \mapsto \tilde{F}_N(U) := \tilde{F}\left(\mu_N\right),$$

where $\tilde{F} : \mathcal{P}_2(\Omega) \to [0, +\infty)$ is a regularized loss functional, typically composed of a risk term and a λ-convex potential term ($\lambda \leq 0$). Considering $\mu_{N,0}$ as a sampling of a probability measure $\mu_0 \in \mathcal{P}_2(\Omega)$, we also prove that the particle gradient flow converges to the unique Wasserstein gradient flow for $\tilde{F}$ starting from μ_0 when N goes to infinity. This result is called many-particle limit and allows us to interpret training of an infinite-width shallow neural network $\Phi_\mu(x) = \int_\Omega \varphi_u(x)\, d\mu(u)$ as a Wasserstein gradient flow governed by the continuity equation $\partial_t \mu_t + \operatorname{div}(v_t \mu_t) = 0$, which can be approximated using the JKO

scheme, i.e. given $\tau > 0$

$$\mu_\tau^k \in \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\Omega)} \left\{ \tilde{F}(\mu) + \frac{1}{2\tau} W_2^2(\mu, \mu_\tau^{k-1}) \right\}.$$

For existence and uniqueness of the Wasserstein gradient flow, and for the results cited above, we follow the setting of [12]. Since stationary points for $\tilde{F}$ are not always global minimizers, we recall from [12] the analytical conditions, specifically regarding the homogeneity of the neuronal map associated to the neural network and the topological separation of $\mathrm{supp}(\mu_0)$, under which the Wasserstein gradient flow is guaranteed to converge to a global minimizer of $\tilde{F}$. The same results show that, in the many-particle limit, the particle gradient flow converges to the same global minimizer. Considering a finite number of neurons, the evolution problems governing the training dynamics using the particle gradient flow and the Wasserstein gradient flow are driven by different geometries: the first evolves according to the Euclidean metric, while the second follows the W_2 metric. In Remark 21, we show that, whenever $[0, +\infty) \ni t \mapsto U(t) \in \Omega^N$ is a particle gradient flow for $\tilde{F}_N$, then $t \mapsto \mu_{N,t} := \frac{1}{N} \sum_{n=1}^N \delta_{u_n(t)} \in \mathcal{P}_2(\Omega)$ is already a Wasserstein gradient flow for $\tilde{F}$. To explore visible differences between these two dynamics, we conduct a series of numerical experiments, where we approximate various target functions, using a neural network with a ReLU activation, which is a standard choice in many applications and it is covered by one of the global convergence results of [12]. To discretize the Wasserstein gradient flow, we use the JKO scheme, calculating the W_2-distance by implementing the Sinkhorn algorithm to reduce the computational cost. To discretize the particle gradient flow, we use the batch gradient descent algorithm. In the experiments, we initialize the parameters according to Theorem 3.3.3, which is the global convergence result that includes the ReLU activation function. We note that the shape of the support of the global minimizer is identical for the two flows, consistent with Theorem 3.3.3. Moreover, the difference between the dynamics is minimal according to the many-particle limit, since we choose a high number of neurons $N = 2000$. Using optimal transport to describe the training of deep neural networks is a natural extension of this thesis, but it remains an open problem in general, since the composition of layers means that the output of a neuron becomes the input of another neuron in the next layer. Even in the simplified case of deep linear networks, where this composition becomes a product of matrices, establishing infinite-width limits requires tools different from optimal transport [29]. Recent works such as [27] and [28] make progress in this direction by considering architectures (infinitely deep ResNets and Transformers, respectively) whose specific structure makes them more tractable with optimal transport tools.

Chapter 1

Gradient flows in $\mathbb{R}^N$

In this chapter we will talk about **gradient flows** in $\mathbb{R}^N$, i.e. first order systems of **ordinary differential equations (ODEs)** of the form

$$\mathbf{u}'(t) = -\nabla \mathbf{F}(t, \mathbf{u}(t)), \tag{1.1}$$

where $\mathbf{u} : [t_0, T) \to \mathbb{R}^N$ and $\mathbf{F} : U \subset \mathbb{R}^{1+N} \to \mathbb{R}^N$ are functions with a **regularity** that we will discuss in the sequel, with $t_0 \in \mathbb{R}$, $T > t_0$ finite or infinite and U being a region (an open and connected subset). $\nabla \mathbf{F}$ denotes the **gradient** of $\mathbf{F}$ with respect to $\mathbf{u}$. The minus sign is conventional and we use it when we want to solve a minimization problem. If we want to solve **ODEs** it is important to understand when we can have **solutions**, so we can study their **uniqueness** (considering some **initial conditions**) and eventually their regularity. This is also important, more generally, for **partial differential equations (PDEs)**, but we will talk about these equations in the next chapter.

1.1 Existence, uniqueness and regularity of solutions

In order to discuss some general results with respect to ODEs, we can consider, instead of a gradient flow, an equation of the form:

$$u' = f(t, u). \tag{1.2}$$

This means that we are considering $N = 1$ and a general function $f : U \subset \mathbb{R}^2 \to \mathbb{R}$, instead of $-\nabla \mathbf{F}$. For now, we omit the dependence of t in the equation to simplify the notation.

It is important to specify that this equation can have an infinite number of **solution curves**, so we have to consider an initial condition to choose a particular one. Given $u_0 \in \mathbb{R}$, we seek a solution to (1.2) for $t > t_0$ such that

$$u(t_0) = u_0. \tag{1.3}$$

The differential equation (1.2) together with the initial condition (1.3) is called an **initial value problem**.

In general, even if $f(\cdot, \cdot)$ is a continuous function, there is no guarantee

that the initial problem possesses a unique solution. Consider, for example, the initial value problem:

$$\begin{cases} u' = u^{2/3}, \\ u(0) = 0; \end{cases}$$

This has at least two solutions: $u(t) \equiv 0$ and $u(t) = \frac{t^3}{27}$.

Fortunately, we can guarantee the existence and uniqueness of a solution to our problem under a further mild condition. The result is encapsulated in the next theorem.

Theorem 1.1.1 (**Picard's Theorem**). *Suppose that $f(\cdot, \cdot)$ is a continuous function of its arguments in a region U of the (t, u) plane that contains the rectangle*

$$R = \{(t, u) \in \mathbb{R}^2 : t_0 \leq t \leq T_M,\ |u - u_0| \leq U_M\},$$

where $T_M > t_0$ and $U_M > 0$ are constants. Suppose also, that there exists a constant $L > 0$ such that

$$|f(t, u) - f(t, v)| \leq L\,|u - v| \tag{1.4}$$

holds whenever (t, u) and (t, v) lie in the rectangle R. Finally, letting

$$M = max\{|f(t, u)| : (t, u) \in R\},$$

suppose that $M(T_M - t_0) \leq U_M$. Then there exists a unique continuously differentiable function $t \mapsto u(t)$, defined on the closed interval $[t_0, T_M]$, which satisfies (1.2) and (1.3).

The expression (1.4) means that f is **Lipschitz continuous** in u with **Lipschitz constant** L independent of t.

We will not see the proof of Picard's Theorem, but we can talk about its logic. The essence of the proof is to consider the sequence of functions $\{u_n\}_{n=0}^{\infty}$, defined recursively through what is known as the *Picard Iteration*:

$$\begin{aligned} &u_0(t) \equiv u_0, \\ &u_n(t) = u_0 + \int_{t_0}^{t} f(\eta, u_{n-1}(\eta))d\eta, \quad n = 1, 2, \ldots, \end{aligned} \tag{1.5}$$

and show, using the conditions of the theorem, that $\{u_n\}_{n=0}^{\infty}$ converges uniformly on the interval $[t_0, T_M]$ to

$$u(t) = u_0 + \int_{t_0}^{t} f(\eta, u(\eta))d\eta.$$

This implies that u is continuously differentiable on $[t_0, T_M]$ and it satisfies our initial value problem.

Picard's Theorem has a natural extension to an initial value problem for a

system of N ODEs of the form

$$\begin{cases} \mathbf{u}' = \mathbf{f}(t, \mathbf{u}), \\ \mathbf{u}(t_0) = \mathbf{u}_0, \end{cases} \tag{1.6}$$

where $\mathbf{y}_0 \in \mathbb{R}^N$ and $f : [t_0, T_M] \times \mathbb{R}^N \to \mathbb{R}^N$. On introducing the **Euclidean norm** on $\mathbb{R}^N$ by

$$\|\mathbf{v}\|_2 = \left(\sum_{i=1}^{N} |v_i|^2 \right)^{\frac{1}{2}}, \quad \mathbf{v} \in \mathbb{R}^N,$$

we can state the following result.

Theorem 1.1.2 (**Picard's Theorem**). *Suppose that $\mathbf{f}(\cdot, \cdot)$ is a continuous function of its arguments in a region U of the $(t, \mathbf{u})$ space that contains the rectangle*

$$R = \{(t, \mathbf{u}) \in \mathbb{R}^{1+N} : t_0 \leq t \leq T_M, \ \|\mathbf{u} - \mathbf{u_0}\|_2 \leq U_M\},$$

where $T_M > t_0$ and $U_M > 0$ are constants. Suppose also, that there exists a constant $L > 0$ such that

$$\|\mathbf{f}(t, \mathbf{u}) - \mathbf{f}(t, \mathbf{v})\|_2 \leq L\|\mathbf{u} - \mathbf{v}\|_2 \tag{1.7}$$

holds whenever (t, u) and (t, v) lie in the rectangle R. Finally, letting

$$M = max\{\|\mathbf{f}(t, \mathbf{u})\|_2 : (t, \mathbf{u}) \in R\},$$

suppose that $M(T_M - t_0) \leq U_M$. Then there exists a unique continuously differentiable function $t \mapsto \mathbf{u}(t)$, defined on the closed interval $[t_0, T_M]$, which satisfies (1.6).

A sufficient condition for (1.7) is that $\mathbf{f}$ is continuous on R, differentiable at each point $(t, \mathbf{u}) \in int(R)$, the interior of R, and there exists $L > 0$ such that

$$\left\| \frac{\partial \mathbf{f}}{\partial \mathbf{u}}(t, \mathbf{u}) \right\| \leq L \qquad \text{for all } (t, \mathbf{u}) \in int(R), \tag{1.8}$$

where $\frac{\partial \mathbf{f}}{\partial \mathbf{u}}$ denotes the $N \times N$ Jacobi matrix of $\mathbf{u} \in \mathbb{R}^N \mapsto \mathbf{f}(t, \mathbf{u}) \in \mathbb{R}^N$, and $\|\cdot\|$ is a matrix norm subordinate to the euclidean vector norm on $\mathbb{R}^N$. This can be proved using the **Mean Value Theorem**.

The converse of this statement is not true; for example, we can take the function $\mathbf{f}(\mathbf{u}) = (|u_1|, \ldots, |u_N|)^T$, with $t_0 = 0$ and $\mathbf{u}_0 = \mathbf{0}$, so we can see that it satisfies (1.7) but not (1.8) because $\mathbf{u} \mapsto \mathbf{f}(\mathbf{u})$ is not differentiable at the point $\mathbf{u} = \mathbf{0}$.

Now we introduce the notion of *stability*.

Definition 1.1.3. *A **solution** $\mathbf{u} = \mathbf{v}(t)$ to (1.6) is said to be **stable** on the interval $[t_0, T_M]$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that for all $\mathbf{z}$ satisfying*

$\|\mathbf{v}(t_0) - \mathbf{z}\|_2 < \delta$ *the solution* $\mathbf{u} = \mathbf{w}(t)$ *to*

$$\begin{cases} \mathbf{u}' = \mathbf{f}(t, \mathbf{u}), \\ \mathbf{u}(t_0) = \mathbf{z}, \end{cases}$$

is defined and satisfies $\|\mathbf{v}(t) - \mathbf{w}(t)\|_2 < \epsilon$ *for all* $t \in [t_0, T_M]$.

A solution that is stable on $[t_0, +\infty)$ *is said to be* ***stable in the sense of Lyapunov****.*

Moreover, if

$$\lim_{t \to +\infty} \|\mathbf{v}(t) - \mathbf{w}(t)\|_2 = 0$$

then the solution $\mathbf{u} = \mathbf{v}(t)$ *is called* ***asymptotically stable****.*

The meaning of the definition of *stable solution* is that a small perturbation of the initial condition implies a small change in the solution. Using this definition, we can state the following theorem.

Theorem 1.1.4. *Under the hypotheses of Picard's theorem, the (unique) solution* $\mathbf{u} = \mathbf{v}(t)$ *to (1.6) is stable on the interval* $[t_0, T_M]$*, (assuming that* $-\infty < t_0 < T_M$*).*

Proof. Since

$$\mathbf{v}(t) = \mathbf{v}(t_0) + \int_{t_0}^{t} \mathbf{f}(\eta, \mathbf{v}(\eta)) d\eta$$

and

$$\mathbf{w}(t) = \mathbf{z} + \int_{t_0}^{t} \mathbf{f}(\eta, \mathbf{w}(\eta)) d\eta,$$

it follows that

$$\begin{aligned} \|\mathbf{v}(t) - \mathbf{w}(t)\|_2 &\leq \|\mathbf{v}(t_0) - \mathbf{z}\|_2 + \int_{t_0}^{t} \|\mathbf{f}(\eta, \mathbf{v}(\eta)) - \mathbf{f}(\eta, \mathbf{w}(\eta))\|_2 \, d\eta \\ &\leq \|\mathbf{v}(t_0) - \mathbf{z}\|_2 + L \int_{t_0}^{t} \|\mathbf{v}(\eta) - \mathbf{w}(\eta)\|_2 \, d\eta. \end{aligned} \tag{1.9}$$

Now, putting $A(t) = \|\mathbf{v}(t) - \mathbf{w}(t)\|_2$ and $a = \|\mathbf{v}(t_0) - \mathbf{z}\|_2$, (1.9) can be written as

$$A(t) \leq a + L \int_{t_0}^{t} A(\eta) d\eta, \qquad t_0 \leq t \leq T_M. \tag{1.10}$$

Multiplying (1.10) by e^{-Lt}, we find that

$$\frac{d}{dt} \left[e^{-Lt} \int_{t_0}^{t} A(\eta) d\eta \right] \leq a e^{-Lt}. \tag{1.11}$$

Integrating (1.11), we deduce that

$$e^{-Lt} \int_{t_0}^{t} A(\eta) d\eta \leq \frac{a}{L} \left(e^{-Lt_0} - e^{-Lt} \right) ,$$

that is

$$L\int_{t_0}^{t} A(\eta)d\eta \leq a\left(e^{L(t-t_0)}-1\right) \ . \tag{1.12}$$

Now substituting (1.12) into (1.10) gives

$$A(t) \leq ae^{L(t-t_0)}, \qquad t_0 \leq t \leq T_M \ . \tag{1.13}$$

The implication (1.10) $\Longrightarrow$ (1.13) is usually referred to as the **Gronwall Lemma**. Returning to our original notation, we conclude from (1.13) that

$$\|\mathbf{v}(t)-\mathbf{w}(t)\|_2 \leq \|\mathbf{v}(t_0)-\mathbf{z}\|_2 \ e^{L(t-t_0)} \ , \qquad t_0 \leq t \leq T_M \ . \tag{1.14}$$

Thus, given $\epsilon > 0$, we can choose $\delta = \epsilon e^{-L(T_M-t_0)}$ to deduce stability. □

We can observe that if either $t_0 = -\infty$ or $T_M = +\infty$, the statement of Theorem 1.1.4 is *false*. For example, the trivial solution $v \equiv 0$ to the differential equation

$$u' = u$$

is unstable on $[t_0, +\infty)$ for any $t_0 > -\infty$. In order to see that, we can consider the initial value problem that admits the trivial solution

$$\begin{cases} u' = u \\ u(t_0) = 0, \end{cases}$$

and, for every $\delta > 0$, the solution $w_\delta(t) = \alpha_\delta e^{t-t_0}$ for the initial value problem

$$\begin{cases} u' = u \\ u(t_0) = \alpha_\delta, \qquad \text{with} \quad 0 < |\alpha_\delta| < \delta. \end{cases}$$

This implies that $|v(t_0) - w_\delta(t_0)| = |\alpha_\delta| < \delta$ and that for $t \geq t_0$:

$$|v(t) - w_\delta(t)| = |w_\delta(t)| = |\alpha_\delta|\left|e^{t-t_0}\right| \geq |\alpha_\delta| =: \epsilon(\delta) > 0.$$

More generally, given the initial value problem

$$\begin{cases} u' = \lambda u \\ u(t_0) = u_0, \end{cases}$$

with $\lambda \in \mathbb{C}$, we have these possible cases for the solution $u(t) = u_0 e^{\lambda(t-t_0)}$:

- u is **unstable** for $Re(\lambda) > 0$,
- u is **stable** in the **sense** of **Lyapunov** for $Re(\lambda) \leq 0$,
- u is **asymptotically stable** for $Re(\lambda) < 0$.

Now we can return to consider a gradient flow of the form

$$\mathbf{u}'(t) = -\nabla \mathbf{F}(t, \mathbf{u}(t)), \tag{1.15}$$

and we can state the next corollary.

Corollary 1.1.5. *Given $t_0 \in \mathbb{R}$ and $\mathbf{u}_0 \in \mathbb{R}^N$, let $\mathbf{F} : U \to \mathbb{R}^N$ be a function of the class $\mathcal{C}^{1,1}(U)$, where $U \subset \mathbb{R}^{1+N}$ is a region that contains the rectangle:*

$$R = \{(t, \mathbf{u}) \in \mathbb{R}^{1+N} : t_0 \leq t \leq T_M,\ \|\mathbf{u} - \mathbf{u_0}\|_2 \leq U_M\},$$

where $T_M > t_0$ and $U_M > 0$ are constants. Letting,

$$M = max\{\|\nabla \mathbf{F}(t, \mathbf{u})\|_2 : (t, \mathbf{u}) \in R\},$$

suppose that $M(T_M - t_0) \leq U_M$. Then there exists a unique continuously differentiable solution $\mathbf{u} : [t_0, T_M] \to \mathbb{R}^N$, that satisfies (1.15) with the initial condition:

$$\mathbf{u}(t_0) = \mathbf{u}_0, \tag{1.16}$$

and that is stable on the interval $[t_0, T_M]$.

Proof. We can directly apply Picard's Theorem to $\mathbf{f} = -\nabla \mathbf{F}$. The thesis is true because, given a constant $L > 0$, $\mathbf{F} \in \mathcal{C}^{1,1}(U)$ implies that:

$$\|\nabla \mathbf{F}(t, \mathbf{u}) - \nabla \mathbf{F}(t, \mathbf{v})\|_2 \leq L\|\mathbf{u} - \mathbf{v}\|_2. \tag{1.17}$$

The stability of the unique solution follows from *Theorem 1.1.4*. □

These are the conditions that, in general, give us the existence and uniqueness of solutions in the context of gradient flows. We can obviously see that, in this case, a sufficient condition for (1.17) is (1.8) with $\mathbf{f} = -\nabla \mathbf{F}$.

1.2 Trajectories and their properties

For this section, let us consider equation (1.15) with $F : \mathbb{R}^N \to \mathbb{R}$:

$$\mathbf{u}'(t) = -\nabla F(t, \mathbf{u}(t)). \tag{1.18}$$

A solution of the equation (1.18) is called **trajectory** or **orbit** of the vector field $-\nabla F$. We can also use the term **integral curve**, if we are interested in its image in $\mathbb{R}^N$ rather than the trajectory itself as a function. In this section, we consider **maximal** solutions, i.e. trajectories $t \mapsto \gamma_{\mathbf{u}_0}(t)$ starting at the point $\mathbf{u}_0$ defined for all t in the maximal interval $[t_0, T_{\mathbf{u}_0})$, where $T_{\mathbf{u}_0}$ can be finite or infinite.

We can call (1.18) the **dynamical gradient system** generated by $-\nabla F$. This terminology suggests an evolution in time (referred to the variable t) of each point of $\mathbb{R}^N$ by the **flow** generated by $-\nabla F$, i.e. a function $\Phi(t, \mathbf{u}_0)$ that associates to each $\mathbf{u}_0 \in \mathbb{R}^N$ and $t \in [t_0, T_{\mathbf{u}_0})$ a point $\gamma_{\mathbf{u}_0}(t)$, where $\gamma_{\mathbf{u}_0}(t)$ is the unique orbit starting at $\mathbf{u}_0$ (under the hypothesis of the previous section). Note that, by Picard's theorem, this flow is unique. We can also calculate the

length of an orbit γ with this formula:

$$\text{length}(\gamma) = \int_{t_0}^{T_{\mathbf{u}_0}} \|\dot{\gamma}(t)\| dt. \tag{1.19}$$

In this section, we consider the **autonomous** version of the system (1.18) :

$$\mathbf{u}'(t) = -\nabla F(\mathbf{u}(t)). \tag{1.20}$$

We can represent the flow of this system by picturing its integral curves in $\mathbb{R}^N$, producing what is called the **phase portrait**. It is also possible to reparametrize the trajectories using a positive smooth function $g : \mathbb{R}^N \to \mathbb{R}$ without affecting the portrait of the system. This means that

$$\mathbf{u}'(t) = -g(\mathbf{u}(t))\, \nabla F(\mathbf{u}(t)) \tag{1.21}$$

has the same portrait as (1.20). The action of g can be interpreted as a change of velocity along the orbits and we can consider $-g$ to invert the direction along each single trajectory. So the system

$$\mathbf{u}'(t) = \nabla F(\mathbf{u}(t)) \tag{1.22}$$

has the same trajectories as (1.20) with the opposite orientation.

Now let us see some elementary properties of gradient systems.

- **Existence of a Lyapunov function**

  In general a *Lyapunov function* of a dynamical system, i.e. a scalar function on $\mathbb{R}^N$ that is strictly decreasing along the integral curves of the system, might not always exist for the general system (1.6). Considering a gradient system, we are lucky because F is a Lyapunov function of (1.20). Now we prove this fact in two easy steps. At first, we can see that $t \mapsto \rho(t) := F(\gamma(t))$, given any orbit $\gamma(t)$, satisfies

  $$\rho'(t) = -\nabla F(\gamma(t)) \cdot \dot{\gamma}(t) = -\|\dot{\gamma}(t)\|^2 \leq 0. \tag{1.23}$$

  If we consider $\gamma_{\mathbf{u}_0}(t)$ defined as in the beginning of the section, i.e. the unique and maximal trajectory of (1.20) starting at $\mathbf{u}_0$, it holds

  $$\nabla F(\mathbf{u}_0) \neq 0 \implies \nabla F(\gamma_{\mathbf{u}_0}(t)) \neq 0 \qquad \text{for all } t \in [0, T_{\mathbf{u}_0}). \tag{1.24}$$

  This follows from the fact that (1.22) has the same trajectories as (1.20) and has a unique flow. Let us denote by S the **set** of **critical** (or **singular**) **points** of F, that is

  $$S = \{\mathbf{u}_0 \in \mathbb{R}^n : \nabla F(\mathbf{u}_0) = 0\}. \tag{1.25}$$

  From (1.24) we can deduce that, unless $\mathbf{u}_0$ is already a singularity (i.e. a singular point of F), $\gamma_{\mathbf{u}_0}$ will never pass through S. This means that $\rho'(t) < 0$ for all $t \in [0, T_{\mathbf{u}_0})$, thus F is a Lyapunov function for (1.20).

- **Level-set parametrization**

  Let us consider $\mathbf{u}(t)$ an orbit of (1.20), and set $\mathbf{u}(0) = \mathbf{u}_0$, $r_0 = F(\mathbf{u}_0)$ and $r_\infty = \lim_{t\to T_{\mathbf{u}_0}} F(\mathbf{u}(t))$. From (1.24) we can see that whenever $\nabla F(\mathbf{u}_0) \neq 0$ the mapping $\rho(t) = F(\mathbf{u}(t))$ is a diffeomorphism between $[0, T_{\mathbf{u}_0})$ and $(r_\infty, r_0]$, and that the curve $\mathbf{w}(r) := \mathbf{u}(\rho^{-1}(r))$ satisfies the differential equation

$$\begin{aligned}\dot{\mathbf{w}}(r) = \frac{d}{dr}\left(\mathbf{u}(\rho^{-1}(r)\right) = \mathbf{u}'(t)\cdot\frac{dt}{dr} = (-\nabla F(\mathbf{u}(t))\cdot\left(\frac{1}{\frac{d}{dt}F(\mathbf{u}(t))}\right)\\ = -\frac{\nabla F(\mathbf{u}(t))}{\langle\nabla F(\mathbf{u}(t)), \mathbf{u}'(t)\rangle} = \frac{\nabla F(\mathbf{w}(r))}{\|\nabla F(\mathbf{w}(r))\|^2}.\end{aligned} \tag{1.26}$$

  Having $\rho'(t) < 0$ for all $t \in [0, T_{\mathbf{u}_0})$ implies that a gradient system does not have periodic orbits nor limit cycles. Given a bounded orbit γ, we can define its ω-limit $\Omega(\gamma)$ as

$$\Omega(\gamma) = \lim_{t_n\to T_{\mathbf{u}_0}} \gamma(t_n) \tag{1.27}$$

  and we can see that it contains only singularities of the system. This implies that, from (1.24), F is constant on $\Omega(\gamma)$ and $dist(\gamma(t), \Omega(\gamma)) \leq dist(\gamma(t), S) \to 0$ as $t \to T_{\mathbf{u}_0}$. Thus, in the context of a minimization problem, we approach a minimizer of F when we follow an orbit over time. It is important to point that $\Omega(\gamma)$ can be either a singleton or infinite.

In the previous section, we discussed the existence and uniqueness of trajectories and saw that F had to be a function of the class $\mathcal{C}^{1,1}$, in order to apply Picard's theorem. Combining that result with the last discussion, we can guarantee the convergence of the unique orbit given by an initial condition to a local minimizer of F as $t \to T_{\mathbf{u}_0}$. At this point, many questions can be asked. Such as the possibility of having the convergence of an orbit to a global minimizer of F (when it does exist), changing the hypothesis made in F. We can also study when these orbits can reach a minimizer in finite time. A natural property that emerges in optimization problems is **convexity**, mainly because, for a convex function, each minimum is actually global. But we will see that they have other useful properties. We recall the definition of **convex function**:

Definition 1.2.1. *A **function** $F : \mathbb{R}^N \to \mathbb{R}$ is said to be **convex** if*

$$F(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda F(\mathbf{x}) + (1-\lambda)F(\mathbf{y}), \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^N$$

*And F is said to be **strictly convex** if the inequality is strict.*

Intuitively, a function is convex if it lies below the chord connecting any two points on its graph.

Assuming that F is convex, we do not really need differentiability to study the integral curves of the gradient system associated with F. Actually, we

can assume that F is **lower semicontinuous** and introduce a **subgradient system**, i.e. a *differential inclusion*

$$\dot{\mathbf{u}}(t) \in -\partial F(\mathbf{u}(t)) \quad \text{(a.e.)} \tag{1.28}$$

where, fixed $\mathbf{x}_0 \in \mathbb{R}^N$, $\partial F(\mathbf{x})$ denotes the **subdifferential** of F at $\mathbf{x}$ and with "a.e." means *almost everywhere* in the sense of the Lebesgue measure. Let us give the following definitions to clarify why we introduce the notion of subdifferential:

Definition 1.2.2. *Given F with the properties assumed above, a **subgradient** of F at $\mathbf{x}$ is a point $\eta \in \mathbb{R}^N$ such that*

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \langle \eta, \mathbf{y} - \mathbf{x} \rangle \quad \textit{for all} \quad \mathbf{y} \in \mathbb{R}^N. \tag{1.29}$$

*For each $\mathbf{x} \in \mathbb{R}^N$, $\partial F(\mathbf{x})$ is the set of all subgradients of F at $\mathbf{x}$ and the **subdifferential** of F is the multivalued mapping $\partial F : \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ which assigns the set $\partial F(\mathbf{x})$ to each $\mathbf{x}$. If $\partial F(\mathbf{x}) \neq \emptyset$, F is said to be **subdifferentiable** at $\mathbf{x}$.*

Inequality (1.29) says that, given $\mathbf{x} \in \mathbb{R}^N$, the graph of the affine function $G(\mathbf{y}) = F(\mathbf{x}) + \langle \eta, \mathbf{y} - \mathbf{x} \rangle$ is a "nonvertical" supporting hyperplane at $(\mathbf{x}, F(\mathbf{x}))$ to the **epigraph** of F, i.e. the convex subset of $\mathbb{R}^{N+1}$ which contains all the points lying above the graph of F.

In this nonsmooth case, we call $\mathbf{x}^* \in \mathbb{R}^N$ a **critical point** of F if $0 \in \partial F(\mathbf{x}^*)$. Moreover, Brezis Theorem guarantees, given initial data, existence and uniqueness of absolutely continuous solution curves of the differential inclusion (1.28) and we can call them **subgradient trajectories**. Given a trajectory $\gamma : [0, T) \to \mathbb{R}^N$ of (1.28), for almost all $t \in (0, T)$ we have

$$\frac{d}{dt}(F \circ \gamma)(t) = \langle \dot{\gamma}(t), \eta \rangle, \quad \text{for all} \quad \eta \in \partial F(\gamma(t)),$$

and the function $t \mapsto \langle \dot{\gamma}(t), \eta \rangle$ is constant on $\partial F(\gamma(t))$.

If we denote by S the set of global minimizers of F, we can take $\mathbf{y} = \gamma(t)$, $\eta = -\dot{\gamma}(t)$ in (1.29) and for every $\mathbf{a} \in S$ we have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\gamma(t) - \mathbf{a}\|^2 = \frac{1}{2}\frac{d}{dt}\langle \gamma(t) - \mathbf{a}, \gamma(t) - \mathbf{a} \rangle &= \langle \dot{\gamma}(t), \gamma(t) - \mathbf{a} \rangle \\ &\leq F(\mathbf{a}) - F(\gamma(t)) \leq 0 \quad \text{a.e. on} \quad (0, +\infty), \end{aligned}$$

so the distance mapping $t \mapsto \|\gamma(t) - \mathbf{a}\|$ is nonincreasing. Since $dist(\gamma(t), S) \to 0$ as $t \to +\infty$, we deduce that every subgradient orbit of a *nonnegative* convex function F converges to a global minimizer of F. If we take the convergence in weak topology, we can guarantee the same result in an infinite-dimensional Hilbert space [4, Theorem 4]. If F is even, we can consider the strong convergence [4, Theorem 5]. In [5] there is an example of a lower semicontinuous function $\varphi : \ell^2(\mathbb{N}) \to \mathbb{R} \cup \{+\infty\}$ with minimum at 0, with a bounded gradient trajectory of infinity length which remains bounded away from 0 and does not converge for the norm topology. It is constructed starting from a family of lower

semicontinuous and convex functions $F_\lambda : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$, given $\lambda \geq 1$,

$$F_\lambda(x,y) = \begin{cases} \left[\arctan\left(\frac{x}{y}\right)\right]^\lambda \sqrt{x^2+y^2}, & \text{if } x,y \geq 0 \\ +\infty, & \text{elsewhere.} \end{cases}$$

In this case we cannot use (1.19) as the definition of length, but we can define the length of a continuous curve $\gamma : I \to \mathbb{R}^N$ as

$$length(\gamma) := \sup\left\{\sum_{i=1}^{k} dist(\gamma(t_i), \gamma(t_{i+1}))\right\}$$

where the supremum is taken over all the finite subdivisions $\{t_i\}_{i=1}^{k+1}$ of an interval $I \subset [0, +\infty)$. This definition is equivalent to (1.19) in the case where γ is at least C^1. Now we introduce a particular type of curve, which is important in the context of optimization.

Definition 1.2.3. *A **curve** $\gamma : I \to \mathbb{R}^N$ is called **self-contracted**, if for every $t_1 \leq t_2 \leq t_3$, with $t_i \in I$, we have*

$$dist(\gamma(t_1), \gamma(t_3)) \geq dist(\gamma(t_2), \gamma(t_3)). \tag{1.30}$$

In other words, for every $[a,b] \subset I$, the map

$$t \in [a,b] \mapsto dist(\gamma(t), \gamma(b))$$

is nonincreasing.

Intuitively, as time goes forward, a self-contracted curve does not fold back on itself. Note that this kind of curve might not be continuous. For example, we can take $\gamma : \mathbb{R} \to \mathbb{R}^2$ defined as

$$\gamma(t) = \begin{cases} (t,1) & \text{if } t \in (-\infty, 0) \\ (0,0) & \text{if } t = 0 \\ (t,-1) & \text{if } t \in (0, +\infty) \end{cases}$$

and see that it is a non-continuous self-contracted curve. Observe that orientation is important in this definition. For example, if $\gamma : (a,b) \subset \mathbb{R} \to \mathbb{R}^N$ is self-contracted, the curve

$$t \in (a,b) \mapsto \gamma(a+b-t)$$

might not be self-contracted.

Definition 1.2.4. *A curve $\gamma : I \to \mathbb{R}^N$ is said to **converge** to a point $\mathbf{x}_0 \in \mathbb{R}^N$ if $\gamma(t) \to \mathbf{x}_0$ as $t \to t_+ := supI$.*

*A curve $\gamma : I \to \mathbb{R}^N$ is said to be **bounded**, if its image $\gamma(I)$ is a bounded subset of $\mathbb{R}^N$.*

The next elementary result about self-contracted curves is presented in [6].

Proposition 1.2.5. *Let $\gamma : I \to \mathbb{R}^N$ be a bounded self-contracted curve and $(a,b) \subset I$. Then, γ has a limit in $\mathbb{R}^N$ whenever $t \in (a,b)$ tends to an endpoint of (a,b).*

In particular, every self-contracted curve can be extended by continuity to the endpoints of I (possibly equal to $\pm\infty$).

We have a direct consequence of Proposition 1.2.5.

Corollary 1.2.6. *Every bounded self-contracted curve $\gamma : (0,+\infty) \to \mathbb{R}^N$ converges to some point $\mathbf{x}_0 \in \mathbb{R}^N$ as $t \to +\infty$. Moreover, the function $t \mapsto dist(\mathbf{x}_0, \gamma(t))$ is non-increasing.*

We can deduce that the trajectories of a general gradient system like (1.20) might not be self-contracted curves. The following example provides a system with non-converging bounded orbits, so they cannot be self-contracted.

Example 1. *Let $F : \mathbb{R}^2 \to \mathbb{R}$ be defined (in polar coordinates) by*

$$F(rcos\theta, rsin\theta) = \begin{cases} exp\left(\dfrac{1}{r^2-1}\right) & \text{if } r < 1 \\ 0 \quad \text{if } r = 0 \\ exp\left(\dfrac{1}{1-r^2}\right) sin\left(\dfrac{1}{r-1} - \theta\right) & \text{if } r > 1 \end{cases}$$

This is a C^∞ function and a deeper description of the associated system (1.20) is given in [7, page 14].

However, we can guarantee that the solution curves are self-contracted with different assumptions on F depending on the system that we are considering [6, Proposition 6.2 & 6.4]:

- for a **subgradient system** it is enough asking F to be continuous and *convex*
- for a **gradient system** it is enough asking $F \in C^{1,1}$ and *quasiconvex*

We recall for a moment the definition of quasiconvex function.

Definition 1.2.7. *A **function** $F : \mathbb{R}^N \to \mathbb{R}$ is said to be **quasiconvex** if its sublevel sets*

$$\left\{\mathbf{x} \in \mathbb{R}^N : F(\mathbf{x}) \leq \lambda\right\}, \quad \lambda \in \mathbb{R}$$

are convex in $\mathbb{R}^N$.

*If F is **differentiable** then,*

$$F \text{ is } \textbf{quasiconvex} \iff \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^N \text{ if } \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle > 0 \text{ , then } F(\mathbf{y}) \geq F(\mathbf{x})$$

Then, we have an important result in the 2-dimensional case [6, Theorem 1.3].

Theorem 1.2.8. *Every bounded continuous self-contracted planar curve γ is of finite length. More precisely,*

$$length(\gamma) \leq (8\pi + 2)D(\gamma)$$

where $D(\gamma)$ is the distance between the endpoints of γ.

As a consequence, we have the following theorem.

Theorem 1.2.9. *Let $F : \mathbb{R}^2 \to \mathbb{R}$ be a smooth convex function with a unique minimum. Then, the trajectories γ of (1.20) have a (uniformly) finite length.*

It is important to note that it is not known whether these last two theorems hold in higher dimensions.

In order to conclude the current section, we use the **Lojasiewickz inequality** to deduce that, given $\mathbf{x}^* \in F^{-1}(0)$ a critical point of a real-analytic function $F : \mathbb{R}^N \to \mathbb{R}$, all gradient orbits of F that converge to $\mathbf{x}^*$ and lie in a neighborhood U of $\mathbf{x}^*$ have finite length.

The inequality asserts that, with the assumptions made above, there exist two constants $\theta \in [\frac{1}{2}, 1)$ and $k > 0$ such that

$$|F(\mathbf{x})|^\theta \leq k\|\nabla F(\mathbf{x})\| \qquad \forall \mathbf{x} \in U. \tag{1.31}$$

Let $\gamma : [0, +\infty) \to U$ be a gradient trajectory of F, that is, $\dot{\gamma}(t) = -\nabla F(\gamma(t))$. Then,

$$\begin{aligned} -\left(\frac{k}{1-\theta}\right)\frac{d}{dt}\left[F(\gamma(t))^{1-\theta}\right] &= -k\langle\dot{\gamma}(t), \nabla F(\gamma(t))\rangle F(\gamma(t))^{-\theta} \\ &= k\|\nabla F(\gamma(t))\|^2 F(\gamma(t))^{-\theta} \geq \|\nabla F(\gamma(t))\| = \|\dot{\gamma}(t)\|, \end{aligned}$$

yielding (since $F(\gamma_\infty) = F(\mathbf{x}^*) = 0$) that

$$length(\gamma) = \int_0^{+\infty} \|\dot{\gamma}(t)\|dt \leq \left(\frac{k}{1-\theta}\right) F(\gamma(0))^{1-\theta} < +\infty.$$

Inequality (1.31) tells us that gradient vanishing is controlled along an orbit over time. Requiring the trajectory to be located in U is not restrictive, because it can be shown that any bounded trajectory of the gradient flow associated to a real-analytic function F is eventually trapped inside a convenient ball of its cluster point. So, the tail of an orbit necessarily has a finite length, and it converges to its cluster point.

1.3 Weak formulation

In this section, we study the existence and uniqueness of solutions for a subgradient system such as (1.28). We make this discussion in order to introduce a technique that will be useful for the next chapters. Therefore, we consider the

following Cauchy problem :

$$\begin{cases} \dot{\mathbf{u}}(t) \in -\partial F(\mathbf{u}(t)) \quad \text{for a.e. } t > 0, \\ \mathbf{u}(0) = \mathbf{u}_0, \end{cases} \tag{1.32}$$

where $F : \mathbb{R}^N \to (-\infty, +\infty]$ and $\mathbf{u} : [0, T] \to \mathbb{R}^N$. We can read this problem as a **weak formulation** of the Cauchy problem associated to a gradient system, and therefore we need different techniques to prove the main result of this section.

First, we recall the following definition.

Definition 1.3.1. *A map $F : \mathbb{R}^N \to (-\infty, +\infty]$ is called **proper** if*

$$dom(F) := \{\mathbf{x} \in \mathbb{R}^N : F(\mathbf{x}) < +\infty\} \neq \emptyset$$

Now, we can prove the following theorem.

Theorem 1.3.2. *Let $F : \mathbb{R}^N \to (-\infty, +\infty]$ proper, convex and lower semicontinuous with*

- $\mathbf{u}_0 \in dom(F)$
- *inf $F > -\infty$.*

Then there exists $u \in H^1(0, T; \mathbb{R}^{\mathbb{N}})$ unique solution of problem (1.32).

Proof. The idea is to divide our interval $[0, T]$ into n subintervals of length $\tau := \frac{T}{n}$, which is a temporal discretization parameter for our equation.

Given $\mathbf{u}_0 \in dom(F)$, we formulate the following iterative scheme:

$$\begin{cases} u^0 := \mathbf{u}_0 \\ u_\tau^k \in \underset{\mathbf{v} \in \mathbb{R}^N}{\operatorname{argmin}} \left\{ F(\mathbf{v}) + \dfrac{1}{2\tau} \left\| \mathbf{v} - u_\tau^{k-1} \right\|^2 \right\} \quad \text{for } k = 1, \dots, n. \end{cases} \tag{1.33}$$

It can be proven that, for each $k = 1, \dots, n$, there exists a unique solution $u_\tau^k \in \mathbb{R}^N$ that satisfies inclusion:

$$-\frac{u_\tau^k - u_\tau^{k-1}}{\tau} \in \partial F(u_\tau^k). \tag{1.34}$$

It is easy to see that, with our assumptions, (1.34) gives us a necessary and sufficient condition for the optimality, i.e. :

$$u_\tau^k \in \underset{\mathbf{v} \in \mathbb{R}^N}{\operatorname{argmin}} \left\{ F(\mathbf{v}) + \frac{1}{2\tau} \left\| \mathbf{v} - u_\tau^{k-1} \right\|^2 \right\} \iff -\frac{u_\tau^k - u_\tau^{k-1}}{\tau} \in \partial F(u_\tau^k). \tag{1.35}$$

If u_τ^k is a minimizer, then, for all $\mathbf{v} \in \mathbb{R}^N$, we have

$$F(u_\tau^k) + \frac{1}{2\tau} \left\| u_\tau^k - u_\tau^{k-1} \right\|^2 \leq F(\mathbf{v}) + \frac{1}{2\tau} \left\| \mathbf{v} - u_\tau^{k-1} \right\|^2 \tag{1.36}$$

Rearranging the terms and using the identity $\frac{1}{2}\|\mathbf{a}\|^2 - \frac{1}{2}\|\mathbf{b}\|^2 = \langle \mathbf{a}, \mathbf{a}-\mathbf{b}\rangle - \frac{1}{2}\|\mathbf{a}-\mathbf{b}\|^2$ with $\mathbf{a} := -(u_\tau^k - u_\tau^{k-1})$ and $\mathbf{b} := -(\mathbf{v} - u_\tau^{k-1})$, we obtain

$$F(\mathbf{v}) \geq F(u_\tau^k) + \langle -\frac{u_\tau^k - u_\tau^{k-1}}{\tau}, \mathbf{v} - u_\tau^k\rangle - \frac{1}{2\tau}\left\|\mathbf{v} - u_\tau^k\right\|^2. \tag{1.37}$$

Now we set $\mathbf{v} := u_\tau^k + \varepsilon(\mathbf{w} - u_\tau^k) = (1-\varepsilon)u_\tau^k + \varepsilon\mathbf{v}$, where $\mathbf{w} \in \mathbb{R}^N$ and $\epsilon > 0$. Using the convexity of F, it follows that:

$$F(\mathbf{v}) \leq (1-\varepsilon)F(u_\tau^k) + \varepsilon F(\mathbf{w}). \tag{1.38}$$

Rearranging the terms in (1.38) and using (1.37) we obtain

$$F(\mathbf{w}) - F(u_\tau^k) \geq \frac{F(\mathbf{v}) - F(u_\tau^k)}{\varepsilon} \geq \langle -\frac{u_\tau^k - u_\tau^{k-1}}{\tau}, \mathbf{w} - u_\tau^k\rangle - \frac{\varepsilon^2}{2\tau}\left\|\mathbf{w} - u_\tau^k\right\|^2 \tag{1.39}$$

For $\varepsilon \to 0^+$ and for all $\mathbf{w} \in \mathbb{R}^N$ we have

$$F(\mathbf{w}) - F(u_\tau^k) \geq \langle -\frac{u_\tau^k - u_\tau^{k-1}}{\tau}, \mathbf{w} - u_\tau^k\rangle \tag{1.40}$$

and, therefore, we have (1.34).

Now we assume that (1.34) holds and, again using the identity $\frac{1}{2}\|\mathbf{a}\|^2 - \frac{1}{2}\|\mathbf{b}\|^2 = \langle \mathbf{a}, \mathbf{a}-\mathbf{b}\rangle - \frac{1}{2}\|\mathbf{a}-\mathbf{b}\|^2$, we obtain (1.36) and so

$$u_\tau^k \in \operatorname*{argmin}_{\mathbf{v}\in\mathbb{R}^N}\left\{F(\mathbf{v}) + \frac{1}{2\tau}\left\|\mathbf{v} - u_\tau^{k-1}\right\|^2\right\}.$$

The formulation (1.33) is called **Minimizing Movement Scheme** and will also be important to us in the sequel. Now we introduce two interpolating functions $\overline{u}_\tau, \hat{u}_\tau : [0,T] \to \mathbb{R}^N$ for these unique solutions $\{u_\tau^k\}_{k=1}^n$, defined $\forall t \in ((k-1)\tau, k\tau]$ as

$$\begin{aligned}\overline{u}_\tau(t) &:= u_\tau^k,\\ \hat{u}_\tau(t) &:= u_\tau^k + \frac{u_\tau^k - u_\tau^{k-1}}{\tau}(t - k\tau).\end{aligned}$$

The first is piecewise constant and therefore is not continuous, while the second is piecewise linear and therefore is continuous. Moreover, we have $\frac{d}{dt}\hat{u}_\tau(t) = \frac{u_\tau^k - u_\tau^{k-1}}{\tau}$ for all $t \in ((k-1)\tau, k\tau]$.

Now, the objective is to prove that (1.32) admits at least a solution by showing the convergence of $\overline{u}_\tau$ and $\hat{u}_\tau$ as $\tau \to 0$, because we want to take the limit of the inequality

$$F(\mathbf{w}) - F(\hat{u}_\tau(t)) \geq \langle -\frac{d}{dt}\hat{u}_\tau(t), \mathbf{w} - \hat{u}_\tau(t)\rangle \tag{1.41}$$

for all $t \in [0,T]$ and for all $\mathbf{w} \in \mathbb{R}^N$. Writing (1.36) with $\mathbf{v} = u_\tau^{k-1}$, we obtain

$$F(u_\tau^k) + \frac{1}{2\tau}\left\|u_\tau^k - u_\tau^{k-1}\right\|^2 \leq F(u_\tau^{k-1}). \tag{1.42}$$

Now we can sum from 0 to n and, remembering that $\alpha = inf\ F > -\infty$ and $\mathbf{u}_0 \in dom(F)$, we have

$$\frac{1}{2}\sum_{k=0}^{n}\frac{1}{\tau}\left\|u_\tau^k - u_\tau^{k-1}\right\|^2 \leq F(\mathbf{u}_0) - F(u_\tau^N) \leq F(\mathbf{u}_0) - \alpha =: C. \tag{1.43}$$

We can rewrite

$$\frac{\left\|u_\tau^k - u_\tau^{k-1}\right\|^2}{\tau} = \tau\left(\frac{\left\|u_\tau^k - u_\tau^{k-1}\right\|}{\tau}\right)^2 = \int_{(k-1)\tau}^{\tau}\left\|\frac{d}{dt}\hat{u}_\tau(t)\right\|^2 dt$$

to obtain a bound independent of τ for $\frac{d}{dt}\hat{u}_\tau$:

$$\frac{1}{2}\int_0^T \left\|\frac{d}{dt}\hat{u}_\tau(t)\right\|^2 dt \leq C. \tag{1.44}$$

This shows that the sequence $\{\hat{u}_\tau\}_{\tau>0}$ is uniformly bounded in $H^1(0,T;\mathbb{R}^N)$ and, therefore, there exists $v \in L^2(0,T;\mathbb{R}^N)$ such that, after extraction of a subsequence,

$$\frac{d}{dt}\hat{u}_\tau \rightharpoonup v \qquad \text{weakly in } L^2(0,T;\mathbb{R}^N). \tag{1.45}$$

Moreover, we observe that for all $t,s \in [0,T]$ with $s < t$ and using (1.43) with Hölder inequality, we have

$$\begin{aligned}\|\hat{u}_\tau(t) - \hat{u}_\tau(s)\| &\leq \int_s^t \left\|\frac{d}{dr}\hat{u}_\tau(r)\right\| dr \leq \left\|\frac{d}{dt}\hat{u}_\tau\right\|_{L^2(0,T;\mathbb{R}^N)} |t-s|^{\frac{1}{2}} \\ &\leq \sqrt{2C}\,|t-s|^{\frac{1}{2}}.\end{aligned} \tag{1.46}$$

This inequality gives us the uniform equicontinuity of the sequence $\{\hat{u}_\tau\}_{\tau>0} \subset C^0([0,T];\mathbb{R}^N)$.We can also notice that (1.44) implies $\hat{u}_\tau \in AC(0,T;\mathbb{R}^N)$ and it is easy to see that $\hat{u}_\tau$ and $\overline{u}_\tau$ are uniformly bounded in $C^0([0,T];\mathbb{R}^N)$:

$$\begin{aligned}&\|\overline{u}_\tau\|_{C^0([0,T];\mathbb{R}^N)} = \sup_{k=1,\dots,n}\left\|u_\tau^k\right\| =: C' \\ &\|\hat{u}_\tau\|_{C^0([0,T];\mathbb{R}^N)} \leq \|\mathbf{u}_0\| + \sqrt{T}\left\|\frac{d}{dt}\hat{u}_\tau\right\|_{L^2(0,T;\mathbb{R}^N)} = \|\mathbf{u}_0\| + \sqrt{2CT}.\end{aligned} \tag{1.47}$$

With (1.46) and the inequality in (1.47), we can use Ascoli-Arzelà Theorem to conclude that there exist a subsequence of $\{\hat{u}_\tau\}_{\tau>0}$ and $u \in C^0([0,T];\mathbb{R}^N)$ such that

$$\begin{aligned}&\hat{u}_\tau \to u \quad \text{in } C^0([0,T];\mathbb{R}^N) \\ &\hat{u}_\tau \rightharpoonup u \quad \text{weakly in } L^2(0,T;\mathbb{R}^N)\end{aligned} \tag{1.48}$$

and using (1.45) we can also conclude that $v = \frac{d}{dt}u$, because $\forall \varphi \in C_c^\infty(0,T;\mathbb{R}^N)$ and for $\tau \to 0$ we have:

$$\int_0^T \langle v, \varphi(t)\rangle dt \leftarrow \int_0^T \langle \frac{d}{dt}\hat{u}_\tau(t), \varphi(t)\rangle dt = -\int_0^T \langle \hat{u}_\tau(t), \varphi'(t)\rangle dt \to -\int_0^T \langle u(t), \varphi'(t)\rangle dt.$$

If we fix $t \in [0,T]$, we can find $k \in \{1,\dots,n\}$ such that $\overline{u}_\tau(t) = \hat{u}_\tau(s)$ for $s = k\tau$ with $|t-s| \leq \tau$. So, we obtain

$$\|\hat{u}_\tau(t) - \overline{u}_\tau(t)\| \leq \sqrt{2C\tau} \quad \text{for all } t \in [0,T] \tag{1.49}$$

and we can conclude that, for $\tau \to 0$

$$\begin{aligned} &\overline{u}_\tau \to u \quad \text{in } C^0([0,T];\mathbb{R}^N), \\ &\overline{u}_\tau \rightharpoonup u \quad \text{weakly in } L^2(0,T;\mathbb{R}^N). \end{aligned} \tag{1.50}$$

From the above discussion, it follows that $u \in H^1(0,T;\mathbb{R}^N)$ and we have to prove that it satisfies (1.32).

Starting from (1.34), we have $\forall \mathbf{v} \in \mathbb{R}^N$ and a.e. $t \in [0,T]$

$$F(\mathbf{v}) - F(\hat{u}_\tau(t)) + \langle \frac{d}{dt}\hat{u}_\tau(t), \mathbf{v} - \hat{u}_\tau(t)\rangle \geq 0. \tag{1.51}$$

We can integrate between t_1 and t_2, where $0 \leq t_1 < t_2 \leq T$, obtaining

$$\int_{t_1}^{t_2} F(\mathbf{v})dt \geq \int_{t_1}^{t_2} F(\hat{u}_\tau(t)) - \langle \frac{d}{dt}\hat{u}_\tau(t), \mathbf{v} - \hat{u}_\tau(t)\rangle dt. \tag{1.52}$$

Now we can take the lim inf for $\tau \to 0$, using the lower semicontinuity of F and the Fatou Lemma owing to (1.51), in order to obtain

$$\begin{aligned} \int_{t_1}^{t_2} F(\mathbf{v})dt &\geq \liminf_{\tau\to 0} \int_{t_1}^{t_2} F(\hat{u}_\tau(t)) - \langle \frac{d}{dt}\hat{u}_\tau(t), \mathbf{v} - \hat{u}_\tau(t)\rangle dt \\ &\geq \int_{t_1}^{t_2} \liminf_{\tau\to 0} F(\hat{u}_\tau(t)) - \langle \frac{d}{dt}\hat{u}_\tau(t), \mathbf{v} - \hat{u}_\tau(t)\rangle dt \\ &\geq \int_{t_1}^{t_2} F(u(t)) - \langle \frac{d}{dt}u(t), \mathbf{v} - u(t)\rangle dt. \end{aligned} \tag{1.53}$$

Thus, we have $\forall t \in [0,T]$

$$F(\mathbf{v}) - \frac{1}{t_1 - t_2}\int_{t_1}^{t_2} F(u(t)) + \langle \frac{d}{dt}u(t), \mathbf{v} - u(t)\rangle dt \geq 0. \tag{1.54}$$

From the density of Lebesgue points of the function $t \mapsto F(\mathbf{v}) - F(u(t)) + \langle \frac{d}{dt}u(t), \mathbf{v} - u(t)\rangle$ in $[0,T]$ we have

$$F(\mathbf{v}) \geq F(u(t)) - \langle \frac{d}{dt}u(t), \mathbf{v} - u(t)\rangle. \tag{1.55}$$

This concludes the proof of the existence.

For the uniqueness, we consider two solutions u_1 and u_2 to (1.32). This means that $\forall \mathbf{v} \in \mathbb{R}^N$ and for a.e. $s \in [0,T]$

$$\begin{aligned} F(\mathbf{v}) &\geq F(u_1(s)) + \langle \frac{d}{ds} u(s), u_1(s) - \mathbf{v} \rangle, \\ F(\mathbf{v}) &\geq F(u_2(s)) + \langle \frac{d}{ds} u(s), u_2(s) - \mathbf{v} \rangle. \end{aligned} \tag{1.56}$$

Now we take $\mathbf{v} = u_2(s)$ in the first inequality, $\mathbf{v} = u_1(s)$ in the second, and sum the two inequalities to obtain

$$\langle \frac{d}{ds}(u_1(s) - u_2(s)), u_1(s) - u_2(s) \rangle \leq 0. \tag{1.57}$$

We can rewrite (1.57) as

$$\frac{1}{2} \frac{d}{ds} \|u_1(s) - u_2(s)\|^2 \leq 0 \tag{1.58}$$

and we can integrate between 0 and $t \in (0,T]$ to obtain

$$\|u_1(t) - u_2(t)\|^2 \leq \|u_1(0) - u_2(0)\|^2 . \tag{1.59}$$

In conclusion, if we take $u_1(0) = u_2(0)$, we have $u_1(t) = u_2(t)$ for all $t \in [0,T]$ and thus the uniqueness. □

Chapter 2

Gradient flows and optimal transport

In this chapter, we will give some preliminary results on optimal transport and its applications to gradient flows. We need these results in the measure formulation of neural network training, which is one of our main focuses, and we will omit unnecessary proofs. The reference for this chapter is [8], where the reader can find all the proofs that we will omit and further details on the theory of optimal transport.

2.1 Measures and optimal transport

For simplicity, we will only consider locally compact, separable, and complete metric spaces. In our case, we introduce X as the space where the source measures are defined and Y as the space where the target measures are defined. We will only consider **Borel measures** and $T : X \to Y$ Borel-measurable maps. Moreover, we denote the spaces of probability (Borel) measures over X and Y with $\mathcal{P}(X)$ and $\mathcal{P}(Y)$, and the Euclidean norm with $|\cdot|$.

Definition 2.1.1. *Given a map $T : X \to Y$ and a probability measure $\mu \in \mathcal{P}(X)$, we can define the **image measure** (or **push-forward measure**) $T_{\#}\mu \in \mathcal{P}(Y)$ as*

$$(T_{\#}\mu)(A) := \mu(T^{-1}(A)) \qquad \text{for any } A \in \mathcal{B}(Y).$$

Where $\mathcal{B}(Y)$ denotes the class of Borel-measurable sets.

This kind of measure can be characterized as follows.

Lemma 2.1.2. *Given $T : X \to Y$, $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, we have*

$$\nu = T_{\#}\mu \quad \iff \quad \int_Y \varphi(y)\, d\nu(y) = \int_X \varphi(T(x))\, d\mu(x) \qquad \forall \varphi : Y \to \mathbb{R} \quad \text{Borel and bounded} \tag{2.1}$$

Definition 2.1.3. *Given $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, a map $T : X \to Y$ that satisfies (2.1) is called a **transport map** from μ to ν.*

Then, we have the following immediate consequence :

Corollary 2.1.4. *For any function $\varphi : Y \to \mathbb{R}$ Borel and bounded, we have*

$$\int_Y \varphi \, d(T_{\#}\mu) = \int_X \varphi \circ T \, d\mu. \tag{2.2}$$

Moreover, there is a simple relation between composition and push-forward:

Lemma 2.1.5. *Let $T : X \to Y$ and $S : Y \to Z$ be measurable functions. Then*

$$(S \circ T)_{\#}\mu = S_{\#}(T_{\#}\mu).$$

Definition 2.1.6. *We call $\gamma \in \mathcal{P}(X \times Y)$ a **transport plan** or a **coupling** of μ and ν if*

$$(\pi_X)_{\#}\gamma = \mu \quad \text{and} \quad (\pi_Y)_{\#}\gamma = \nu,$$

where

$$\pi_X(x, y) = x, \quad \pi_Y(x, y) = y \quad \forall\, (x, y) \in X \times Y.$$

This is equivalent to requiring that

$$\int_{X\times Y} \varphi(x) \, d\gamma(x, y) = \int_{X\times Y} \varphi \circ \pi_X(x, y) \, d\gamma(x, y) = \int_X \varphi(x) \, d\mu(x)$$

for all $\varphi : X \to \mathbb{R}$ Borel and bounded, and

$$\int_{X\times Y} \psi(y) \, d\gamma(x, y) = \int_{X\times Y} \psi \circ \pi_Y(x, y) \, d\gamma(x, y) = \int_Y \psi(y) \, d\nu(y)$$

*for all $\psi : Y \to \mathbb{R}$ Borel and bounded. We denote by $\Gamma(\mu, \nu)$ the **set** of **couplings** of μ and ν.*

We can easily prove that a transport map induces a coupling.

Remark 1. Let $T : X \to Y$ satisfy $T_{\#}\mu = \nu$. If we consider, for example, the map $Id \times T : X \to X \times Y$, i.e., $x \mapsto (x, T(x))$, then we can define

$$\gamma_T := (Id \times T)_{\#}\mu \in \mathcal{P}(X \times Y).$$

Using Lemma 2.1.5, we have

$$(\pi_X)_{\#}\gamma_T = (\pi_X)_{\#}(Id \times T)_{\#}\mu = (\pi_X \circ (Id \times T))_{\#}\mu = Id_{\#}\mu = \mu$$
$$(\pi_Y)_{\#}\gamma_T = (\pi_Y)_{\#}(Id \times T)_{\#}\mu = (\pi_Y \circ (Id \times T))_{\#}\mu = T_{\#}\mu = \nu.$$

Thus, we have $\gamma_T \in \Gamma(\mu, \nu)$.

Remark 2. If we consider $\mathcal{M}(X)$, the space of all finite signed measures on X, it is clear that $\mathcal{P}(X) \subset \mathcal{M}(X)$. Moreover, by *Riesz representation theorem*, we have that

$$\mathcal{M}(X) \cong C_c(X)^* \cong C_0(X)^*.$$

This is true if we consider $\|\cdot\|_\infty$, remembering that $C_c(X)$ is not complete with respect to this norm, if X is not compact, and $C_0(X) = \overline{C_c(X)}^{\|\cdot\|_\infty}$, if X is at least locally compact and Hausdorff, which is true in our case.

The considerations in Remark 2 allow us to introduce a weak* convergence on $\mathcal{M}(X)$. If we consider a sequence $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{P}(X)$, we have $\mu_n(X) = 1$ for all $n \in \mathbb{N}$ and thus we see that $\{\mu_n\}_{n\in\mathbb{N}}$ is uniformly bounded in $\mathcal{M}(X)$. Using **Banach-Alaoglu-Bourbaki theorem**, we conclude that there exists a subsequence $\{\mu_{n_k}\}_{k\in\mathbb{N}}$ and a measure $\mu \in \mathcal{M}(X)$ such that

$$\mu_{n_k} \overset{*}{\rightharpoonup} \mu,$$

that means

$$\int_X \varphi \, d\mu_{n_k} \to \int_X \varphi \, d\mu \quad \text{for any } \varphi \in C_c(X).$$

Unfortunately, with this convergence, we cannot guarantee that $\mu \in \mathcal{P}(X)$ if $\mu_n \in \mathcal{P}(X)$ and we can see this fact in the following example.

Example 2. *Let $X = \mathbb{R}$ and $\mu_n = \delta_n$ for $n \in \mathbb{N}$. Then, for any $\varphi \in C_c(\mathbb{R})$,*

$$\int_{\mathbb{R}} \varphi \, d\mu_n = \varphi(n) \to 0, \quad \text{when } n \to \infty.$$

Thus $\mu_n \overset{*}{\rightharpoonup} 0 \notin \mathcal{P}(X)$.

To avoid this problem, we introduce a different, but stronger, notion of convergence.

Definition 2.1.7. *Let $C_b(X)$ be the set of continuous bounded functions. We say that μ_n **converges** to μ **narrowly** if*

$$\int_X \varphi \, d\mu_n \to \int_X \varphi \, d\mu \quad \text{for any } \varphi \in C_b(X).$$

We denote this convergence by $\mu_n \rightharpoonup \mu$.

Remark 3. Consider $\{\mu_n\}_n \subset \mathcal{P}(X)$ such that $\mu_n \rightharpoonup \mu$. Then, taking $\varphi \equiv 1$, we have:

$$\mu_n(X) = \int_X d\mu_n \to \int_X d\mu = \mu(X).$$

Thus $\mu \in \mathcal{P}(X)$. This shows that the narrow limits of probability measures are still probability measures.

In Example 2, we also saw that weak* convergence may let some mass of μ_n escape to ∞ or to the boundary of X, if X is open. The introduction of the narrow convergence does not solve this problem because the same sequence in that example does not have a narrow limit. Thus, we need a condition that guarantees the trapping of almost all the mass of μ_n in a fixed compact set.

Definition 2.1.8. *Let $\mathcal{F} \subset \mathcal{P}(X)$ be a family of probability measures. We say that $\mathcal{F}$ is **tight** if for any $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset X$ such that $\mu(X \setminus K_\varepsilon) < \varepsilon$ for any $\mu \in \mathcal{F}$.*

The following theorem states that tightness is a necessary and sufficient condition for compactness with respect to narrow convergence.

Theorem 2.1.9 (**Prokhorov**). *Let $\mathcal{F} \subset \mathcal{P}(X)$. Then $\mathcal{F}$ is **tight** if and only if for every sequence $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ there exist a subsequence $(\mu_{n_k})_{k\in\mathbb{N}}$ and a probability measure $\mu \in \mathcal{P}(X)$ such that*

$$\mu_{n_k} \rightharpoonup \mu.$$

Moreover, it can be shown that if we have the weakly* convergence of a sequence of probability measures and we can guarantee that the limit is a probability measure, then the convergence is narrow. We formulate this result in the following lemma in order to recall it in the sequel.

Lemma 2.1.10. *Consider $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{P}(X)$ such that $\mu_n \overset{*}{\rightharpoonup} \mu$ for some $\mu \in \mathcal{P}(X)$. Then $\mu_n \rightharpoonup \mu$.*

After these preliminary results, we can now explain what optimal transport is. Consider a *source measure* $\mu \in \mathcal{P}(X)$, a *target measure* $\nu \in \mathcal{P}(Y)$ and $c : X \times Y \to [0, +\infty]$ lower semicontinuous, which is called a **cost function**. The original formulation of the optimal transport problem was made by Gaspard Monge in 1781 and consists of finding a transport map T that minimizes the transportation cost among all transport maps:

$$C_M(\mu, \nu) := \inf \left\{ \int_X c(x, T(x))\, d\mu(x) \mid T_{\#}\mu = \nu \right\}. \tag{2.3}$$

This problem can be very difficult to solve because an optimal transport map may not exist. To avoid these difficulties, in 1942 Leonid Kantorovich revisited this problem. The new formulation consisted of minimizing the transport cost among all couplings (or transport plans) between μ and ν:

$$C_K(\mu, \nu) := \inf \left\{ \int_X c(x, y)\, d\gamma(x, y) \mid \gamma \in \Gamma(\mu, \nu) \right\}. \tag{2.4}$$

In Remark 1, we saw that a transport map T always induces a coupling γ_T and it is easy to see that it happens with the same cost. Thus, we are sure that

$$C_M(\mu, \nu) \geq C_K(\mu, \nu).$$

Kantorovich's problem is simpler because it is convex, by the convexity of $\Gamma(\mu, \nu)$ and the linearity in γ of the problem. Moreover, we can prove that we can always find at least a minimum.

Lemma 2.1.11. *The set $\Gamma(\mu, \nu) \subset \mathcal{P}(X \times Y)$ is **tight** and **closed under narrow convergence**.*

Proof. We omit the proof of the tightness and prove that the set is narrowly closed. Take a sequence of couplings $\{\gamma_n\}_{n\in\mathbb{N}} \subset \Gamma(\mu, \nu)$, and assume that $\gamma_n \rightharpoonup \gamma \in \mathcal{P}(X \times Y)$. Then, for any $\varphi \in C_b(X)$ we have

$$\int_X \varphi(x)\, d\mu(x) = \int_{X\times Y} \varphi(x)\, d\gamma_n(x, y) \to \int_{X\times Y} \varphi(x)\, d\gamma(x, y),$$

thus $(\pi_X)_{\#}\gamma = \mu$. If we consider the integral on Y, we also have $(\pi_Y)_{\#}\gamma = \nu$ and therefore $\gamma \in \Gamma(\mu, \nu)$. □

We state the following lemma to prove the next theorem.

Lemma 2.1.12. *Let $\gamma_n \rightharpoonup \gamma$, and let $c : X \times Y \to [0, +\infty]$ be a lower semicontinuous function. Then*

$$\liminf_{n\to\infty} \int_{X\times Y} c\, d\gamma_n \geq \int_{X\times Y} c\, d\gamma.$$

Theorem 2.1.13. *Consider $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a lower semicontinuous cost function $c : X \times Y \to [0, +\infty]$. Then there exists a coupling $\tilde{\gamma}$ that is a minimizer for (2.5).*

Proof. If $\inf_{\gamma\in\Gamma(x,y)} \int_{X\times Y} c\, d\gamma = +\infty$, then every $\gamma \in \Gamma(\mu, \nu)$ is a minimizer. Hence, we can assume that $\alpha := \inf_{\gamma\in\Gamma(x,y)} \int_{X\times Y} c\, d\gamma < +\infty$.

Let $\{\gamma_n\}_{n\in\mathbb{N}} \subset \Gamma(\mu, \nu)$ be a minimizer sequence, i.e.

$$\int_{X\times Y} c\, d\gamma_n \to \alpha \quad \text{as } n \to \infty.$$

By Lemma 2.1.11, it follows that $\{\gamma_n\}_{n\in\mathbb{N}}$ is tight and, by Theorem 2.1.9, there exists a subsequence $\{\gamma_{n_k}\}_{k\in\mathbb{N}}$ and a coupling $\tilde{\gamma}$ such that $\gamma_{n_k} \rightharpoonup \tilde{\gamma}$. Since c is nonnegative and lower semicontinuous, it follows from Lemma 2.1.12 that

$$\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c\, d\gamma = \alpha = \liminf_{k\to\infty} \int_{X\times Y} c\, d\gamma_{n_k} \geq \int_{X\times Y} c\, d\tilde{\gamma} \geq \alpha.$$

The last inequality follows from $\tilde{\gamma} \in \Gamma(\mu, \nu)$, thanks to Lemma 2.1.11. This shows that $\int_{X\times Y} c\, d\tilde{\gamma} = \alpha$, so $\tilde{\gamma}$ is a minimizer for (2.5). □

In general, this minimizer may not be unique and may not be induced by a transport map.

2.2 Wasserstein spaces and JKO Scheme

Now we can introduce Wasserstein distances and focus on the gradient flow formulation in the space of probability measures.

Definition 2.2.1. *Let (X, d) be a locally compact and separable metric space. Given $1 \leq p < \infty$, we can define*

$$\mathcal{P}_p(X) := \left\{ \mu \in \mathcal{P}(X) \mid \int_X d(x, x_0)^p\, d\mu(x) < +\infty \text{ for some } x_0 \in X \right\}. \quad (2.5)$$

*This is called the **set of probability measures with finite p-moment**.*

Remark 4. We can easily see that (2.6) is a good definition, i.e. it is independent of the base point x_0. Given $x_1 \in X$, we have

$$d(x, x_1)^p \leq (d(x, x_0) + d(x_0, x_1))^p \leq 2^{p-1} \left(d(x, x_0)^p + d(x_0, x_1)^p \right).$$

Where we used the triangle inequality and the convexity of the map $x \mapsto d(x, x_1)^p$. Thus, if we consider $\mu \in \mathcal{P}(X)$ such that $\int_X d(x, x_0)^p \, d\mu(x) < +\infty$, we have

$$\int_X d(x, x_1)^p \, d\mu(x) \leq 2^{p-1} \left(\int_X d(x, x_0)^p \, d\mu(x) + C \right) < +\infty.$$

Definition 2.2.2. *Given $\mu, \nu \in \mathcal{P}_p(X)$, we define their **p-Wasserstein distance** as*

$$W_p(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{X \times X} d(x, y)^p \, d\gamma(x, y) \right)^{\frac{1}{p}}. \tag{2.6}$$

Remark 5. *Given $\mu, \nu \in \mathcal{P}(X)$, we have, for all $\gamma \in \Gamma(\mu, \nu)$*

$$\begin{aligned} \int_{X \times X} d(x, y)^p \, d\gamma(x, y) &\leq 2^{p-1} \int_{X \times X} \left(d(x, x_0)^p + d(x_0, y)^p \right) \, d\gamma(x, y) \\ &= 2^{p-1} \left(\int_X d(x, x_0)^p \, d\mu(x) \, + \, \int_X d(y, x_0)^p \, d\nu(y) \right) < +\infty. \end{aligned}$$

Thus, W_p is finite on $\mathcal{P}_p(X) \times \mathcal{P}_p(X)$.

It can be proved that W_p is effectively a distance on $\mathcal{P}_p(X)$, but the proof is not required for our discussion. Moreover, W_p induces a topology that is connected to the weak* topology.

Theorem 2.2.3. *Given $1 \leq p < \infty$ and $x_0 \in X$, consider $\{\mu_n\}_{n \in \mathbb{N}} \subset \mathcal{P}_p(X)$ a sequence of probability measures with finite p-moment and $\mu \in \mathcal{P}_p(X)$. The following statements are equivalent:*

1. *$\mu_n \overset{*}{\rightharpoonup} \mu$ and $\int_X d(x, x_0)^p \, d\mu_n \to \int_X d(x, x_0)^p \, d\mu$.*
2. *$W_p(\mu_n, \mu) \to 0$.*

If X is a compact metric space and $x_0 \in X$ a base point, the function $x \mapsto d(x, x_0)^p$ has compact support (because X is compact and the map is continuous). Thus, if $\mu_n \overset{*}{\rightharpoonup} \mu$, we have $\int_X d(x, x_0)^p \, d\mu_n \to \int_X d(x, x_0)^p \, d\mu$ as a consequence, because we can use $x \mapsto d(x, x_0)^p$ as a test function for the weak* convergence. In this case, we deduce that Wasserstein convergence and weak* convergence are equivalent.

In the previous chapter, we saw that convexity plays an important role in the study of minimization problems. Therefore, it will be very useful to define a notion of convexity in Wasserstein spaces. It is important to note that $\mathcal{P}_p(X)$, endowed with the distance W_p, does not have a vector structure, but only a metric one. Thus, convexity must be understood in terms of the geodesics of space.

Remark 6 (**Construction of geodesics**)**.** Using a generic metric space X, an explicit form for geodesics is not given, hence we consider $X = \mathbb{R}^d$. Now, fix $\mu_0, \mu_1 \in \mathcal{P}_p(\mathbb{R}^d)$ and let $\gamma \in \Gamma(\mu_0, \mu_1)$ be an optimal coupling for W_p. The goal is to construct a curve of measures and to give a condition to establish when it

can be called a **geodesic**. Fixed $t \in [0,1]$, consider a map $\pi_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ that realizes the convex combination of entries, i.e. $\pi_t(x,y) := (1-t)x + ty$. Therefore, we have

$$\begin{cases} (\pi_0)_{\#}\gamma = \mu_0, \\ (\pi_1)_{\#}\gamma = \mu_1. \end{cases}$$

Define $\mu_t := (\pi_t)_{\#}\gamma$ and let $\gamma_{s,t} := (\pi_s, \pi_t)_{\#}\gamma \in \Gamma(\mu_s, \mu_t)$ for all $s,t \in [0,1]$. Then

$$\begin{aligned} W_p(\mu_s,\mu_t) &\leq \left(\int_{X\times X} |z-z'|^p \, d\gamma_{s,t}(z,z') \right)^{\frac{1}{p}} \\ &= \left(\int_{X\times X} |\pi_s(x,y) - \pi_t(x,y)|^p \, d\gamma(x,y) \right)^{\frac{1}{p}} \\ &= |t-s| \left(\int_{X\times X} |x-y|^p \, d\gamma(x,y) \right)^{\frac{1}{p}} = |x-y| \, W_p(\mu_0,\mu_1). \end{aligned}$$

Now we consider $0 \leq s \leq t \leq 1$ and apply the above bound on the intervals $[0,s], [s,t]$, and $[t,1]$, obtaining

$$\begin{aligned} W_p(\mu_0,\mu_s) + W_p(\mu_s,\mu_t) + W_p(\mu_t,\mu_1) &\leq (s + (t-s) + 1 - t) \, W_p(\mu_0,\mu_1) \\ &= W_p(\mu_0,\mu_1). \end{aligned}$$

Since we have the converse inequality, by the triangle inequality, we can conclude that

$$W_p(\mu_s,\mu_t) = |t-s| \, W_p(\mu_0,\mu_1) \quad \forall \, 0 \leq s,t \leq 1. \tag{2.7}$$

Definition 2.2.4. *A curve of measures $(\mu_t)_{t\in[0.1]} \subset \mathcal{P}_p(\mathbb{R}^d)$ that satisfies (2.7) is called a* ***constant speed geodesic****.*

From the above discussion, we can also conclude that any optimal coupling γ induces a geodesic $(\mu_t)_{t\in[0,1]} := ((\pi_t)_{\#}\gamma)_{t\in[0,1]}$. One can also show, when $X = \mathbb{R}^d$, that any geodesic is induced by an optimal coupling. Moreover, when the coupling is induced by a transport map, i.e. $\gamma = (Id \times T)_{\#}\mu$, the geodesic μ_t takes the form

$$\mu_t = (\pi_t)_{\#}(Id \times T)_{\#}\mu = (\pi_t \circ (Id \times T))_{\#}\, \mu = (T_t)_{\#}\mu,$$

where $T_t(x) := (1-t)x + tT(x)$ is the linear interpolation between the identity and the transport map T.

In the previous chapter, we also saw the weak formulation of gradient flows in $\mathbb{R}^d$ and now we would see an analogue in Wasserstein spaces. In this case, it is useful to recall a result related to gradient flows in Hilbert spaces.

Remark 7. Consider the space $L^2(\mathbb{R}^d)$ and the **Dirichlet energy functional**

$$\Phi(u) := \begin{cases} \dfrac{1}{2} \displaystyle\int_{\mathbb{R}^d} |\nabla u|^2 \, dx & \text{if } u \in H^1(\mathbb{R}^d), \\ +\infty & \text{otherwise.} \end{cases}$$

Then, for a.e. $t > 0$ and for a.e. $x \in \mathbb{R}^d$, we have

$$\partial_t u(t) \in -\partial\Phi(u(t)) \quad \Longleftrightarrow \quad \partial_t u(t,x) = \Delta u(t,x). \tag{2.8}$$

This means that the **gradient flow** of Φ with respect to the L^2-norm is the **heat equation**.

The above result can be obtained by proving the analog of Theorem 1.3.2 in the context of infinite-dimensional Hilbert spaces. Then, the idea is to consider the Minimizing Movement Scheme (1.35) using the L^2-norm, with the optimality condition

$$u_\tau^k \in \operatorname*{argmin}_{v \in L^2} \left\{ \Phi(v) + \frac{1}{2\tau} \left\| v - u_\tau^{k-1} \right\|_{L^2}^2 \right\} \iff -\frac{u_\tau^k - u_\tau^{k-1}}{\tau} \in \partial\Phi(u_\tau^k). \tag{2.9}$$

In [9], the authors developed a new way to construct solutions of the heat equation as gradient flows, replacing the L^2-norm with the 2-Wasserstein distance W_2, and the Dirichlet energy functional with the **entropy functional** $\int \rho \log(\rho)$. More precisely, they introduced the so-called **JKO Scheme**

$$\rho_\tau^k \in \operatorname*{argmin}_{\rho \in \mathcal{P}(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \rho \, \log(\rho) \, dx + \frac{1}{2\tau} W_2^2(\rho, \rho_\tau^{k-1}) \right\}. \tag{2.10}$$

In this context, we denote by ρ a probability density and adopt the convention that $\int_{\mathbb{R}^d} \rho \log(\rho)\, dx := +\infty$ if $\rho \in \mathcal{P}(\mathbb{R}^d)$ is not absolutely continuous with respect to the Lebesgue measure. Moreover, given two probability densities ρ and $\tilde{\rho}$, we identify them with the probability measures $\rho\, dx$ and $\tilde{\rho}\, dx$, hence

$$W_2(\rho, \tilde{\rho}) := W_2(\rho\, dx, \tilde{\rho}\, dx) = \inf_{\gamma \in \Gamma(\rho\, dx, \tilde{\rho}\, dx)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \, d\gamma \right)^{\frac{1}{2}}.$$

In [9], the authors also applied the W_2-interpretation of the heat equation to the **Fokker-Planck Equation**, which cannot be possible with the L^2-interpretation. To end this section, we will prove the existence of a solution for (2.11) and then state the associated optimality condition with the link to the heat equation. The authors in [9] worked with probability densities in $\mathbb{R}^d$, but we will use the setting in [8], considering probability densities in a convex bounded domain $\Omega \subset \mathbb{R}^d$.

Lemma 2.2.5. *Consider ρ_0 a probability density in Ω ($\rho_0 \in \mathcal{P}(\Omega)$, using the identification between ρ_0 and $\rho\, dx$) with* finite entropy, *i.e.*

$$\int_\Omega \rho_0 \log(\rho_0)\, dx < +\infty. \tag{2.11}$$

Fix $\tau > 0$, $\rho_\tau^0 := \rho_0$, and $\rho_\tau^{k-1} \in \mathcal{P}(\Omega)$. Then, for any $k \geq 1$ there exists ρ_τ^k that satisfies (2.11).

Proof. Fix $k \geq 1$ and take $\{\rho_n\}_{n\in\mathbb{N}} \subset \mathcal{P}(\Omega)$ a minimizing sequence, that is

$$\frac{W_2^2(\rho_n, \rho_\tau^{k-1})}{2\tau} + \int_\Omega \rho_n \log(\rho_n)\, dx \to \inf_{\rho\in\mathcal{P}(\Omega)} \left\{ \frac{W_2^2(\rho, \rho_\tau^{k-1})}{2\tau} + \int_\Omega \rho \log(\rho)\, dx \right\}.$$

For all $N \in \mathbb{N}$, the sequence $\{\rho_n \wedge N\}_{n\in\mathbb{N}}$ is bounded in $L^\infty(\Omega)$. In particular, fixed $N \in \mathbb{N}$, we have $\|\rho_n \wedge N\|_{L^\infty(\Omega)} \leq N$. Hence, by the Banach-Alaoglu theorem, the sequence is weakly* compact in $L^\infty(\Omega)$. Thus, we can use a diagonal argument to find a subsequence n_l independent of N and $\rho_N \in L^\infty(\Omega)$ such that

$$\rho_{n_l} \wedge N \overset{*}{\rightharpoonup} \rho_N \quad \text{in } L^\infty(\Omega). \tag{2.12}$$

Moreover, since $x - N \leq \frac{x\log(x)}{log(N)}$ for all $x > 0$, we have

$$\begin{aligned}\int_\Omega (\rho_{n_l} - \rho_{n_l} \wedge N)\, dx &= \int_{\Omega\cap\{\rho_{n_l}\geq N\}} (\rho_{n_l} - N)\, dx \\ &\leq \frac{1}{log(N)} \int_{\Omega\cap\{\rho_{n_l}\geq N\}} \rho_{n_l} \log(\rho_{n_l})\, dx \\ &\leq \frac{1}{log(N)} \int_{\Omega\cap\{\rho_{n_l}\geq N\}} (\rho_{n_l} \log(\rho_{n_l}) + 1)\, dx.\end{aligned}$$

Then, since $x\, log(x) + 1 \geq 0$ for all $x > 0$, we have

$$\begin{aligned}\frac{1}{\log(N)} \int_{\Omega\cap\{\rho_{n_l}\geq N\}} (\rho_{n_l} \log(\rho_{n_l}) + 1)\, dx &\leq \frac{1}{\log(N)} \int_\Omega (\rho_{n_l} \log(\rho_{n_l}) + 1)\, dx \\ &\leq \frac{C}{\log(N)},\end{aligned}$$

where, in the last inequality, we used the fact that ρ_{n_l} is a minimizing sequence and Ω is bounded. Hence, we proved the following bound

$$\|\rho_{n_l} - \rho_{n_l} \wedge N\|_{L^1(\Omega)} \leq \frac{C}{\log(N)}. \tag{2.13}$$

Now, setting $\rho_\infty := \sup_N \rho_N$, we have $\rho_\infty \in L^1(\Omega)$, because $\|\rho_N\|_{L^1(\Omega)} \leq 1$ for all $N \in \mathbb{N}$. Then, by monotone convergence theorem, we have

$$\rho_N \overset{L^1}{\to} \rho_\infty. \tag{2.14}$$

Combining (2.13) and (2.15), we can find a sequence of indices $\{n_{l_N}\}_{N\in\mathbb{N}}$, with $n_{l_N} \to \infty$, such that $\rho_{n_{l_N}} \wedge N \rightharpoonup \rho_\infty$ in $L^1(\Omega)$ as $N \to \infty$. Then, using (2.14), we conclude that $\rho_{n_{lN}} \rightharpoonup \rho_\infty$ in $L^1(\Omega)$. Now, we want to prove that ρ_∞ is still a probability density, since, taking this limit, some mass may have "escaped" from Ω. For all $\varepsilon > 0$, set $A_\varepsilon := \{x \in \Omega \,|\, dist(x, \partial\Omega) < \varepsilon\}$. Then, we have

$|A_\varepsilon| \leq C\varepsilon$ and, for $L := \frac{1}{\varepsilon|\log(\varepsilon)|}$, it follows that

$$\begin{aligned}\int_{A_\varepsilon} \rho_n &\leq \int_{A_\varepsilon \cap \{\rho_n \leq L\}} \rho_n + \int_{A_\varepsilon \cap \{\rho_n \geq L\}} \rho_n \frac{\log(\rho_n)}{\log(L)} \\ &\leq L|A_\varepsilon| + \frac{C}{\log(L)} \leq C\left(\varepsilon L + \frac{1}{\log(L)}\right) \leq \frac{C}{|\log(\varepsilon)|} \quad \forall n \in \mathbb{N}.\end{aligned}$$

In particular, we have

$$\int_{\Omega \setminus A_\varepsilon} \rho_{n_{l_N}} \geq 1 - \frac{C}{|log(\varepsilon)|}$$

and thus

$$\int_{\Omega \setminus A_\varepsilon} \rho_\infty \geq 1 - \frac{C}{|log(\varepsilon)|}.$$

Taking $\varepsilon \to 0$, we conclude that ρ_∞ is still a probability density. By Lemma 2.1.10, we have $\rho_{n_{l_N}} \rightharpoonup \rho_\infty$ and, thanks to Theorem 2.1.9, the family $\{\rho_{n_{l_N}}\}_{n\in\mathbb{N}}$ is tight. Now we want to show the lower semicontinuity of the functional defined in (2.11). We start observing that, for all $t \geq 0$, we have

$$t\log(t) \geq t\,(\alpha + 1) - e^\alpha, \quad \forall \alpha \in \mathbb{R}, \quad \text{with equality for } \alpha = \log(t). \tag{2.15}$$

Thus, choosing $t = \rho(x)$ and $\alpha = \phi(x)$, we have, for any $\phi \in C(\Omega)$

$$\begin{aligned}\liminf_{N\to\infty} \int_\Omega \rho_{n_{l_N}}\, log(\rho_{n_{l_N}}) &\geq \liminf_{N\to\infty} \int_\Omega (\rho_{n_{l_N}}(x)(\phi(x)+1) - e^{\phi(x)})\,dx \\ &= \int_\Omega (\rho_\infty(x)(\phi(x)+1) - e^{\phi(x)})\,dx.\end{aligned} \tag{2.16}$$

Where the last equality follows from the narrow convergence of $\rho_{n_{l_N}}$ to ρ_∞. Now, taking a sequence $\{\phi_j\}_{j\in\mathbb{N}}$ converging to $\log(\rho_\infty)$ when $j \to \infty$, we obtain

$$\int_\Omega \rho_\infty \log(\rho_\infty) \leq \liminf_{N\to\infty} \int_\Omega \rho_{n_{l_N}} \log(\rho_{n_{l_N}}). \tag{2.17}$$

Consider $\gamma_N \in \Gamma(\rho_{n_{l_N}}, \rho_\tau^{k-1})$. Then, combining Lemma 2.1.11 with the tightness of $\{\rho_{n_{l_N}}\}_{n\in\mathbb{N}}$, we see that γ_N is also tight. Hence, being Γ narrowly closed, we have, up to a subsequence, $\gamma_N \rightharpoonup \gamma_\infty$ when $N \to \infty$, with

$$(\pi_1)_\#\gamma_\infty = \rho_\infty, \quad (\pi_2)_\#\gamma_\infty = \rho_\tau^{k-1}.$$

Then, it follows that $\gamma_\infty \in \Gamma(\rho_\infty, \rho_\tau^{k-1})$. Moreover, the function $(x, y) \mapsto |x-y|^2$ is continuous and bounded on $\Omega \times \Omega$, so we have

$$W_2^2(\rho_{n_{l_N}}, \rho_\tau^{k-1}) = \int_{\Omega\times\Omega} |x-y|^2\, d\gamma_N \to \int_{\Omega\times\Omega} |x-y|^2\, d\gamma_\infty \geq W_2^2(\rho_\infty, \rho_\tau^{k-1}). \tag{2.18}$$

Thus, combining (2.18) and (2.19), we get the lower semicontinuity of the functional. Since ρ_n was a minimizing sequence, we conclude that ρ_∞ is a minimizer. The proof ends with the definition of $\rho_\tau^k := \rho_\infty$. □

The minimality equation satisfied by ρ_τ^k is defined in the following lemma.

Lemma 2.2.6. *For any vector field $\xi \in C^\infty(\Omega, \mathbb{R}^d)$ tangent to the boundary of Ω, we have*

$$\int_\Omega \rho_\tau^k \, (\nabla \cdot \xi) \, dx = \frac{1}{\tau} \int_\Omega \langle \xi \circ T_k, T_k - x \rangle \rho_\tau^{k-1} \, dx,$$

where $T_k : \Omega \to \Omega$ is the optimal map from ρ_τ^{k-1} to ρ_τ^k.

Combining the two previous results, one can prove the following theorem.

Theorem 2.2.7. *Given $\tau > 0$, let $\rho_\tau : [0, +\infty) \to \mathcal{P}(\Omega)$ be the curve of probability measures given by*

$$\begin{cases} \rho_0 & \text{for } t = 0, \\ \rho_\tau^k & \text{for } t \in ((k-1)\tau, k\tau],\ k \geq 1. \end{cases}$$

Then there exists a curve of probability measures $\rho \in L^1_{loc}([0, +\infty) \times \Omega)$ such that, up to a subsequence in τ,

$$\rho_\tau \rightharpoonup \rho \quad \text{weakly in } L^1_{loc}([0, +\infty) \times \Omega).$$

*Moreover, ρ satisfies the **heat equation** (in distributional sense) with initial data ρ_0 and zero Neumann boundary conditions.*

2.3 Gradient flows in Wasserstein spaces

After the discussion in the previous section, the question is : can we interpret the heat equation as the gradient flow of the entropy with respect to the W_2 distance? In this section, we will introduce a differential structure on $\mathcal{P}_2(\Omega)$ that allows us to realize this interpretation. Moreover, one can also use this formulation for the interpretation of other evolution equations as gradient flows.

For the following discussion, consider a convex set $\Omega \subset \mathbb{R}^d$, $\tilde{\rho}_0 \in \mathcal{P}_2(\Omega)$ and let $v : [0, T] \times \Omega \to \mathbb{R}^d$ be a smooth bounded vector field tangent to $\partial\Omega$. Moreover, we denote the flow of v by $X(t, x)$, which is a function such that

$$\begin{cases} \dot{X}(t, x) = v(t, X(t, x)), \\ X(0, x) = x, \end{cases}$$

and set $\rho_t(\cdot) := (X(t))_\# \tilde{\rho}_0(\cdot)$.

Remark 8. With the assumptions made above, we have $\rho_t \in \mathcal{P}_2(\Omega)$. Indeed, since v is tangent to $\partial\Omega$, the flow remains inside Ω. To see this fact, consider the function $d(y) := dist(y, \Omega^c)$. Since Ω is smooth, d is also smooth in a neighborhood of any $y \in \partial\Omega$ and $\nabla d(y) = \nu(y)$ for all $y \in \partial\Omega$. Starting from $x \in \Omega$, if the flow of v reaches $\partial\Omega$, i.e. $X(t, x) \in \partial\Omega$ for some $t > 0$, we have

$$\frac{d}{dt} d(X(t, x)) = \dot{X}(t, x) \cdot \nabla d(X(t, x)) = v(t, X(t, x)) \cdot \nu(X(t, x)) = 0.$$

Hence, once the flow reached $\partial\Omega$, it cannot leave Ω. Moreover, since $v_t(\cdot) := v(t,\cdot)$ is bounded, we have $|X(t,x) - x| \leq Ct$, hence

$$\begin{aligned}\int_\Omega |x|^2 \rho_t(x)\,dx &= \int_\Omega |X(t,x)|^2 \tilde{\rho}_0(x)\,dx \\ &\leq 2\int_\Omega \left(|x|^2 + (Ct)^2\right)\tilde{\rho}_0(x)\,dx.\end{aligned}$$

Having $\tilde{\rho}_0 \in \mathcal{P}_2(\Omega)$, it follows that $\rho_t \in \mathcal{P}_2(\Omega)$.

Lemma 2.3.1. *With the above assumptions, the* ***continuity equation***

$$\partial_t \rho_t + \operatorname{div}(v_t \rho_t) = 0 \tag{2.19}$$

holds in the distributional sense.

Proof. Let $\varphi \in C_c^\infty(\Omega)$ and consider $t \mapsto \int_\Omega \rho_t(x)\varphi(x)\,dx$. Then, using the definitions of X and ρ_t, we get

$$\begin{aligned}\int_\Omega \partial_t \rho_t(x)\varphi(x)\,dx &= \frac{d}{dt}\int_\Omega \rho_t(x)\varphi(x)\,dx = \frac{d}{dt}\int_\Omega \varphi(X(t,x))\tilde{\rho}_0(x)\,dx \\ &= \int_\Omega \nabla\varphi(X(t,x)) \cdot \dot{X}(t,x)\tilde{\rho}_0(x)\,dx \\ &= \int_\Omega \nabla\varphi(X(t,x)) \cdot v_t(X(t,x))\tilde{\rho}_0(x)\,dx \\ &= \int_\Omega \nabla\varphi(x) \cdot v_t(x)\rho_t(x)\,dx = -\int_\Omega \varphi(x)\operatorname{div}(v_t\rho_t)\,dx.\end{aligned}$$

□

Definition 2.3.2. *Given a pair (ρ_t, v_t) solving the continuity equation (2.20), with $v_t \cdot \nu\,|_{\partial\Omega} = 0$ (namely, v_t is tangent to $\partial\Omega$), we define its* ***action*** *as*

$$S[\rho_t, v_t] := \int_0^1 \int_\Omega |v_t(x)|^2 \rho_t(x)\,dx\,dt.$$

Now, we state the following theorem that links the continuity equation and the W_2 distance.

Theorem 2.3.3 (**Benamou-Brenier formula**). *Given $\tilde{\rho}_0$, $\tilde{\rho}_1 \in \mathcal{P}_2(\Omega)$, it holds that*

$$W_2^2(\tilde{\rho}_0, \tilde{\rho}_1) = \inf_{\rho_t, v_t} \left\{ S[\rho_t, v_t] \,|\, \rho_0 = \tilde{\rho}_0,\ \rho_1 = \tilde{\rho}_1,\ \partial_t\rho_t + div(v_t\rho_t) = 0,\ v \cdot \nu\,|_{\partial\Omega} = 0 \right\}.$$

The next step is the introduction of a "norm" in the space $\mathcal{P}_2(\Omega)$. The idea is to interpret $(\mathcal{P}_2(\Omega), W_2)$ as an **infinite-dimensional Riemannian manifold**. Under this interpretation, the space can be treated locally as a Hilbert space.

Remark 9 (**Wasserstein norm of $\partial_t\rho_t$**). We start manipulating the Benamou-

Brenier formula

$$\begin{aligned}
W_2^2(\tilde\rho_0,\tilde\rho_1) &= \inf_{\rho_t,v_t}\left\{\int_0^1\left(\int_\Omega |v_t|^2\rho_t\,dx\right)dt \;\middle|\; \partial_t\rho_t+\operatorname{div}(v_t\rho_t)=0,\ v_t\cdot\nu|_{\partial\Omega}=0,\right.\\
&\qquad\qquad \left.\rho_0=\tilde\rho_0,\ \rho_1=\tilde\rho_1\right\}\\
&= \inf_{\rho_t}\left\{\inf_{v_t}\int_0^1\left(\int_\Omega |v_t|^2\rho_t\,dx\right)dt \;\middle|\; \partial_t\rho_t+\operatorname{div}(v_t\rho_t)=0,\ v_t\cdot\nu|_{\partial\Omega}=0,\right.\\
&\qquad\qquad \left.\rho_0=\tilde\rho_0,\ \rho_1=\tilde\rho_1\right\}\\
&= \inf_{\rho_t}\left\{\int_0^1\inf_{v_t}\left\{\int_\Omega |v_t|^2\rho_t\,dx \;\middle|\; \operatorname{div}(v_t\rho_t)=-\partial_t\rho_t,\ v_t\cdot\nu|_{\partial\Omega}=0\right\}dt,\right.\\
&\qquad\qquad \left.\rho_0=\tilde\rho_0,\ \rho_1=\tilde\rho_1\right\}.
\end{aligned}$$

Now, it is natural to define the **Wasserstein norm** of $\partial_t\rho_t$ as

$$\|\partial_t\rho_t\|_{\rho_t}^2 := \inf_{v_t}\left\{\int_\Omega |v_t|^2\rho_t\,dx \;\middle|\; \operatorname{div}(v_t\rho_t)=-\partial_t\rho_t,\ v_t\cdot\nu|_{\partial\Omega}=0\right\}. \tag{2.20}$$

Hence, the continuity equation gives, at each time t, a constraint on $\operatorname{div}(v_t\rho_t)$, and we obtain

$$W_2^2(\tilde\rho_0,\tilde\rho_1) = \inf_{\rho_t}\left\{\int_0^1\|\partial_t\rho_t\|_{\rho_t}^2\,dt \;\middle|\; \rho_0=\tilde\rho_0,\ \rho_1=\tilde\rho_1\right\}. \tag{2.21}$$

Now, we want to find a better formula for the Wasserstein norm of $\partial_t\rho_t$. Thus, given ρ_t and $\partial_t\rho_t$, let v_t be a minimizer, and consider a vector field w such that $\operatorname{div}(w)\equiv 0$ and $w\cdot\nu=0$. Then, for every $\varepsilon>0$, we have

$$\operatorname{div}\left(\left(v_t+\varepsilon\frac{w}{\rho_t}\right)\rho_t\right) = -\partial_t\rho_t.$$

Hence $v_t+\varepsilon\frac{w}{\rho_t}$ is an admissible vector field in the minimization problem (2.21), and so by minimality of v_t we get

$$\begin{aligned}
\int_\Omega |v_t|^2\,\rho_t\,dx &\le \int_\Omega\left|v_t+\varepsilon\frac{w}{\rho_t}\right|^2\rho_t\,dx\\
&= \int_\Omega |v_t|^2\,\rho_t\,dx + 2\varepsilon\int_\Omega\langle v_t,w\rangle\,dx + \varepsilon^2\int_\Omega\frac{|w|^2}{\rho_t}\,dx.
\end{aligned}$$

Dividing by ε and taking $\varepsilon\to 0$, we have

$$\int_\Omega\langle v_t,w\rangle\ge 0.$$

Choosing $-w$, we obtain the opposite inequality and so

$$\int_\Omega \langle v_t, w\rangle = 0.$$

By the Helmholtz decomposition, we get

$$v_t \in \{w \,|\, \operatorname{div}(w) = 0\}^\perp = \{\nabla\psi \,|\, \psi : \Omega \to \mathbb{R}\}.$$

Therefore, there exists a function ψ_t such that $v_t = \nabla\psi$. Moreover, since $\operatorname{div}(v_t\rho_t) = -\partial_t\rho_t$ and $v_t \cdot \nu\,|_{\partial\Omega} = 0$, then ψ_t is a solution of

$$\begin{cases} \operatorname{div}(\rho_t\nabla\psi_t) = -\partial_t\rho_t & \text{in } \Omega, \\ \dfrac{\partial\psi_t}{\partial\nu} = 0 & \text{on } \partial\Omega. \end{cases} \tag{2.22}$$

Note that if ρ_t is positive and smooth, then (2.23) is a uniformly elliptic equation with Neumann boundary conditions for ψ_t, and the solution ψ_t is unique up to a constant. Hence, we can define

$$\|\partial_t\rho_t\|^2_{\rho_t} = \int_\Omega |\nabla\psi_t|^2 \, \rho_t \, dx.$$

Moreover, given $\rho \in \mathcal{P}_2(\Omega)$ and $h : \Omega \to \mathbb{R}$ such that $\int_\Omega h = 0$, we can define

$$\|h\|^2_\rho := \int_\Omega |\nabla\psi|^2 \, \rho \, dx,$$

where

$$\begin{cases} \operatorname{div}(\rho\nabla\psi) = -h & \text{in } \Omega, \\ \dfrac{\partial\psi}{\partial\nu} = 0 & \text{on } \partial\Omega. \end{cases}$$

In conclusion, we got a nice expression for the Wasserstein norm of the derivative of a curve of probability measures.

Now, we can canonically construct the Wasserstein scalar product.

Definition 2.3.4. *Given two functions $h_1, h_2 : \Omega \to \mathbb{R}$ with $\int_\Omega h_1 = \int_\Omega h_2 = 0$, we define their **Wasserstein scalar product** at ρ as*

$$\langle h_1, h_2\rangle_\rho := \int_\Omega \nabla\psi_1 \cdot \nabla\psi_2 \, \rho \, dx, \tag{2.23}$$

where, for $i = 1, 2$, it holds that

$$\begin{cases} \operatorname{div}(\rho\nabla\psi_i) = -h_i & \text{in } \Omega, \\ \dfrac{\partial\psi_i}{\partial\nu} = 0 & \text{on } \partial\Omega. \end{cases}$$

Finally, we can define the gradient flow of a functional in $(\mathcal{P}_2(\Omega), W_2)$. For our purposes, we will only consider probability measures which are absolutely

continuous with respect to the Lebesgue measure, i.e. $\rho = \rho(x)\,dx$, and functionals $\mathcal{F} : \mathcal{P}_2(\Omega) \to \mathbb{R}$ which have an integral representation.

Definition 2.3.5. *Given a functional $\mathcal{F} : \mathcal{P}_2(\Omega) \to \mathbb{R}$, its **gradient** with respect to the **Wasserstein scalar product** at $\tilde{\rho} \in \mathcal{P}_2(\Omega)$ is the unique function $\mathrm{grad}_{W_2} \mathcal{F}[\tilde{\rho}]$ (if it exists) such that*

$$\left\langle \mathrm{grad}_{W_2} \mathcal{F}[\tilde{\rho}], \left.\frac{\partial \rho_\varepsilon}{\partial \varepsilon}\right|_{\varepsilon=0} \right\rangle_{\tilde{\rho}} = \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \mathcal{F}[\rho_\varepsilon] \tag{2.24}$$

for any smooth curve $\rho_\varepsilon : (-\varepsilon_0, \varepsilon_0) \to \mathcal{P}_2(\Omega)$ with $\rho_0 = \tilde{\rho}$.

Given a functional $\mathcal{F} : \mathcal{P}_2(\Omega) \to \mathbb{R}$ and $\tilde{\rho} \in \mathcal{P}_2(\Omega)$ with the assumptions made above, we can obtain a more explicit formula for the Wasserstein gradient of $\mathcal{F}$ in $\tilde{\rho}$. Let us denote by $\frac{\delta \mathcal{F}[\tilde{\rho}]}{\delta \rho}$ the first L^2-variation of $\mathcal{F}$ (if it exists), which is the function in $L^2(\Omega)$ such that

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \mathcal{F}[\rho_\varepsilon] = \int_\Omega \frac{\delta \mathcal{F}[\tilde{\rho}]}{\delta \rho}(x) \left.\frac{\partial \rho_\varepsilon(x)}{\partial \varepsilon}\right|_{\varepsilon=0} dx$$

for any smooth curve $\rho_\varepsilon : (-\varepsilon_0, \varepsilon_0) \to \mathcal{P}_2(\Omega)$ with $\rho_0 = \tilde{\rho}$. Then, by (2.25), we get

$$\left\langle \mathrm{grad}_{W_2} \mathcal{F}[\tilde{\rho}], \left.\frac{\partial \rho_\varepsilon}{\partial \varepsilon}\right|_{\varepsilon=0} \right\rangle_{\tilde{\rho}} = \int_\Omega \frac{\delta \mathcal{F}[\tilde{\rho}]}{\delta \rho} \left.\frac{\partial \rho_\varepsilon}{\partial \varepsilon}\right|_{\varepsilon=0} dx.$$

Hence, denoting by ψ the solution of $\mathrm{div}(\nabla \psi\, \tilde{\rho}) = -\frac{\partial \rho_\varepsilon}{\partial \varepsilon}|_{\varepsilon=0}$ with zero Neumann boundary conditions, we have

$$\begin{aligned} \int_\Omega \frac{\delta \mathcal{F}[\tilde{\rho}]}{\delta \rho} \left.\frac{\partial \rho_\varepsilon}{\partial \varepsilon}\right|_{\varepsilon=0} dx &= -\int_\Omega \frac{\delta \mathcal{F}[\tilde{\rho}]}{\delta \rho} \,\mathrm{div}(\nabla \psi\, \tilde{\rho})\, dx \\ &= \int_\Omega \nabla \frac{\delta \mathcal{F}[\tilde{\rho}]}{\delta \rho} \cdot \nabla \psi\, \tilde{\rho}\, dx. \end{aligned}$$

In conclusion, by definition of the Wasserstein scalar product, we have

$$\mathrm{grad}_{W_2} \mathcal{F}[\tilde{\rho}] = -\,\mathrm{div}\left(\tilde{\rho}\, \nabla \left(\frac{\delta \mathcal{F}[\tilde{\rho}]}{\delta \rho}\right)\right). \tag{2.25}$$

Example 3. *If $\mathcal{F}[\rho] = \int_\Omega U(\rho(x))\,dx$ with $U : \mathbb{R} \to \mathbb{R}$, then for any smooth variation $\varepsilon \mapsto \rho_\varepsilon$ such that $\rho_0 = \tilde{\rho}$ we have*

$$\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \int_\Omega U(\rho_\varepsilon(x))\, dx = \int_\Omega U'(\tilde{\rho}(x)) \left.\frac{\partial \rho_\varepsilon(x)}{\partial \varepsilon}\right|_{\varepsilon=0} dx.$$

Hence, the first L^2-variation of $\mathcal{F}[\rho]$ at $\tilde{\rho} \in \mathcal{P}_2(\Omega)$ is given by

$$\frac{\delta \mathcal{F}[\tilde{\rho}]}{\delta \rho}(x) = U'(\tilde{\rho}(x)).$$

Using (2.26), we obtain the Wasserstein gradient of $\mathcal{F}$

$$\operatorname{grad}_{W_2} \mathcal{F}[\tilde{\rho}] = -\operatorname{div}\left(\tilde{\rho}\,\nabla\left[U'(\tilde{\rho})\right]\right) = -\operatorname{div}\left(\tilde{\rho}\,\mathrm{U"}(\tilde{\rho})\,\nabla\tilde{\rho}\right). \tag{2.26}$$

If we choose $U(s) = s\,log(s)$ *(thus* $\mathcal{F}$ *is the entropy) we get*

$$\operatorname{grad}_{W_2} \mathcal{F}[\tilde{\rho}] = -\operatorname{div}\left(\tilde{\rho}\frac{d}{d\tilde{\rho}}\left[\log(\tilde{\rho}) + \tilde{\rho}\,\frac{1}{\tilde{\rho}}\right]\nabla\tilde{\rho}\right) = -\Delta\tilde{\rho}.$$

Now we can give the definition of gradient flow in the 2-Wasserstein space.

Definition 2.3.6. *Given a functional* $\mathcal{F} : \mathcal{P}_2(\Omega) \to \mathbb{R}$*, a curve of probability measures* $\rho : [0,T) \to \mathcal{P}_2(\Omega)$ *is a* ***gradient flow*** *of* $\mathcal{F}$ *with respect to* W_2 *and with starting point* $\tilde{\rho}_0$ *if*

$$\begin{cases} \partial_t \rho_t = -\operatorname{grad}_{W_2} \mathcal{F}[\rho_t] \\ \rho_0 = \tilde{\rho}_0. \end{cases} \tag{2.27}$$

Remark 10. If we choose the entropy functional $\mathcal{F}[\rho] = \int_\Omega \rho\,log(\rho)\,dx$, we get that its Wasserstein gradient flow is the heat equation

$$\partial_t \rho_t = -\operatorname{grad}_{W_2} \mathcal{F}[\rho_t] = \Delta\rho_t,$$

as expected from what we proved in the previous section.

To conclude this chapter, we want to briefly discuss the convexity properties of functionals in the Wasserstein space.

Definition 2.3.7. *A functional* $\mathcal{F} : \mathcal{P}_2(\Omega) \to \mathbb{R}$ *is* W_2-***convex****, or* ***displacement convex****, if the function*

$$[0,1] \ni t \mapsto \mathcal{F}[\rho_t]$$

is convex for any W_2*-geodesic* $\rho : [0,1] \to \mathcal{P}_2(\Omega)$*, using Definition 2.2.4.*

Remark 11. An alternative definition of displacement convexity can be given requiring that, for any two measures $\mu, \nu \in \mathcal{P}_2(\Omega)$, the functional is convex along at least one geodesic between μ and ν. This is called **weak displacement convexity**, and, in general, it is not equivalent to Definition 2.3.7. On the other hand, if the functional can be written as

$$\mathcal{F}[\rho] = \int_\Omega U(\rho(x))\,dx,$$

for a convex Lipschitz-continuous function $U : [0,+\infty) \to \mathbb{R}$ with $U(0) = 0$, then the two definitions are equivalent. Note that this is true for the entropy functional.

With extra work, the following result can be proven.

Proposition 2.3.8. *Consider a function* $U : [0,+\infty) \to \mathbb{R}$ *such that*

$$(0,+\infty)s \ni \mapsto U\left(\frac{1}{s^d}\right)s^d \quad \textit{is convex and nonincreasing.}$$

Then the functional $\mathcal{F}[\rho] := \int_\Omega U(\rho)\,dx$ *is displacement convex.*

u

Chapter 3

Neural network training

In this chapter, our focus will be on the mathematical formulation of neural network training. In particular, we are interested in the interpretation of neural network training as a Wasserstein gradient flow. Generally speaking, an artificial neural network is a supervised machine learning algorithm that emulates biological neural processes. From a theoretical point of view, it can be understood as a parametrized function whose training dynamics arise from the minimization of a suitable functional.

3.1 Classical formulation

First, consider a set of parameters $u := (w, \theta, b) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+2}$ and define $\varphi_u : \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathbb{R}^d \ni x \mapsto \varphi_u(x) = w\,\sigma(\theta \cdot x + b), \tag{3.1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is called **activation function** and there are many different suitable options for its choice, depending on the problem we want to solve. Usually, σ is a smooth non-linear function like $\sigma(x) = \tanh(x)$, although one of the most used activation functions in applications is $\sigma(x) = \text{Relu}(x) = \max(x, 0)$. A function defined as in (3.1) is called a **neuron**, and we can combine several of them to build a so-called **neural network**. One simple but very common type of neural network is the so-called **feedforward** network, characterized by having neurons grouped in layers, and the input to neurons in the $(l+1)-$st layer are exclusively neurons from the $l-$th layer. In particular, we want to work on **shallow** feedforward neural networks, i.e. feedforward neural networks with only one hidden layer.

Definition 3.1.1 (**Shallow feedforward neural network**)**.** *Let N be the number of neurons and $U := (u_1, \dots, u_N)$ be the collection of neurons parameters, i.e. $u_n := (w_n, \theta_n, b_n)$ are the parameters of the n-th neuron. A **shallow feedforward neural network** is a function $\Phi_U : \mathbb{R}^d \to \mathbb{R}$ such that*

$$\mathbb{R}^d \ni\ x \mapsto \Phi_U(x) := \frac{1}{N} \sum_{n=1}^{N} \varphi_{u_n}(x). \tag{3.2}$$

In this case, $\mathbb{R}^d$ *takes the role of* ***input space*** *and* $x \in \mathbb{R}^d$ *is called an* ***input data****.*

In applications, it is used to apply an extra activation function after the linear combination of the neurons' outputs. From a mathematical point of view, this could over complicate our study, so we prefer to work with functions like (3.2).

Remark 12. Fixing a parametric configuration U, by Definition 3.1.1, determines a shallow neural network Φ_U. As a consequence, varying the parameter configuration while keeping the number of neurons and the activation function fixed produces different networks that share the same structure. In the literature and in applications, the term **architecture** refers to this fixed structural part and we can denote it as $\tilde{\Phi}$, to indicate that no parameters are specified. Let $\mathcal{H}_N$ denote the class of all shallow neural networks with N neurons, that is,

$$\mathcal{H}_N := \{\, \Phi_U : U \in \mathbb{R}^{N(d+2)} \,\}.$$

So, we can see that architecture induces a map $\tilde{\Phi} : \mathbb{R}^{N(d+2)} \to \mathcal{H}_N$ such that

$$\mathbb{R}^{N(d+2)} \ni U \mapsto \tilde{\Phi}(U) := \Phi_U.$$

In the context of supervised learning, we have a collection of input data (or **dataset**) $\{x_m\}_{m=1}^K \subset \mathbb{R}^d$ with **labels** $\{y_m\}_{m=1}^K \subset \mathbb{R}$. The goal is to split the dataset into two subsets, called the **training set** and the **test set**, using the former to train the neural network and the latter to test its ability to generalize. So, we can choose $1 \leq M < K$, obtaining the training set $\{x_m\}_{m=1}^M \subset \mathbb{R}^d$ with labels $\{y_m\}_{m=1}^M \subset \mathbb{R}$.

Definition 3.1.2 (**Loss function**)**.** *Fixed an architecture* $\tilde{\Phi}$ *and given a training set* $\{x_m\}_{m=1}^M$ *with labels* $\{y_m\}_{m=1}^M$*, we can define the associated* L^2 ***loss function*** *as* $L : \mathbb{R}^{N(d+2)} \to \mathbb{R}_+$ *such that*

$$\mathbb{R}^{N(d+2)} \ni U \mapsto \frac{1}{M} \sum_{m=1}^{M} |\Phi_U(x_m) - y_m|^2 , \tag{3.3}$$

where $\Phi_U = \tilde{\Phi}(U)$*.*

The function in Definition 3.1.2 gives us a measure of the neural network's prediction error, and it is also called **empirical risk**.

Remark 13. Different loss functions can be defined, but (3.3) is the simplest to formulate gradient flows for neural network training. Motivation comes from its convexity and smoothness with respect to $\Phi_U(x_m)$, and linearity of the gradient with respect to the prediction error. Other losses are not differentiable everywhere and may introduce additional non-linearities, complicating the analysis of gradient flows. It is important to note that, considering a non-linear activation function, (3.3) is not convex or smooth with respect to U, and this motivates the introduction of a **continuous formulation** of neural network training.

Given a loss function L, the training problem consists in solving the parametric minimization

$$\min_{U \in \mathbb{R}^{N(d+2)}} L(U), \tag{3.4}$$

that is, minimizing the empirical risk on all networks in $\mathcal{H}_N$. Usually in applications, to avoid **overfitting** in the training set, a convex **regularization** term $V : \mathbb{R}^{N(d+2)} \to \mathbb{R}_+$ is added. Then, the minimization problem (3.4) becomes

$$\min_{U \in \mathbb{R}^{N(d+2)}} F(U), \tag{3.5}$$

where $F := L + V : \mathbb{R}^{N(d+2)} \to \mathbb{R}_+$.

Example 4. *In applications, the most used regularization (or potential) terms are*

- *L^2-regularization or **weight decay**: $V(U) = \lambda \|U\|_2^2$,*
- *L^1-regularization or **LASSO** (Least Absolute Shrinkage and Selection Operator): $V(U) = \lambda \|U\|_1$.*

*Both potentials are convex, but only the first one is smooth. Depending on the structure of the dataset and on the learning task, one may prefer L^1, which promotes sparsity, or L^2, which penalizes large weights more uniformly. Modern optimizers such as **AdamW** use an **adaptive** form of L^2-regularization, where weight decay is applied parameter-wise and decoupled from the gradient update. Such adaptive schemes improve stability and generalization in large-scale neural network training.*

The standard approach to neural network training starts with the initialization of the parameters U at time $t = 0$ and consists of the implementation of a **gradient descent algorithm**, inspired by the minimization movement scheme seen in the previous chapters. From a theoretical point of view, we want to find a curve $U : [0, T) \to \mathbb{R}^{N(d+2)}$, which is a solution of the Cauchy problem

$$\begin{cases} \dot{U}(t) = -\partial^0 F(U(t)), \\ U(0) = U_0, \end{cases} \tag{3.6}$$

where U_0 represents the initial values of the parameters and $\partial^0 F(U)$ is the element with minimum Euclidean norm in $\partial F(U)$, i.e.

$$\partial^0 F(U) := \operatorname{argmin} \{|\xi| : \xi \in \partial F(U)\}.$$

Hence, training can be seen as a gradient flow with respect to the Euclidean norm in $\mathbb{R}^{N(d+2)}$. As a consequence of Remark 12, F is not convex with respect to U. Thus, we cannot use results like Theorem 1.3.2 to conclude that (3.6) has a unique solution.

3.2 Reformulation with probability measures

Now we want to reformulate neural network training in terms of probability measures and study the limit of the network as the number of neurons goes to infinity. We start by treating the parameters of each neuron as physical particles that live in $\mathbb{R}^{d+2}$, denoting by $\Omega \subset \mathbb{R}^{d+2}$ a closed convex set that contains all of them, and we suppose that all the data live in a bounded domain $D \subset \mathbb{R}^d$. Now, we fix an architecture $\tilde{\Phi}$ for our neural network and choose a suitable functional space $\mathcal{F}$ as an ambient space for possible realizations of the network, and from the previous section, we know that the architecture induces a map $\tilde{\Phi} : \Omega^N \to \mathcal{F}$. In applications, shallow networks defined as in (3.2) are generally continuous, but, for our purposes, it is sufficient to consider $\mathcal{F} = L^2(D)$. This choice is motivated by the fact that $L^2(D)$ is a separable Hilbert space, so it has many useful properties that simplify our discussion.

Remark 14. Let $\mathcal{F}$ be a separable Banach space. Then, by Pettis Theorem, a function $G : \Omega^N \to \mathcal{F}$ is **strongly measurable** (or Bochner measurable) if and only if it is **weakly measurable**. In particular, for $\mathcal{F} = L^2(D)$, which is a separable Hilbert space, it suffices to verify weak measurability to prove the Bochner measurability. Let $\tilde{\Phi} : \Omega^N \to L^2(D)$ be the map induced by the architecture of a neural network. Fixing a continuous activation function σ, for all $f \in L^2(D)$ the map

$$\begin{aligned}\Omega^N \ni U \mapsto (f, \Phi_U)_{L^2(D)} &= \int_D f(x)\, \Phi_U(x)\, dx \\ &= \frac{1}{N} \sum_{n=1}^{N} \int_D f(x)\, \varphi_{u_n}(x)\, dx \\ &= \frac{1}{N} \sum_{n=1}^{N} w_n \int_D f(x)\, \sigma(\theta_n \cdot x + b_n)\, dx\end{aligned}$$

is Lebesgue measurable in Ω^N. Thus, $\tilde{\Phi}$ is weakly measurable, so it is Bochner measurable. Moreover, if Ω is bounded, we also have

$$\int_{\Omega^N} \|\Phi_U\|_{L^2(D)}\, dU = \int_{\Omega^N} \left(\int_D |\Phi_U(x)|^2\, dx \right)^{\frac{1}{2}} dU < +\infty,$$

so $\tilde{\Phi} \in L^1(\Omega^N, L^2(D))$, i.e. $\tilde{\Phi}$ is Bochner integrable. Note that, if Ω is unbounded, the Bochner integrability can be ensured in other ways, as we will see in the sequel.

Then, we consider the empirical measure associated with the particles $u_1, \dots, u_n$

$$\mu_N = \frac{1}{N} \sum_{n=1}^{N} \delta_{u_n}, \tag{3.7}$$

where δ_{u_n} is the **Dirac's Delta measure** centered in u_n, i.e. the measure such

that

$$\delta_{u_n}(E) = \begin{cases} 1 & \text{if } u_n \in E, \\ 0 & \text{if } u_n \notin E, \end{cases}$$

for all $E \in \mathcal{B}(\Omega)$ (Borel-measurable subsets of Ω). Now, we can rewrite a shallow feedforward neural network Φ_U in terms of (3.7), using the subscript N to point out the number of its neurons

$$\Phi_N := \Phi_U = \frac{1}{N} \sum_{n=1}^{N} \varphi_{u_n} = \int_\Omega \varphi_u \, d\mu_N(u). \tag{3.8}$$

Remark 15. The expression (3.8) can be seen as a Bochner integral of the function $\varphi : \Omega \ni u \mapsto \varphi_u \in L^2(D)$ such that, considering $u = (w, \theta, b) \in \Omega$

$$D \ni x \mapsto \varphi_u(x) := w\, \sigma(\theta \cdot x + b). \tag{3.9}$$

Indeed, assuming, for example, that φ is weakly continuous, we have

$$\int_\Omega \|\varphi_u\|_{L^2(D)} \, d\mu_N(u) = \frac{1}{N} \sum_{n=1}^{N} \|\varphi_{u_n}\|_{L^2(D)} < +\infty.$$

We call φ a **neuronal map**, since it maps each set of parameters living in Ω to a neuron as defined in (3.1). The weak continuity of φ means that $\Omega \ni u \mapsto (\varphi_u, g)_{L^2} \in \mathbb{R}$ is continuous for all $g \in L^2(D)$. We could also obtain the existence of the Bochner integral in (3.8) assuming $\varphi_u \in L^2(D)$ for all $u \in \Omega$, but we need the weak continuity for the next result. If we suppose, as in different applications, that σ is continuous, we automatically have the strong continuity of φ and therefore its weak continuity.

If we want to study the limit of a shallow neural network as the number of neurons N goes to infinity, we must consider a sequence of networks $\{\Phi_N\}_{N\in\mathbb{N}}$ defined in (3.8) and study its convergence as $N \to \infty$.

Proposition 3.2.1. *Let $\Omega \subset \mathbb{R}^{d+2}$ be a closed bounded set and consider a sequence of shallow neural networks $\{\Phi_N\}_{N\in\mathbb{N}}$ defined as in (3.8) such that the neuronal map $\varphi : \Omega \to L^2(D)$ is weakly continuous. Then, there exists a subsequence $\{\Phi_{N_k}\}_k$ and $\mu \in \mathcal{P}(\mathbb{R}^{d+2})$, supported in Ω, such that*

$$\Phi_{N_k} \xrightarrow{k\to\infty} \Phi_\mu \ \text{ weakly in } L^2(D), \ \text{ where } \Phi_\mu := \int_\Omega \varphi_u \, d\mu(u). \tag{3.10}$$

*Here, $\Phi_\mu \in L^2(D)$ represents the **infinite-width limit** of the network and μ is called the **mean field limit** of $(\mu_N)_{N\in\mathbb{N}}$.*

Proof. Since Ω is a closed and bounded subset of $\mathbb{R}^{d+2}$, it is compact. Therefore, since all μ_N are supported in Ω, we have

$$\mu_N\left(\mathbb{R}^{d+2} \setminus \Omega\right) = 0 \ \text{ for all } N \in \mathbb{N}.$$

Thus, the sequence $(\mu_N)_{N\in\mathbb{N}}$ is tight and, by Theorem 2.1.9 (Prokhorov), there

exists a subsequence $(\mu_{N_k})_k$ and a probability measure $\mu \in \mathcal{P}(\mathbb{R}^{d+2})$, supported in Ω, such that

$$\mu_{N_k} \xrightarrow{k\to\infty} \mu.$$

Now, considering a test function $g \in L^2(D)$, we have

$$\begin{aligned}(\Phi_{N_k},\, g)_{L^2(D)} &= \left(\int_\Omega \varphi_u \, d\mu_{N_k}(u),\, g\right)_{L^2(D)} \\ &= \int_\Omega (\varphi_u,\, g)_{L^2(D)} \, d\mu_{N_k}(u).\end{aligned}$$

Since Ω is compact, there exists $C > 0$ such that

$$|(\varphi_u,\, g)| \le \|\varphi_u\|_{L^2} \, \|g\|_{L^2} \le C,$$

since the map $u \mapsto \varphi_u \mapsto \|\varphi_u\|_{L^2}$ is weakly lower semicontinuous. Hence, the function $\Omega \ni u \mapsto (\varphi_u,\, g)_{L^2(D)}$ is bounded and is also continuous, since φ is weakly continuous. So we have

$$\begin{aligned}\int_\Omega (\varphi_u,\, g)_{L^2(D)} \, d\mu_{N_k}(u) \xrightarrow{k\to\infty} \int_\Omega (\varphi_u,\, g)_{L^2(D)} \, d\mu(u) &= \left(\int_\Omega \varphi_u \, d\mu(u),\, g\right)_{L^2(D)} \\ &=: (\Phi_\mu,\, g)_{L^2(D)}\,.\end{aligned}$$

In conclusion, we proved $\Phi_{Nk} \xrightarrow{k\to\infty} \Phi_\mu$ weakly in $L^2(D)$. □

Remark 16. It is important to note that Proposition 3.2.1. does not ensure the uniqueness of the infinite-width limit, since convergence is obtained only up to subsequences. Although infinite-width limits are classical objects in the literature, their existence in this setting is usually not stated explicitly, probably because it is regarded as a straightforward consequence of compactness arguments. In works like [12], the existence of the limit is derived through probabilistic arguments or through the analysis of the network training dynamics. We used a functional-analytic viewpoint, because it is interesting to see that, given a shallow neural network, under suitable assumptions and independently of the training dynamics, its limit as the number of neurons goes to infinity exists. Therefore, the existence of an infinite-width limit can be conceptually separated from the study of training dynamics. Note that while this result relies on the compactness of Ω, in the sequel we will admit an unbounded space of parameters, overcoming the lack of compactness by introducing a localization framework using a family of nested closed convex subsets of Ω.

Now, we can give a continuous formulation of the loss function seen in Definition 3.1.2.

Definition 3.2.2 (**Squared loss function**). *Given a closed convex set $\Omega \subset \mathbb{R}^{d+2}$ for the parameters and a domain $D \subset \mathbb{R}^d$ for the data, we call* ***squared loss function*** *the map $R : L^2(D) \to [0, +\infty)$ such that*

$$L^2(D) \ni \Phi \mapsto R(\Phi) := \int_D |\Phi(x) - y(x)|^2 \, dx = \|\Phi - y\|^2_{L^2(D)}\,, \tag{3.11}$$

*where $y : D \to \mathbb{R}$ is the **label map** and, for consistency, we require $y \in L^2(D)$.*

Remark 17. We call y a label map, since its role is to map all possible data in the corresponding labels. Moreover, we can also call y an **objective map**, since it is a function that we want to approximate with a neural network. In [12], as in other works, the authors consider a joint distribution $\rho \in \mathcal{P}(D \times \mathbb{R})$ of input data $x \in D$ and labels $y \in \mathbb{R}$, with $\rho_x \in \mathcal{P}(D)$ being the marginal distribution of input data, defining the squared loss function as the expected risk

$$L^2(\rho_x) \ni \Phi \mapsto R(\Phi) := \int_{D\times\mathbb{R}} |\Phi(x) - y|^2 \, d\rho(x, y).$$

In our setting, we consider the deterministic scenario in which there is no label noise. Furthermore, we assume that the marginal distribution of the input data ρ_x is the uniform measure on D, which is proportional to the Lebesgue measure. Under these assumptions, up to the normalization constant $|D|^{-1}$, the expected risk is reduced to our formulation.

Finally, using Definition 3.2.2, we can introduce a regularized loss function in the context of signed measures.

Definition 3.2.3 (**Regularized loss function**). *Given a closed convex subset $\Omega \subset \mathbb{R}^{d+2}$ and a weakly continuous neuronal map $\varphi : \Omega \to L^2(D)$, we define a function $\tilde{L} : \mathcal{M}(\Omega) \to [0, +\infty)$ such that*

$$\mathcal{M}(\Omega) \ni \mu \mapsto \tilde{L}(\mu) := R(\Phi_\mu) := R\left(\int_\Omega \varphi_u \, d\mu(u)\right). \tag{3.12}$$

Moreover, consider a suitable potential function $V : \Omega \to [0, +\infty)$ and the map $\tilde{V} : \mathcal{M}(\Omega) \to [0, +\infty)$ such that

$$\mathcal{M}(\Omega) \ni \mu \mapsto \tilde{V}(\mu) := \int_\Omega V(u) \, d\mu(u). \tag{3.13}$$

*We call **regularized loss function** the map $\tilde{F} := \tilde{L} + \tilde{V} : \mathcal{M}(\Omega) \to [0, +\infty)$.*

Remark 18. Note that Φ_μ defined in (3.11), given a measure $\mu \in \mathcal{M}(\Omega)$ and a neuronal map φ, is unique. This is true because the Bochner integral

$$\Phi_\mu = \int_\Omega \varphi_u \, d\mu(u)$$

is uniquely defined in $L^2(D)$. Note that if $\mu = \mu_N = \frac{1}{N}\sum_{n=1}^N \delta_{u_n}$, we obtain a shallow neural network with N neurons, and, in general, Φ_μ can also be an infinite-width shallow neural network. Moreover, if φ is not a neuronal map, we cannot ensure that Φ_μ is a neural network.

Example 5. *A natural choice of regularizer $\tilde{V}$ is the **total variation norm***

$$\tilde{V}(\mu) := \|\mu\|_{TV} = |\mu|(\Omega) = \int_\Omega d\,|\mu|. \tag{3.14}$$

This kind of regularization is widely used in applications that benefit from sparsity, because it can be seen as a natural generalization of the L^1-norm to the space of signed measures.

In order to study the training dynamics of a shallow neural network in terms of probability measures, we need to introduce the so-called **particle gradient flow**. We start by defining $\tilde{F}_N : \Omega^N \to [0, +\infty)$ such that

$$\Omega^N \ni U \mapsto \tilde{F}_N(U) := \tilde{F}\left(\frac{1}{N}\sum_{n=1}^{N}\delta_{u_n}\right). \tag{3.15}$$

Definition 3.2.4 (**Particle gradient flow**). *A gradient flow for the functional $\tilde{F}_N$ is a locally absolutely continuous curve $U : [0, +\infty) \to \Omega^N$ which satisfies, for a.e. $t \geq 0$*

$$U'(t) \in -N\partial\tilde{F}_N(U(t)) \tag{3.16}$$

Now, we make some technical assumptions to prove the basic properties of this object that we need for the main results of this section.

Proposition 3.2.5. *Given a closed convex set $\Omega \subset \mathbb{R}^{d+2}$ and a Fréchet differentiable neuronal map $\varphi : \Omega \to L^2(D)$, consider a convex differentiable loss function $R : L^2(D) \to [0, +\infty)$ as defined in (3.11), with a differential dR that is Lipschitz on bounded sets and bounded on sublevel sets, and a λ_V-convex potential term $V : \Omega \to [0, +\infty)$, i.e $x \mapsto V(x) - \frac{\lambda_V}{2}|x|^2$ is convex with $\lambda_V \leq 0$. Assume that there exists a family $(Q_r)_{r>0}$ of nested nonempty closed convex subsets of Ω (not necessarily bounded) such that:*

1. *$\{u \in \Omega : dist(u, Q_r) \leq r'\} \subset Q_{r+r'}$ for all $r, r' > 0$*
2. *φ and V are bounded and $d\varphi$ is Lipschitz on each Q_r*
3. *there exists C_1, $C_2 > 0$ such that $\sup_{u\in Q_r}\left(\|d\varphi_u\|_{\mathcal{L}(\mathbb{R}^{d+2},L^2(D))} + |\partial V(u)|\right) \leq C_1 + C_2 r$ for all $r > 0$, where $|\partial V(u)|$ stands for the maximal euclidean norm for an element in $\partial V(u)$.*

Then, for any initialization $U(0) \in \Omega^N$, there exists a unique particle gradient flow $U : [0, +\infty) \to \Omega^N$ for $\tilde{F}_N$. Moreover, for a.e. $t > 0$, it holds

$$\frac{d}{ds}\tilde{F}_N(U(s))|_{s=t} = -\frac{1}{N}\left|U'(t)\right|^2, \tag{3.17}$$

and the velocity of the n-th particle is given by $u_n'(t) = v_t(u_n(t))$, where, denoting $\mu_{N,t} := \frac{1}{N}\sum_{n=1}^{N}\delta_{u_n(t)}$, the velocity field $v_t : \Omega \to \mathbb{R}^{d+2}$ is defined as

$$v_t(u) := \tilde{v}_t(u) - proj_{\partial V(u)}(\tilde{v}_t(u)), \tag{3.18}$$

with $\tilde{v}_t(u)$ being the negative gradient of the loss term with respect to the parameter u, explicitly given by the vector

$$\tilde{v}_t(u) := -\left[\langle dR\left[\Phi_{\mu_{N,t}}\right], \partial_j\varphi_u\rangle\right]_{j=1}^{d+2}, \tag{3.19}$$

where $dR[\Phi_{\mu_N,t}]$ denotes the differential of R applied to $\Phi_{\mu_N,t}$ and $\partial_j\varphi_u$ denotes the differential $d\varphi_u$ applied to the j-th vector of the canonical basis of $\mathbb{R}^{d+2}$.

Remark 19. The expression for the velocity field v_t highlights the dynamics of gradient flow. The projection operator $\mathrm{proj}_{\partial V(u)}$ appears because, when the potential V is non-smooth (as in the case of L^1-regularization), the gradient flow selects the subgradient of minimal norm. Furthermore, due to our specific choice of the loss function R in Definition 3.2.2, its Fréchet differential evaluated at Φ applied to a variation $h \in L^2(D)$ is given by $\langle dR[\Phi], h\rangle = 2(\Phi - y, h)_{L^2(D)}$. Therefore, the j-th component of $\tilde{v}_t(u)$ is directly computable as:

$$\left(\tilde{v}_t(u)\right)_j = -2\int_D \left(\Phi_{\mu_{N,t}}(x) - y(x)\right)\partial_j\varphi_u(x)\,dx,$$

which perfectly matches the classical backpropagation updates used in gradient descent algorithms for neural networks. Moreover, when V is differentiable, we have $v_t(u) = \tilde{v}_t(u) - \nabla V(u)$ and we recover the classical gradient of $\tilde{F}_N$. The crucial advantage of using the family $(Q_r)_{r>0}$ instead of standard closed balls is that the sets Q_r are not necessarily bounded. This allows us to naturally include directions of invariance or flat regions in Ω, where the neuronal map φ remains bounded even if the parameters grow to infinity. For example, consider a single neuron $\varphi_u(x) = w\,\sigma(\theta\cdot x + b)$. If $w = 0$, the inner parameters θ and b can diverge to infinity while the neuronal map remains identically zero. Similarly, for ReLU activation, a sufficiently negative b yields $\varphi_u \equiv 0$ regardless of the magnitude of θ. In these unbounded regions of Ω, φ and its differential are perfectly bounded and behave well.

Proof. Since R is convex and differentiable with a locally Lipschitz differential and the map $U \mapsto \Phi_U$ has a locally Lipschitz differential (due to the assumptions made on $d\varphi$), their composition $U \mapsto R(\Phi_U)$ is locally λ-convex for some $\lambda \in \mathbb{R}$. As the sum of a locally λ-convex function and a λ_V-convex potential V, $\tilde{F}_N$ is locally $(\lambda + \lambda_V)$-convex. It is easy to extend Theorem 1.3.2 to this case, see [14, Sec. 2.1], proving the existence of a unique particle gradient flow $U : [0, T) \to \Omega^N$ with the claimed properties, where $[0, T)$ is a maximal interval. Now, a general property of gradient flows for semiconvex functions is that, for a.e. $t > 0$, the derivative $U'(t)$ selects the negative subgradient of minimal norm in the set $-N\partial\tilde{F}_N(U(t))$. Since the functional $\tilde{F}_N$ is the sum of a smooth loss and a non-smooth potential, its subdifferential with respect to $U = (u_1, \ldots, u_N)$ splits component-wise. By applying the chain rule, the gradient of the loss term with respect to a single particle u_n is exactly $-\frac{1}{N}\tilde{v}_t(u_n)$, while the subdifferential of the potential term is $\frac{1}{N}\partial V(u_n)$. Since the particle gradient flow is defined by the inclusion $u_n'(t) \in -N[\partial\tilde{F}_N(U(t))]_n$, the velocity $v_t(u_n)$ of a single particle must be selected from the set:

$$-N\left(-\frac{1}{N}\tilde{v}_t(u_n) + \frac{1}{N}\partial V(u_n)\right) = \tilde{v}_t(u_n) - \partial V(u_n).$$

This leads to the explicit formula for the velocity field with pointwise minimal

norm for each particle u:

$$\begin{aligned} v_t(u) &= \arg\min\left\{|v|^2 : \tilde{v}_t(u) - v \in \partial V(u)\right\} \\ &= \tilde{v}_t(u) - \arg\min\left\{|\tilde{v}_t(u) - z|^2 : z \in \partial V(u)\right\} \\ &= \left(\mathrm{id} - \mathrm{proj}_{\partial V(u)}\right)(\tilde{v}_t(u)). \end{aligned}$$

Finally, in the specific case of gradient flows of lower bounded functions, we can derive estimates that imply $T = +\infty$ (even if $\tilde{F}_N$ is not globally semiconvex). Indeed, for all $t > 0$ and applying Jensen's inequality, we obtain:

$$\begin{aligned} \tilde{F}_N(U(0)) - \tilde{F}_N(U(t)) = -\int_0^t \frac{d}{ds}\tilde{F}_N(U(s))\,ds &= \int_0^t \frac{1}{N}\left|U'(s)\right|^2 ds \\ &\geq \frac{1}{Nt}\left(\int_0^t \left|U'(s)\right| ds\right)^2. \end{aligned}$$

Since the loss function and the regularizer are non-negative, $\tilde{F}_N$ is lower bounded by zero. Thus, the inequality proves that the gradient flow has a finite length on bounded time intervals. If T were finite, since Ω is a closed subset of $\mathbb{R}^{d+2}$ (hence complete), the limit $U^* := \lim_{t\to T^-} U(t)$ would exist and belong to Ω^N. By Theorem 1.3.2, the flow could then be extended beyond T taking U^* as a new initial condition, which contradicts the assumption that $[0, T)$ is the maximal interval of existence. Hence $T = +\infty$ and the particle gradient flow is globally defined. □

The expression of the velocity of each particle as the evaluation of a velocity field will help us to generalize the particle gradient flow to arbitrary measures, not only atomic ones. Moreover, there is a direct link between the velocity field and the functional $\tilde{F}$. The differential of $\tilde{F}$ evaluated on a measure $\mu \in \mathcal{M}(\Omega)$ is represented by the function $\tilde{F}'(\mu) : \Omega \to \mathbb{R}$ defined as

$$u \mapsto \tilde{F}'(\mu)(u) := \langle dR\left[\Phi_\mu\right], \varphi_u\rangle + V(u). \tag{3.20}$$

Thus, $-v_t$ is simply the field of minimal norm subgradients of $\tilde{F}'(\mu_{N,t})$. We write this relation $v_t \in -\partial\tilde{F}'(\mu_{N,t})$, where the set $\partial\tilde{F}'$ denotes the **Wasserstein subdifferential** of $\tilde{F}$, as it can be interpreted as the subdifferential of $\tilde{F}$ relative to the Wasserstein metric on $\mathcal{P}_2(\Omega)$.

Remark 20. It is easy to see that the differential of $\tilde{F}$ is defined as in (3.20). Considering $\mu, \nu \in \mathcal{M}(\Omega)$, the directional derivative of $\tilde{F}$ at μ along the direction

ν is computed as

$$\begin{aligned}\frac{d}{d\varepsilon}\tilde{F}(\mu+\varepsilon\nu)\Big|_{\varepsilon=0} &= \frac{d}{d\varepsilon}\left[R(\Phi_{\mu+\varepsilon\nu}) + \int_\Omega V(u)\,d(\mu+\varepsilon\nu)(u)\right]_{\varepsilon=0} \\ &= \left\langle dR[\Phi_\mu], \frac{d}{d\varepsilon}\left[\int_\Omega \varphi_u\,d(\mu+\varepsilon\nu)(u)\right]_{\varepsilon=0}\right\rangle \\ &\quad + \frac{d}{d\varepsilon}\left[\int_\Omega V(u)\,d(\mu+\varepsilon\nu)(u)\right]_{\varepsilon=0} \\ &= \left\langle dR[\Phi_\mu], \int_\Omega \varphi_u\,d\nu(u)\right\rangle + \int_\Omega V(u)\,d\nu(u) \\ &= \int_\Omega \left[\langle dR[\Phi_\mu], \varphi_u\rangle + V(u)\right] d\nu(u).\end{aligned}$$

By the definition of the first variation in $\mathcal{M}(\Omega)$, $\tilde{F}'(\mu):\Omega\to\mathbb{R}$ is the unique function that satisfies

$$\frac{d}{d\varepsilon}\tilde{F}(\mu+\varepsilon\nu)\Big|_{\varepsilon=0} = \int_\Omega \tilde{F}'(\mu)(u)\,d\nu(u) \quad \text{for all } \nu\in\mathcal{M}(\Omega).$$

By identifying the integrand, we immediately obtain (3.20).

Definition 3.2.6 (**Wasserstein gradient flow**). *A Wasserstein gradient flow for the functional $\tilde{F}$ on the interval $[0,T)$ is an absolutely continuous path $[0,T)\ni t\mapsto\mu_t\in\mathcal{P}_2(\Omega)$ that satisfies, distributionally on $[0,T)\times\Omega$, the continuity equation*

$$\partial_t\mu_t = -\operatorname{div}(v_t\mu_t) \quad \textit{where} \quad v_t\in-\partial\tilde{F}'(\mu_t). \tag{3.21}$$

This means that for all smooth test functions $\psi:(0,T)\times\mathbb{R}^{d+2}\to\mathbb{R}$ with compact support, it holds

$$\int_0^T\int_{\mathbb{R}^{d+2}} \left(\partial_t\psi(t,u)+\nabla_u\psi(t,u)\cdot v_t(u)\right)\,d\mu_t(u)dt = 0. \tag{3.22}$$

The integrability condition $\int_0^{t_0}\int_{\mathbb{R}^{d+2}}|v_t(u)|\,d\mu_t(u)dt<+\infty$ for all $t_0<T$ should also hold.

Remark 21. Note that Definition 2.3.6 of a Wasserstein gradient flow only involves probability measures which are absolutely continuous with respect to the Lebesgue measure; hence, this new definition is more general and is the most useful for our discussion. We can also see that this is a proper generalization of Definition 3.2.4 since, whenever $[0,+\infty)\ni t\mapsto U(t)\in\Omega^N$ is a particle gradient flow for $\tilde{F}_N$, then $t\mapsto\mu_{N,t}:=\frac{1}{N}\sum_{n=1}^N\delta_{u_n(t)}\in\mathcal{P}_2(\Omega)$ is a Wasserstein gradient flow of $\tilde{F}$. To see this, consider the velocity field defined in (3.18). It is easy to see that $t\mapsto\mu_{N,t}$ is absolutely continuous for W_2, since $t\mapsto U(t)$ is absolutely

continuous and, for each $t, s \in [0, +\infty)$, with $t > s$, it holds

$$\begin{aligned} W_2^2(\mu_{N,t}, \mu_{N,s}) &= \min_{\sigma \in S_N} \frac{1}{N} \sum_{n=1}^{N} \left| u_n(t) - u_{\sigma(n)}(s) \right|^2_{\mathbb{R}^{d+2}} \\ &\leq \frac{1}{N} \sum_{n=1}^{N} |u_n(t) - u_n(s)|^2_{\mathbb{R}^{d+2}} \\ &= \frac{1}{N} |U(t) - U(s)|^2_{\mathbb{R}^{N(d+2)}} \leq \frac{1}{N} (t-s) \int_s^t |U'(r)|^2_{\mathbb{R}^{N(d+2)}} \, dr, \end{aligned}$$

where the first equality holds because we are considering atomic measures and S_N denotes the set of all permutations of the indices $\{1, \dots, N\}$. Moreover, for any smooth function $\psi : (0, +\infty) \times \mathbb{R}^{d+2} \to \mathbb{R}$ with compact support, we have

$$\begin{aligned} 0 &= \frac{1}{N} \sum_{n=1}^{N} \int_0^{+\infty} \frac{d}{dt} \psi(t, u_n(t)) \, dt \\ &= \frac{1}{N} \sum_{n=1}^{N} \int_0^{+\infty} \left(\partial_t \psi(t, u_n(t)) + \nabla_u \psi(t, u_n(t)) \cdot v_t(u_n(t)) \right) dt \\ &= \int_0^{+\infty} \int_\Omega \left(\partial_t \psi(t, u) + \nabla_u \psi(t, u) \cdot v_t(u) \right) d\mu_{N,t} dt, \end{aligned}$$

which means that (3.22) holds. Note that $(\mu_{N,t})_t$ has the same number of atoms throughout the dynamic. In particular, if no minimizers of $\tilde{F}$ is an atomic measure with at most N atoms, then $(\mu_{N,t})_t$ is guaranteed not to converge to a minimizer of $\tilde{F}$.

Now, we want to prove the existence and uniqueness of Wasserstein gradient flows for the functional $\tilde{F}$. Our strategy is to use, as an intermediate step, the Wasserstein gradient flows for the family of "localized" functionals $\tilde{F}^{(r)} : \mathcal{P}_2(\Omega) \to \mathbb{R}$, defined for $r > 0$, as

$$\tilde{F}^{(r)}(\mu) := \begin{cases} \tilde{F}(\mu) & \text{if } \mu(Q_r) = 1, \\ +\infty & \text{otherwise} \end{cases} \tag{3.23}$$

where $(Q_r)_{r>0}$ is the nested family of subsets of $\mathbb{R}^{d+2}$ that appear in Proposition 3.2.5. For $r > 0$, we say that $\gamma \in \mathcal{P}(\Omega \times \Omega)$ is an admissible transport plan if both its marginals are concentrated on Q_r and have finite second moments. We denote by $C_p(\gamma) := \left(\int |x - y|^p \, d\gamma(x, y) \right)^{1/p}$ the transport cost associated with γ and, identifying $L^2(D)$ with its dual space, introduce the quantities

$$\|d\varphi\|_{\infty,r} = \sup_{u \in Q_r} \|d\varphi_u\|_{\mathcal{L}(\mathbb{R}^{d+2}, L^2(D))} \qquad L_{d\varphi} = \sup_{\substack{u, \tilde{u} \in Q_r \\ u \neq \tilde{u}}} \frac{\|d\varphi_{\tilde{u}} - d\varphi_u\|_{\mathcal{L}(\mathbb{R}^{d+2}, L^2(D))}}{|\tilde{u} - u|}$$

$$\|dR\|_{\infty,r} = \sup_{f \in \mathcal{F}_r} \|dR_f\|_{L^2(D)} \qquad L_{dR} = \sup_{\substack{f, g \in \mathcal{F}_r \\ f \neq g}} \frac{\|dR_f - dR_g\|_{L^2(D)}}{\|f - g\|_{L^2(D)}}$$

where $\mathcal{F}_r := \{\Phi_\mu \in L^2(D) : \mu \in \mathcal{P}(\Omega), \mu(Q_r) = 1\}$ is bounded in $L^2(D)$. Due to assumptions made in Proposition 3.2.5, these quantities are finite for all $r > 0$.

Lemma 3.2.7. *Under the assumptions made in Proposition 3.2.5, suppose, for simplicity, that $V = 0$. For all $r > 0$, $\tilde{F}^{(r)}$ is proper and continuous for W_2 on its closed domain. Moreover,*

1. *there exists $\lambda_r > 0$ such that for any admissible transport plan γ, considering the transport interpolation $\mu_t^\gamma := ((1-t)\pi_X + t\pi_Y)_\#\gamma$, the function $t \mapsto \tilde{F}^{(r)}(\mu_t^\gamma)$ is differentiable with a $\lambda_r C_2^2(\gamma)$-Lipschitz derivative;*

2. *for μ concentrated on Q_r, a vector field $\xi \in L^2(\mu, \mathbb{R}^{d+2})$ (i.e., the space of real vector fields that are square integrable with respect to μ) satisfies, for any admissible transport plan γ with first marginal μ,*

$$\tilde{F}((\pi_Y)_\#\gamma) \geq \tilde{F}(\mu) + \int \xi(u) \cdot (\tilde{u} - u) d\gamma(u, \tilde{u}) + o(C_2(\gamma))$$

if and only if $\xi(u) \in \partial(\tilde{F}'(\mu) + i_{Q_r})(u)$ for μ-almost every $u \in \Omega$, where i_{Q_r} is the convex function on Ω that takes value 0 on Q_r and $+\infty$ outside.

Proof. Clearly, fixed $r > 0$, we see that $\tilde{F}^{(r)}$ is proper, because $\tilde{F}^{(r)}(\delta_{u_0}) = R(\varphi_{u_0}) = \|\varphi_{u_0} - y\|^2_{L^2(D)}$ is finite whenever $u_0 \in Q_r$. To see its continuity for W_2, consider $(\mu_k)_{k\in\mathbb{N}}, \bar{\mu} \in \mathcal{P}_2(\Omega)$ such that $W_2(\mu_k, \bar{\mu}) \to 0$. From the assumptions made previously in Proposition 3.2.5, we know that the differential $d\varphi$ has at most a linear growth. Since Ω is convex, by a direct consequence of the Mean Value Theorem in Banach spaces (see [18, Thm. 3.3.2]), the map φ is locally Lipschitz with a linearly growing constant. That is, there exists $C > 0$ such that for all $u, v \in \Omega$:

$$\|\varphi_u - \varphi_v\|_{L^2(D)} \leq C(1 + |u| + |v|)|u - v|.$$

Given an optimal transport plan π_k between μ_k and $\bar{\mu}$ for the W_2 distance, using the Bochner integral property shown in [16, Prop. E.5] in the first inequality, we have

$$\begin{aligned}
\|\Phi_{\mu_k} - \Phi_{\bar{\mu}}\|_{L^2(D)} &= \left\| \int_\Omega \varphi_u \, d\mu_k(u) - \int_\Omega \varphi_v \, d\bar{\mu}(v) \right\|_{L^2(D)} \\
&= \left\| \int_{\Omega\times\Omega} (\varphi_u - \varphi_v) \, d\pi_k(u, v) \right\|_{L^2(D)} \\
&\leq \int_{\Omega\times\Omega} \|\varphi_u - \varphi_v\|_{L^2(D)} \, d\pi_k(u, v) \\
&\leq C \int_{\Omega\times\Omega} (1 + |u| + |v|)|u - v| \, d\pi_k(u, v) \\
&\leq C \left(\int_{\Omega\times\Omega} (1 + |u| + |v|)^2 \, d\pi_k(u, v) \right)^{\frac{1}{2}} \left(\int_{\Omega\times\Omega} |u - v|^2 \, d\pi_k(u, v) \right)^{\frac{1}{2}}.
\end{aligned}$$

Since μ_k converges to $\bar{\mu}$ in $\mathcal{P}_2(\Omega)$, their second moments are uniformly bounded. Hence, the first integral is bounded by a constant $\tilde{C} > 0$, so we have

$$\|\Phi_{\mu_k} - \Phi_{\bar{\mu}}\|_{L^2(D)} \leq \tilde{C}\, W_2(\mu_k, \bar{\mu}) \to 0 \quad \text{as } k \to +\infty.$$

Thus, $\Phi_{\mu_k} \to \Phi_{\bar{\mu}}$ strongly in $L^2(D)$ and, since R is continuous in the strong topology of $L^2(D)$, we have $\tilde{F}(\mu_k) \to \tilde{F}(\bar{\mu})$. This implies that $\tilde{F}^{(r)}$ is continuous on its closed domain $\{\mu \in \mathcal{P}_2(\Omega) : \mu(Q_r) = 1\}$.

Proof of 1. Consider $h(t) := \tilde{F}^{(r)}(\mu_t^\gamma)$. Since dR is Lipschitz on $\mathcal{F}_r$ and $d\varphi$ is Lipschitz on Q_r, $h(t)$ is differentiable with its derivative given by the chain rule

$$h'(t) = \frac{d}{dt}\tilde{F}^{(r)}(\mu_t^\gamma) = \left\langle dR\left[\Phi_{\mu_t^\gamma}\right], \int_{\Omega\times\Omega} d\varphi_{(1-t)x+ty}(y-x)d\gamma(x,y)\right\rangle.$$

In particular, we can differentiate $t \mapsto \int_\Omega \varphi\, d\mu_t^\gamma = \int_{\Omega\times\Omega} \varphi_{(1-t)x+ty} d\gamma(x,y)$ because all $(\mu_t^\gamma)_t$ are supported on Q_r, where $d\varphi$ is uniformly bounded and Bochner integrals admit a dominated convergence theorem [16, Thm. E6], so we can safely differentiate under the integral sign. For $0 \leq t_1 < t_2 < 1$, we have the bounds

$$|h'(t_2) - h'(t_1)| \leq (I) + (II)$$

where, on the one hand,

$$\begin{aligned}
(I) &= \left|\left\langle dR\left[\Phi_{\mu_{t_2}^\gamma}\right] - dR\left[\Phi_{\mu_{t_1}^\gamma}\right], \int_{\Omega\times\Omega} d\varphi_{(1-t_2)x+t_2y}(y-x)d\gamma(x,y)\right\rangle\right| \\
&\leq \left[L_{dR}\cdot\left\|\Phi_{\mu_{t_2}^\gamma} - \Phi_{\mu_{t_1}^\gamma}\right\|_{L^2(D)}\right]\cdot[\|d\varphi\|_{\infty,r}\cdot C_1(\gamma)] \\
&\leq \left[L_{dR}\cdot\int_{\Omega\times\Omega}\left\|\varphi_{(1-t_2)x+t_2y} - \varphi_{(1-t_1)x+t_1y}\right\|_{L^2(D)} d\gamma(x,y)\right]\cdot[\|d\varphi\|_{\infty,r}\cdot C_1(\gamma)] \\
&\leq [L_{dR}\cdot\|d\varphi\|_{\infty,r}\cdot|t_2-t_1|\cdot C_1(\gamma)]\cdot[\|d\varphi\|_{\infty,r}\cdot C_1(\gamma)] \\
&\leq L_{dR}\cdot\|d\varphi\|_{\infty,r}^2\cdot C_2^2(\gamma)\cdot|t_2-t_1|
\end{aligned}$$

where we used Jensen's inequality to obtain $C_1^2(\gamma) \leq C_2^2(\gamma)$ in the last line. On the other hand,

$$\begin{aligned}
(II) &= \left|\left\langle dR\left[\Phi_{\mu_{t_1}^\gamma}\right], \int_{\Omega\times\Omega}[d\varphi_{(1-t_2)x+t_2y} - d\varphi_{(1-t_1)x+t_1y}](y-x)d\gamma(x,y)\right\rangle\right| \\
&\leq \|dR\|_{\infty,r}\cdot L_{d\varphi}\cdot C_2^2(\gamma)\cdot|t_1-t_2|
\end{aligned}$$

As a consequence, h' is $\lambda_r\cdot C_2^2(\gamma)$ Lipschitz with $\lambda_r := L_{dR}\,\|d\varphi\|_{\infty,r}^2 + L_{d\varphi}\,\|dR\|_{\infty,r}$. These bounds may explode when r goes to infinity, and this is why we work with measures supported on Q_r.

Proof of 2. The proof is similar, with the difference that this property is a local one. Thanks to the assumptions made on the neuronal map φ and on the loss R, we can consider their first-order Taylor expansions for $\tilde{u}, u \in Q_r$ and

$f, g \in \mathcal{F}_r$,

$$\varphi_{\tilde{u}} = \varphi_u + d\varphi_u(\tilde{u} - u) + M(u, \tilde{u})$$
$$R(g) = R(f) + \langle dR[f], g - f\rangle + N(f, g)$$

where the remainders M and N satisfy $\|M(\tilde{u}, u)\| \leq \frac{1}{2} L_{d\varphi} \cdot |\tilde{u} - u|^2$ and $\|N(f, g)\| \leq \frac{1}{2} L_{dR} \|f - g\|^2$. We denote by μ and ν, respectively, the first and second marginals of γ. Then, we assume that they are both concentrated on Q_r, and obtain, by composition, the Taylor expansion

$$\tilde{F}^{(r)}(\nu) = \tilde{F}^{(r)}(\mu) + \int_{\Omega\times\Omega} \langle dR[\Phi_\mu], d\varphi_u(\tilde{u} - u)\rangle \, d\gamma(\tilde{u}, u) + (I) + (II),$$

where $(I) = \left\langle dR[\Phi_\mu], \int_{\Omega\times\Omega} M(u, \tilde{u}) d\gamma(\tilde{u}, u) \right\rangle$, so

$$|(I)| \leq \frac{1}{2} \|dR\|_{\infty,r} \cdot L_{d\varphi} \cdot C_2^2(\gamma) = o(C_2(\gamma))$$

and $(II) = N(\Phi_\mu, \Phi_\nu)$, so

$$\begin{aligned}
|(II)| &\leq \frac{1}{2} L_{dR} \left\| \int d\varphi_u(\tilde{u} - u) d\gamma(\tilde{u}, u) + \int M(u, \tilde{u}) d\gamma(\tilde{u}, u) \right\|^2 \\
&\leq \frac{1}{2} L_{dR} \left[\|d\varphi\|_{\infty,r} \cdot C_1(\gamma) + \frac{1}{2} L_{d\varphi} \cdot C_2^2(\gamma) \right]^2 \\
&\leq \frac{1}{2} L_{dR} \left[\|d\varphi\|_{\infty,r} \cdot C_2(\gamma) + \frac{1}{2} L_{d\varphi} \cdot C_2^2(\gamma) \right]^2 = o(C_2(\gamma))
\end{aligned}$$

where we used Hölder's inequality in the last line. As a consequence,

$$\tilde{F}^{(r)}(\nu) = \tilde{F}^{(r)}(\mu) + \int \langle dR[\Phi_\mu], d\varphi_u(\tilde{u} - u)\rangle \, d\gamma(\tilde{u}, u) + o(C_2(\gamma))$$

and remember that the j-th component of $\nabla \tilde{F}'(\mu)$ is $u \mapsto \langle dR[\Phi_\mu], d\varphi_u(\mathbf{e}_j)\rangle$, where $\mathbf{e}_j$ is the j-th vector of the canonical basis of $\mathbb{R}^{d+2}$. This means that a velocity field $v \in L^2(\mu, \mathbb{R}^{d+2})$ satisfies 2. in the interior of Q_r if and only if $v(u) \in \nabla \tilde{F}'(\mu)(u)$ for μ-almost every $u \in \mathring{Q}_r$. If we now consider the boundary of Q_r, we have more freedom in the choice of $v(u)$, since ν is constrained to be supported on Q_r. Therefore, $v(u) - \nabla \tilde{F}'(\mu)(u)$ can live in the normal cone of Q_r at u, which is the set $\partial \iota_{Q_r}(u)$. Thus, the condition relaxes as $v(u) \in \partial(\tilde{F}'(\mu) + \iota_{Q_r})(u)$. □

Remark 22. Note that, as seen in Remark 6, μ_t^γ corresponds to a constant-speed geodesic in $\mathcal{P}_2(\Omega)$ provided that γ is an optimal transport plan. Under this optimality condition, we have $C_2^2(\gamma) = W_2^2(\pi^1_\#\gamma, \pi^2_\#\gamma)$. Hence, the first statement of Lemma 3.2.7 tells us that the functional $\tilde{F}$ is differentiable along the geodesics of $\mathcal{P}_2(\Omega)$ with a $\lambda_r W_2^2(\pi^1_\#\gamma, \pi^2_\#\gamma)$-Lipschitz derivative. The second statement shows a local property of $\tilde{F}$, which, as illustrated in Figure 3.1, allows us to characterize its subdifferential both in the interior and on the boundary of Q_r, given $r > 0$.

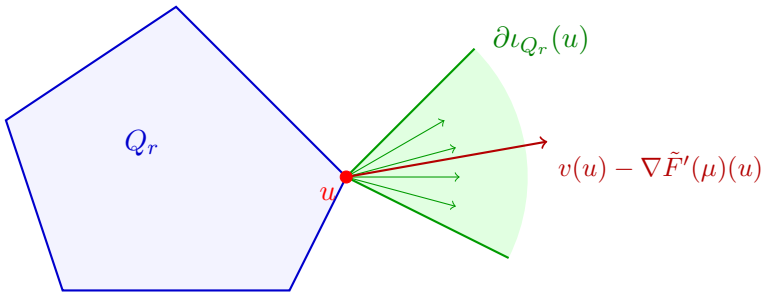

Figure 3.1: Geometric intuition of the subdifferential condition on ∂Q_r.

The previous properties are sufficient to guarantee that Wasserstein gradient flows for the functionals $\tilde{F}^{(r)}$ are well defined.

Lemma 3.2.8. *Under the assumptions made in Proposition 3.2.5, there exists a unique Wasserstein gradient flow for $\tilde{F}^{(r)}$ starting from any $\mu_0 \in \mathcal{P}_2(\Omega)$ concentrated on Q_r, i.e. a curve $(\mu_t^{(r)})_{t\geq 0}$, continuous in $\mathcal{P}_2(\Omega)$, that solves the continuity equation*

$$\partial_t \mu_t^{(r)} + \operatorname{div}(v_t^{(r)} \mu_t^{(r)}) = 0 \tag{3.24}$$

where, for all $t > 0$,

$$v_t^{(r)}(u) \in -\partial(\tilde{F}'(\mu_t^{(r)}) + \iota_{Q_r})(u) \quad \text{for } \mu_t^{(r)}\text{-a.e } u \in \Omega. \tag{3.25}$$

Proof. Consider the constant-speed geodesic $\mu_t^\gamma = ((1-t)\pi^1 + t\pi^2)_\#\gamma$, where γ is the **optimal** transport plan between $\mu := \pi^1_\#\gamma$ and $\nu := \pi^2_\#\gamma$. In particular, this implies $C_2^2(\gamma) = W_2^2(\mu,\nu)$. In the case $V = 0$, we already know, by Lemma 3.2.7, that $h(t) := \tilde{F}^{(r)}(\mu_t^\gamma)$ is differentiable with $\lambda_r \cdot C_2^2(\gamma)$-Lipschitz derivative, so for all $t_1, t_2 \in [0,1]$ we have

$$|h'(t_2) - h'(t_1)| \leq \lambda_r C_2^2(\gamma)|t_2 - t_1|. \tag{3.26}$$

Now, it is easy to see that $g(t) := h(t) + \frac{\lambda_r \cdot C_2^2(\gamma)}{2} t^2$ is convex. Indeed, $g'(t) = h'(t) + \lambda_r \cdot C_2^2(\gamma) t$ is monotonically non-decreasing, because, for $t_2 \geq t_1$ we have

$$\begin{aligned} g'(t_2) - g'(t_1) &= h'(t_2) - h'(t_1) + \lambda_r \cdot C_2^2(\gamma)(t_2 - t_1) \\ &\geq -\lambda_r \cdot C_2^2(\gamma)(t_2 - t_1) + \lambda_r \cdot C_2^2(\gamma)(t_2 - t_1) = 0, \end{aligned}$$

where we used (3.26) in the last line. Therefore, we obtain $(-\lambda_r)$-semiconvexity of $\tilde{F}^{(r)}$ along geodesics:

$$\begin{aligned} \tilde{F}^{(r)}(\mu_t^\gamma) &\leq (1-t)\tilde{F}^{(r)}(\mu) + t\tilde{F}^{(r)}(\nu) + \frac{\lambda_r \cdot C_2^2(\gamma)}{2} t(1-t) \\ &= (1-t)\tilde{F}^{(r)}(\mu) + t\tilde{F}^{(r)}(\nu) + \frac{\lambda_r}{2} t(1-t) W_2^2(\mu,\nu). \end{aligned}$$

Now suppose that the regularization potential V is λ_V-semiconvex, i.e. for all

$x, y \in \Omega$ and for all $t \in [0,1]$ we have

$$V((1-t)x + ty) \leq (1-t)V(x) + tV(y) - \frac{\lambda_V}{2} t(1-t)|x-y|^2.$$

Integrating this inequality along the optimal transport plan γ, we obtain

$$\begin{aligned}\tilde{V}(\mu_t^\gamma) &\leq (1-t)\tilde{V}(\mu) + t\tilde{V}(\nu) - \frac{\lambda_V}{2} t(1-t)C_2^2(\gamma) \\ &= (1-t)\tilde{V}(\mu) + t\tilde{V}(\nu) - \frac{\lambda_V}{2} t(1-t)W_2^2(\mu,\nu),\end{aligned}$$

which proves the λ_V-semiconvexity of $\tilde{V}$ along geodesics. Combining these results, we find that the total functional $\tilde{F}^{(r)}$ is $(\lambda_V - \lambda_r)$-semiconvex along geodesics. Applying [15, Thm 11.2.1], we conclude that there exists a unique Wasserstein gradient flow for $\tilde{F}^{(r)}$. Thanks to Lemma 3.2.7, easily readapted with the potential term $\tilde{V}$, this flow is characterized as in (3.25). □

Now we can prove the well-posedness of Wasserstein gradient flows for the original functional $\tilde{F}$. Note that, by the characterization in Lemma 3.2.8, the Wasserstein gradient flows for the functionals $\tilde{F}^{(r)}$ all coincide for $r > r_0 > 0$ on $[0,T]$ if $\mu_t^{(2r_0)}$ is concentrated in Q_{r_0} for all $t \in [0,T]$. So, our strategy is simply to ensure that for all time T, such a r_0 exists, i.e. to ensure that the support of gradient flows does not grow too fast.

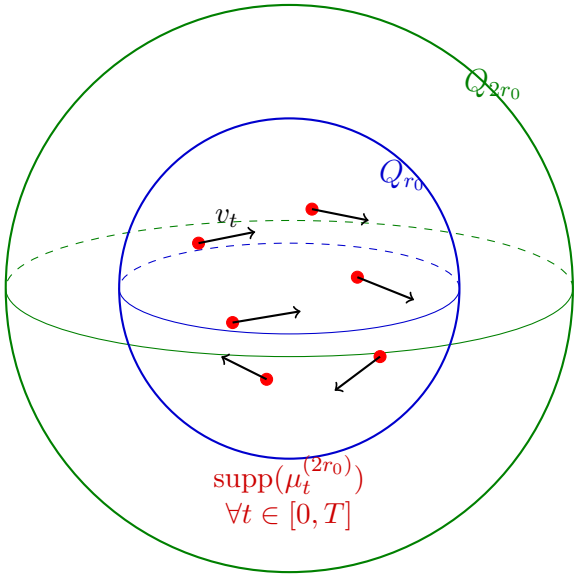

Figure 3.2: Geometric interpretation of the interior confinement explained above. If the support of the measure $\mu_t^{(2r_0)}$ remains confined in the interior of the smaller set Q_{r_0} for all $t \in [0,T]$, the particles never interact with the boundary of Q_{2r_0}. Consequently, the subdifferential condition reduces to the pure gradient $v_t(u) = -\nabla \tilde{F}'(\mu_t)(u)$, and the flows associated with different $r > r_0$ coincide.

Proposition 3.2.9 (**Existence and Uniqueness**). *Under the assumptions made in Proposition 3.2.5, if $\mu_0 \in \mathcal{P}_2(\Omega)$ is concentrated on a set $Q_{r_0} \subset \Omega$, then there exists a unique Wasserstein gradient flow $(\mu_t)_{t\geq 0}$ for $\tilde{F}$ starting from μ_0. It satisfies the continuity equation with the velocity field $v_t : \Omega \to \mathbb{R}^{d+2}$ defined as*

$$v_t(u) := \tilde{v}_t(u) - \mathrm{proj}_{\partial V(u)}(\tilde{v}_t(u)), \tag{3.27}$$

with $\tilde{v}_t(u)$ being the negative gradient of the loss term with respect to the parameter u, explicitly given by the vector

$$\tilde{v}_t(u) := -\left[\langle dR\left[\Phi_{\mu_t}\right], \partial_j \varphi_u\rangle\right]_{j=1}^{d+2}, \tag{3.28}$$

where $dR[\Phi_{\mu_t}]$ denotes the differential of R applied to Φ_{μ_t} and $\partial_j \varphi_u$ denotes the differential $d\varphi_u$ applied to the j-th vector of the canonical basis of $\mathbb{R}^{d+2}$.

Proof. Given $r_0 > 0$ such that μ_0 is concentrated on Q_{r_0}, by Lemma 3.2.8 we know that, for all $r > r_0$, there exists a unique, globally defined, Wasserstein gradient flow $(\mu_t^{(r)})_{t\geq 0}$ for $\tilde{F}^{(r)}$. For all $r > r_0$, consider the first time exit from Q_r:

$$t_r := \inf\{t > 0 : \mu_t^{(2r)}(Q_r) < 1\}.$$

Note that the definition of t_r involves the flow $(\mu_t^{(2r)})_t$ but in fact, for all $\bar{r} > r$ and $0 \leq t \leq t_r$, it holds $\mu_t^{(2r)} = \mu_t^{(\bar{r})}$ by the uniqueness given in Lemma 3.2.8. Hence, if $t_r > 0$, we have the existence and uniqueness of a Wasserstein gradient flow of $\tilde{F}$ on $[0, t_r]$. We only have to prove that $t_r \to +\infty$ whenever $r \to +\infty$, so that the Wasserstein gradient flow can be defined at all times. Given the property of $v^{(r)}$ in Lemma 3.2.8, for all $0 \leq t \leq t_r$, we have $v_t^{(r)} \in \partial \tilde{F}'(\mu_t^{(r)})$ in $L^2(\mu_t^{(r)}; \mathbb{R}^{d+2})$. Therefore, using the assumptions made in Proposition 3.2.5 (boundedness of dR on sublevel sets and assumption 3.), we obtain, for all $0 \leq t \leq t_r$,

$$|v_t^{(r)}(u)| \leq C_1 + C_2 r$$

with constants C_1 and C_2 independent of u, r and t. Now, if we apply Grönwall's lemma to the flow of characteristics of the velocity field, we find that $\mu_t^{(r)}$ is concentrated on $\{u \in \Omega : \mathrm{dist}(u, Q_{r_0}) \leq (r_0 + C_1/C_2)e^{tC_2}\}$ and thus, for all $T > 0$ there exists $r > 0$ such that $t_r > T$. Hence $\lim_{r\to+\infty} t_r = +\infty$ and the Wasserstein gradient flow is uniquely well-defined on $[0, T)$ for T arbitrary large. □

Note that the condition on the initialization is automatically satisfied in Proposition 3.2.5, since for a particle gradient flow the initial measure has a finite discrete support and, therefore, it is contained in any Q_r for $r > 0$ large enough. Now, we state a useful lemma that allows us to represent the Wasserstein gradient flow as the pushforward of the initial measure μ_0 by the flow of the velocity fields, which we used at the end of the previous proof.

Lemma 3.2.10. *Under the assumptions made in Proposition 3.2.9, let $(\mu_t)_{t\geq 0}$ be the Wasserstein gradient flow of $\tilde{F}$ and $(v_t)_t$ the associated velocity fields. Consider the flow $X : \mathbb{R}_+ \times \Omega \to \Omega$ that for all $u \in \Omega$, is an absolutely continuous*

solution of

$$\begin{cases} \partial_t X(t,u) = v_t(X(t,u)) & \text{for a.e. } t \geq 0, \\ X(0,u) = u. \end{cases}$$

Then X is uniquely well-defined and continuous, $X(t,\cdot)$ is Lipschitz on Q_r uniformly on compact time intervals for all $r > 0$ and $\mu_t = (X_t)_{\#}\mu_0$.

Now, we are ready to characterize the many-particle limit of classical gradient flows as the unique Wasserstein gradient flow of $\tilde{F}$. Recalling that each particle represents a single neuron of our shallow neural network, this limit corresponds to studying the training dynamics as the network's width goes to infinity. Thus, we are going to prove that the classical gradient descent used to train a finite network converges to a continuous Wasserstein gradient flow in the infinite-width regime.

Theorem 3.2.11 (**Many-particle limit**)**.** *Under the assumptions made in Proposition 3.2.5, consider $\{t \mapsto U(t) = (u_1(t), \dots, u_N(t))\}_{N\in\mathbb{N}}$ a sequence of gradient flows of $\tilde{F}_N$ initialized in a set $Q_{r_0} \subset \Omega$. If $\mu_{N,0} \to \mu_0 \in \mathcal{P}_2(\Omega)$ for the Wasserstein distance W_2, then $(\mu_{N,t})_t \to (\mu_t)_t$ as $N \to +\infty$, where $(\mu_t)_t$ is the unique Wasserstein gradient flow of $\tilde{F}$ starting from μ_0.*

Proof. While we could rely on stability results for Wasserstein gradient flows [15, Thm.11.2.1 (Stability)], we will follow the proof shown in [12], which is direct and gives an independent argument for the existence of Wasserstein gradient flows, distinct from the standard one : it involves a discretization in space instead of the classical discretization in time.

Step 1. First, we show that, at least on a small interval $[0, t_r]$, the paths are contained in Q_r for some $r > r_0$. Let us introduce t_r, the first exit time from Q_r

$$t_r := \inf\{t > 0 : \exists N \in \mathbb{N} \quad \text{s.t. } \mu_{N,t}(Q_r) < 1\}.$$

In order to show that $t_r > 0$, it is sufficient to bound the velocity of individual particles $u_n(t)$ before t_r. Consider $L_{V,r}$ the Lipschitz constant of V on Q_r. By Proposition 3.2.5, for each particle in Q_r, we have

$$\begin{aligned} |u_n'(t)| = |v_t(u_n(t))| &\leq |\tilde{v}_t(u_n(t))| + |\partial V(u_n(t))| \\ &\leq \|dR\|_{\infty,r}\|d\varphi\|_{\infty,r} + L_{V,r}. \end{aligned}$$

For each particle starting from Q_{r_0}, the minimum travel distance required to leave Q_r is $r - r_0$. Hence, we have the following bound for the travel distance

$$\begin{aligned} 0 < r - r_0 \leq |u_n(t_r) - u_n(0)| &\leq \int_0^{t_r} |u_n'(s)| ds \\ &\leq t_r \cdot \left(\|dR\|_{\infty,r}\|d\varphi\|_{\infty,r} + L_{V,r}\right), \end{aligned}$$

which means that $t_r > 0$.

Step 2. Now we want to prove the existence of a limit curve $[0, t_r] \ni t \mapsto \mu_t \in$

$\mathcal{P}_2(\Omega)$. Considering $0 \le t_1 < t_2 \le t_r$, we have the bound

$$W_2(\mu_{N,t_1}, \mu_{N,t_2})^2 \le \frac{1}{N}\sum_{n=1}^{N} |u_n(t_2) - u_n(t_1)|^2 = \frac{1}{N}\sum_{n=1}^{N} \left| \int_{t_1}^{t_2} u_n'(s)\, ds \right|^2$$
$$\le \frac{t_2 - t_1}{N} \sum_{n=1}^{N} \int_{t_1}^{t_2} |u_n'(s)|^2 ds = (t_2 - t_1) \int_{t_1}^{t_2} -\frac{d}{ds}\tilde{F}(\mu_{N,s})\, ds,$$

where we matched each particle at t_1 with its future position t_2 in the first inequality, and we used Jensen's inequality with the energy decreasing condition (3.17) for a particle gradient flow in the second line. Therefore, it follows

$$W_2(\mu_{N,t_1}, \mu_{N,t_2}) \le \sqrt{t_2 - t_1} \left(\sup_N \tilde{F}(\mu_{N,0}) - \inf_{\mu \in \mathcal{P}(\mathbb{R}^{d+2})} \tilde{F}(\mu) \right)^{1/2}$$

and thus the family of curves $(t \mapsto \mu_{N,t})_{N\in\mathbb{N}}$ is equicontinuous in W_2 on $[0, t_r]$, uniformly in N. Moreover, for all $t \in [0, t_r]$, the family $(\mu_{N,t})_N$ lies in a W_2 ball, which is narrowly precompact thanks to Theorem 2.1.9 (Prokhorov) (but a priori not W_2 precompact, since the narrow topology is weaker than the topology of W_2). By the Ascoli-Arzelà theorem, we can extract a subsequence converging narrowly to a curve $(\mu_t)_{t\ge 0}$ continuous in the narrow topology, which is concentrated in Q_r at all time. We also have uniform convergence in the **bounded Lipschitz metric**, which, for any couple of measures $\mu, \nu \in \mathcal{M}(\Omega)$, is defined as

$$d_{BL}(\mu, \nu) := \sup_{\substack{\|f\|_\infty \le 1 \\ \mathrm{Lip}(f) \le 1}} \left| \int f\, d\mu - \int f\, d\nu \right| \tag{3.29}$$

and metrizes the narrow convergence of probability measures. We can also define the Bounded Lipschitz norm of $\mu \in \mathcal{M}(\Omega)$, as

$$\|\mu\|_{BL} := \sup_{\substack{\|f\|_\infty \le 1 \\ \mathrm{Lip}(f) \le 1}} \left\{ \int f\, d\mu \right\}$$

where $Lip(f)$ is the smallest Lipschitz constant of f. In the sequel, we only consider this subsequence, still denoted by $(\mu_N)_N$.

Step 3. Now we show that the limit curve $(\mu_t)_{t\ge 0}$ satisfies a continuity equation such as (3.21). Consider the velocity fields $v_{N,t}$ associated with the measures $\mu_{N,t}$, defined in (3.18), and define v_t the analog of the limit curve $(\mu_t)_t$. We want to show that the sequence $(E_N)_N$ of momenta, the vector valued measures on $[0, t_r] \times \Omega$ defined by $E_N := v_{N,t}\mu_{N,t}dt$, converges narrowly to $E := v_t\mu_t$. Notice that these measures are also concentrated on Q_r. For any $\psi \in C_b([0, t_r] \times$

$\mathbb{R}^{d+2}; \mathbb{R}^{d+2}$), adding and subtracting $v_t d\mu_{N,t} dt$ in the differential, we have

$$\begin{aligned}\left|\int_{\Omega\times[0,t_r]} \psi \cdot d(E_N - E)\right| \leq \|\psi\|_\infty \int |v_{N,t}(u) - v_t(u)|\, d\mu_{N,t} dt \\ + \left|\int \psi \cdot v_t \, d(\mu_{N,t} - \mu_t) dt\right|.\end{aligned} \tag{3.30}$$

First, we prove that the first term in (3.30) tends to 0 when $N \to +\infty$. Since all $(\mu_{N,t})_{N,t}$ are concentrated on Q_r, it is sufficient to show that the sequence of velocity fields $(t,u) \mapsto v_{N,t}(u)$ converges uniformly on $[0,t_r] \times Q_r$ to $(t,u) \mapsto v_t(u)$, when $N \to +\infty$. Since a projection on a convex set is 1-Lipschitz, we have

$$|v_{N,t}(u) - v_t(u)| \leq 2|\tilde{v}_{N,t}(u) - \tilde{v}_t(u)| \leq 2\|d\varphi\|_{\infty,r} \|dR[\Phi_{\mu_{N,t}}] - dR[\Phi_{\mu_t}]\|.$$

Moreover, for all $t \in [0,t_r]$, we have

$$\begin{aligned}\|dR[\Phi_{\mu_{N,t}}] - dR[\Phi_{\mu_t}]\| &\leq \|dR\|_{\infty,r} \cdot \left\| \int \varphi_u d\mu_{N,t}(u) - \int \varphi_u d\mu_t(u) \right\| \\ &\leq \|dR\|_{\infty,r} \cdot \sup_{f \in L^2(D), \|f\| \leq 1} \int \langle f, \varphi_u \rangle \, d(\mu_{N,t} - \mu_t)(u) \\ &\leq \|dR\|_{\infty,r} \cdot \max\{\|\varphi\|_{\infty,r}, \|d\varphi\|_{\infty,r}\} \cdot \|\mu_{N,t} - \mu_t\|_{BL}\end{aligned}$$

where $\|\varphi\|_{\infty,r} = \sup_{u \in Q_r} \|\varphi_u\|_{L^2(D)}$ and we used the dual characterization of the norm in the second inequality. Since $(t \mapsto \mu_{N,t})_N$ uniformly converges to $t \mapsto \mu_t$ in the Bounded Lipschitz norm, this proves uniform convergence of the velocity fields and the convergence of the first term in (3.30) to 0. The second term in (3.30) also converges to 0, because $(t,u) \mapsto \psi(t,u) \cdot v_t(u)$ is continuous and bounded. Hence, $E_N \xrightarrow{N\to\infty} E$ narrowly and, therefore, the continuity equation is satisfied in the limit. As $(v_t)_t$ is bounded on Q_r, uniformly in time, we have $\int_0^{t_r} \int_\Omega |v_t(u)|^2 \, d\mu_t(u) dt < +\infty$ which proves, by [15, Thm.8.3.1], that $(\mu_t)_t$ is absolutely continuous in W_2.

Step 4. So far, we have shown the convergence, up to a subsequence, to a Wasserstein gradient flow on $[0,t_r]$: it remains to show that $\lim_{t\to+\infty} t_r = +\infty$. Since $\tilde{F}$ is proper and continuous for W_2, we have $\tilde{F}(\mu_{N,0}) \to \tilde{F}(\mu_0)$. We also know, by Proposition 3.2.5, that all paths $(\mu_{N,t})_t$ decrease monotonically the value of $\tilde{F}$. Therefore, everything lies in a sublevel of R, where dR is bounded. So, we can conclude with the same argument shown in the end of the proof of Proposition 3.2.9. The theorem follows by combining this result with the uniqueness obtained in Proposition 3.2.9. □

Remark 23. Remember that in Proposition 3.2.1, we proved the existence of the infinite-width limit of a sequence of shallow neural networks with simple assumptions without considering the training dynamics, but we could not ensure the uniqueness of the limit. Thanks to Theorem 3.2.11, we can ensure uniqueness by using the training dynamics and more technical assumptions.

3.3 Convergence to global minimizers

As can be seen from Definition 3.2.6, a probability measure $\mu \in \mathcal{P}_2(\Omega)$ is a stationary point of a Wasserstein gradient flow if and only if $0 \in \partial \tilde{F}'(\mu)(u)$ for μ-a.e. $u \in \Omega$. In [17] it is proven that these stationary points are, in some cases, optimal over probability measures that have a smaller support. However, in general they are not global minimizers of $\tilde{F}$ over $\mathcal{M}_+(\Omega)$, even if R is convex. Note that a Wasserstein gradient flow is only defined in $\mathcal{P}_2(\Omega)$ to preserve optimal transport, but the actual optimization landscape of neural networks lacks total mass constraints, making $\mathcal{M}_+(\Omega)$ the natural domain for the global problem. Such global minimizers are indeed characterized as follows.

Proposition 3.3.1. *Consider a convex loss function R. A measure $\mu \in \mathcal{M}_+(\Omega)$ such that $\tilde{F}(\mu) < +\infty$ minimizes $\tilde{F}$ on $\mathcal{M}_+(\Omega)$ if and only if*

1. *$\tilde{F}'(\mu) \geq 0$ in Ω*
2. *$\tilde{F}'(\mu)(u) = 0$ for μ-a.e. $u \in \Omega_+$, where $\Omega_+ = \Omega \setminus \Omega_0$ and Ω_0 is the largest open set such that $\mu(\Omega_0) = 0$.*

Proof. Recall that, by Remark 20, we have that for all $\mu, \sigma \in \mathcal{M}(\Omega)$ such that $\tilde{F}(\mu), \tilde{F}(\sigma) < +\infty$, it holds $\int_\Omega |\tilde{F}'(\mu)(u)| d\sigma(u) < +\infty$ and

$$\frac{d}{d\varepsilon}\tilde{F}(\mu+\varepsilon\sigma)|_{\varepsilon=0} = \int_\Omega \tilde{F}'(\mu)(u)\, d\sigma(u) \quad \text{with} \quad \tilde{F}'(\mu) : u \mapsto \langle dR[\Phi_\mu], \varphi_u \rangle + V(u).$$

Let $\mu, \nu \in \mathcal{M}_+(\Omega)$ be such that $\tilde{F}(\mu), \tilde{F}(\nu) < +\infty$, consider $\sigma := \nu - \mu$ and its Lebesgue decomposition $\sigma = f\mu + \sigma^\perp$ where $f \in L^1(\mu), \sigma^\perp \in \mathcal{M}_+(\Omega)$ is singular to σ (see [16, Thm. 4.3.2]). Therefore, if μ is a minimizer of $\tilde{F}$ in $\mathcal{M}_+(\Omega)$, we have

$$0 \leq \frac{d}{d\varepsilon}\tilde{F}(\mu + \varepsilon(f\mu + \sigma^\perp))|_{\varepsilon=0} = \int_\Omega \tilde{F}'(\mu)(u)\, d\sigma(u),$$

thus we must have $\tilde{F}'(\mu) \geq 0$ and the equality when $\tilde{F}'(\mu)(u) = 0$ for μ-a.e. $u \in \Omega_+$. Conversely, since R is convex, for all $t \in [0,1]$, we have

$$\begin{aligned} 0 \leq \frac{d}{dt}\tilde{F}(\mu + t\sigma)|_{t=0} &= \frac{d}{dt}\tilde{F}((1-t)\mu + t\nu)|_{t=0} \\ &\leq \frac{d}{dt}((1-t)\tilde{F}(\mu) + t\tilde{F}(\nu))|_{t=0} = \tilde{F}(\nu) - \tilde{F}(\mu), \end{aligned}$$

and, therefore, μ is a minimizer of $\tilde{F}$ in $\mathcal{M}_+(\Omega)$. □

Remark 24. Note that if R is convex, then $\tilde{F} : \mathcal{M}_+(\Omega) \to \mathbb{R}$ is convex. Indeed, for any measure $\mu \in \mathcal{M}_+(\Omega)$ and for all $t \in [0,1]$ we have

$$\begin{aligned} \tilde{L}((1-t)\mu + t\nu) = R(\Phi_{(1-t)\mu+t\nu}) &= R\left((1-t)\int_\Omega \varphi_u d\mu(u) + t\int_\Omega \varphi_u d\nu(u)\right) \\ &\leq R((1-t)\Phi_\mu + t\Phi_\nu) \\ &\leq (1-t)R(\Phi_\mu) + tR(\Phi_\nu) = (1-t)\tilde{L}(\mu) + t\tilde{L}(\nu) \end{aligned}$$

and we also have

$$\begin{aligned}\tilde{V}((1-t)\mu + t\nu) &= \int_\Omega V(u)\, d((1-t)\mu + t\nu)(u) \\ &= (1-t)\int_\Omega V(u)\, d\mu + t\int_\Omega V(u)\, d\nu(u) \\ &= (1-t)\tilde{V}(\mu) + t\tilde{V}(\nu).\end{aligned}$$

Hence, $\tilde{F} = \tilde{L} + \tilde{V}$ is convex and this justifies the end of the proof of Proposition 3.3.1.

Despite these strong differences between stationarity and global optimality, we show in this section that Wasserstein gradient flows converge to global minimizers, under two main assumptions:

- the neuronal map φ and the potential term V must share a homogeneity direction (see Definition 3.3.2 below), and
- the support of the initialization of the Wasserstein gradient flow satisfies a "separation" property. This is preserved throughout the dynamic and, combined with homogeneity, allows to escape from neighborhoods of non-optimal points.

To introduce these assumptions, we give the following definition.

Definition 3.3.2. *(**Homogeneity**) Given a vector space $\mathcal{V}$. A function $f : \mathbb{R}^{d+1} \to \mathcal{V}$ is said **positively p-homogeneous**, with $p \geq 0$ if for all $x \in \mathbb{R}^{d+1}$ and $\lambda > 0$ it holds*

$$f(\lambda x) = \lambda^p f(x). \tag{3.31}$$

Example 6. If we consider the ReLU activation function $\sigma(s) = \max\{0, s\}$, the corresponding neuronal map $\varphi : u \in \mathbb{R}^{d+2} \mapsto \varphi_u \in L^2(D)$ is 2-homogeneous. To show this fact, we define $\tilde{\theta} := (\theta, b)$, so we can rewrite $u = (w, \tilde{\theta})$ and

$$\varphi_u : x \mapsto w\, \sigma((x, 1) \cdot \tilde{\theta}).$$

We conclude seeing that, for all $\lambda > 0$ and for all $x \in D$, we have

$$\begin{aligned}\varphi_{\lambda u}(x) &= \lambda w\, \sigma((x, 1) \cdot (\lambda\tilde{\theta})) \\ &= \lambda^2 w\, \sigma((x, 1) \cdot \tilde{\theta}) = \lambda^2 \varphi_u.\end{aligned}$$

If we consider the sigmoid activation function $\sigma(s) = \frac{1}{1+e^{-s}}$, the corresponding neuronal map $\varphi : u \in \mathbb{R}^{d+2} \mapsto \varphi_u \in L^2(D)$ is partially 1-homogeneous, since we have homogeneity only with respect to the parameter w:

$$\varphi_{\lambda u}(x) = \lambda w\, \sigma(\lambda(\theta \cdot x\ + b)).$$

We recall that we denote the "collection" of parameters of a single neuron $x \mapsto \varphi_u(x) = w\, \sigma(\theta \cdot x + b)$ by $u = (w, \theta, b) \in \mathbb{R}^{d+2}$ and, in the previous section, we considered a general closed convex subset $\Omega \subset \mathbb{R}^{d+2}$ that contains

all parameters of the network. For this discussion, we must give a more precise structure to Ω to ensure that φ and V share a direction of homogeneity. We will only see the 2-homogeneous and the partially 1-homogeneous case, following the framework shown in [12]. This is motivated by Example 6, since ReLU and sigmoid activation functions are standard in many applications. Moreover, the authors in [12] proved their results only in these two settings. In the 2-homogeneous case, a rich structure emerges, where the $(d+1)$-dimensional sphere $\mathbb{S}^{d+1} \subset \mathbb{R}^{d+2}$ plays a special role. Therefore, we make the following assumptions, considering a convex loss function R as we did in the results of the previous section.

Assumptions 1. *The domain is $\Omega = \mathbb{R}^{d+2}$ and φ is differentiable with $d\varphi$ locally Lipschitz and V is λ_V-convex. V and φ are both positively 2-homogeneous. Moreover,*

(i) (smooth convex loss) The loss R is convex, differentiable with dR Lipschitz on bounded sets and bounded on sublevel sets,

(ii) (Sard-type regularity) For all $f \in L^2(D)$, the set of regular values of $\mathbb{S}^{d+1} \ni u \mapsto \langle f, \varphi_u \rangle + V(u)$ is dense in its range (it is in fact sufficient that this holds for functions of the form $f = dR[\Phi_\mu]$ for some $\mu \in \mathcal{M}_+(\Omega)$).

Remark 25. Taking the balls of radius $r > 0$ as the family $(Q_r)_{r>0}$, these assumptions imply the assumptions made in Proposition 3.2.5. It is clear that our quadratic loss $R : \Phi \mapsto \|\Phi - y\|^2_{L^2(D)}$ satisfies Assumption 1-(i), but we will state the results in a more general context, as we did in the previous section. Assumption 1-(ii) substantially means that we are requiring the validity of the thesis of Sard's Theorem without verifying the hypothesis. The classical Sard's Theorem states that the set of critical values of a function $G \in C^k(M, N)$ has zero Lebesgue measure if $k \geq \max\{1, m-n+1\}$, where M, N are smooth manifolds of dimensions (respectively) m, n and $y \in N$ is a critical value of G if there exists $x \in M$ such that the differential $dG_x : T_xM \to T_yN$ is not surjective. In practice, modern neural network architectures rarely satisfy the requirements of Sard's Theorem, since even C^1 regularity is unattainable when employing standard activation functions like ReLU, making the higher-order smoothness required by Sard's Theorem in high-dimensional spaces structurally impossible. This Sard-type regularity has become a standard in some machine learning literature and an interested reader can find a more exhaustive discussion of this assumption in [20]. However, this theoretical gap is naturally bridged in algorithmic implementations: Stochastic Gradient Descent (SGD) noise introduces a diffusive behavior that allows the system to escape saddle points and flat regions, rendering the assumption superfluous [21, 22].

Now, we state the first global convergence result. It involves a condition on the initialization of the Wasserstein gradient flow and a separation property that can only be satisfied in the many-particle limit. In an ambient space Ω, we say that a set C separates the sets A and B if any continuous path in Ω with endpoints in A and B intersects C.

Theorem 3.3.3. *Under Assumptions 1, let $(\mu_t)_{t\geq 0}$ be a Wasserstein gradient flow of $\tilde{F}$ such that, for some $0 < r_a < r_b$, the support of μ_0 is contained in the ball of radius r_b centered in $0 \in \mathbb{R}^{d+2}$, i.e. $B(0, r_b)$, and separates the spheres $r_a\mathbb{S}^{d+1}$ and $r_b\mathbb{S}^{d+1}$. If $(\mu_t)_t$ converges to μ_∞ in W_2 when $t \to \infty$, then μ_∞ is a global minimizer of $\tilde{F}$ over $\mathcal{M}_+(\Omega)$. In particular, if $(U^{(N)}(t))_{N\in\mathbb{N}}, t \geq 0$ is a sequence of gradient flows for the functionals $\tilde{F}_N$, initialized in $B(0, r_b)$ such that $\mu_{N,0}$ weakly converges to μ_0, then (limits can be interchanged)*

$$\lim_{t\to\infty}\lim_{N\to\infty} \tilde{F}(\mu_{N,t}) = \lim_{N\to\infty}\lim_{t\to\infty} \tilde{F}(\mu_{N,t}) = \min_{\mu\in\mathcal{M}_+(\Omega)} \tilde{F}(\mu). \tag{3.32}$$

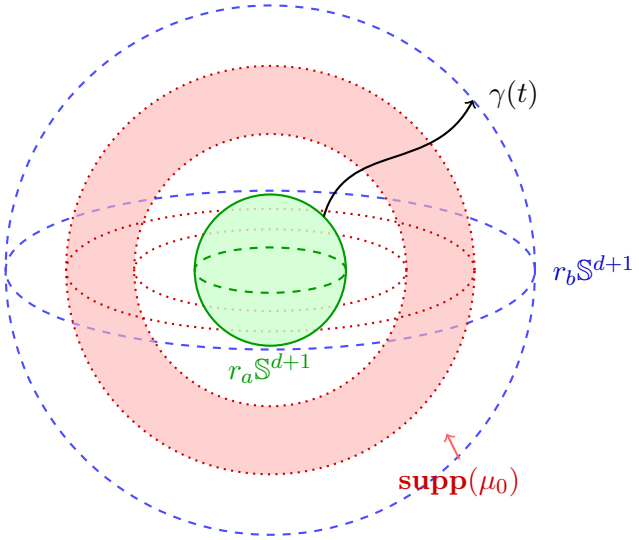

Figure 3.3: Geometric representation of the topological separation property required by Theorem 3.3.3, illustrated on a 2D cross-section. Any continuous path $\gamma(t)$ connecting the inner sphere $r_a\mathbb{S}^{d+1}$ to the outer sphere $r_b\mathbb{S}^{d+1}$ must necessarily intersect the support of the initial measure $\text{supp}(\mu_0)$.

Remark 26. A proof of Theorem 3.3.3 and stronger statements are presented in [12, Appendix C]. There, the authors give a criterion for Wasserstein gradient flows to escape neighborhoods of non-optimal measures, also valid in the finite-width setting, and then show that it is always satisfied by the flow defined above. They also weaken the assumption that μ_t converges to μ_∞ in W_2 when $t \to \infty$, showing that it is sufficient to have the weak convergence of a certain projection of μ_t. For the 2-homogeneous case, they consider the projection operator $h^2 : \mathcal{M}_+(\mathbb{R}^{d+2}) \to \mathcal{M}_+(\mathbb{S}^{d+1})$ which is characterized by the relationship, for all continuous and bounded functions $\psi : \mathbb{S}^{d+1} \to \mathbb{R}$ (extending the integrand to 0 when $u = 0$) and for all measures $\mu \in \mathcal{M}_+(\mathbb{R}^{d+2})$:

$$\int_{\mathbb{S}^{d+1}} \psi(\alpha) dh^2(\alpha) = \int_{\mathbb{R}^{d+2}} |u|^2 \psi\left(\frac{u}{|u|}\right) d\mu(u).$$

This operator is well-defined if and only if μ has finite second order moments. The authors in [12] assume the convergence of the Wasserstein gradient flow to

avoid other stronger requirements and obtain more reasonable assumptions.

Remark 27. Note that the ReLU activation function $\sigma(s) = \max\{0, s\}$ can be chosen in Theorem 3.3.3 only if we consider an alternative form of the neuronal map, since $d\varphi$ is discontinuous when φ is defined as in Example 6 (see [12, Lemma D.3]). To overcome this singularity without losing 2-homogeneity, the authors introduce an alternative differentiable parameterization of φ. This is achieved by dropping the explicit weight w and embedding the scaling factor directly into the internal parameters by applying the signed square function $s(t) = t|t| = \operatorname{sign}(t) \cdot t^2$ to each entry of $\tilde{\theta} = (\theta, b)$ (see [12, Lemma D.5]). This reparameterization ensures that $d\varphi$ is continuous, while the optimization dynamics of the Wasserstein gradient flow remain equivalent, so we can apply Theorem 3.3.3.

In the partially 1-homogeneous case, we have to make different assumptions.

Assumptions 2. *The domain is $\Omega = \mathbb{R} \times \Theta$ with $\Theta \subset \mathbb{R}^{d+1}$, $\varphi_u = w \cdot \phi(\tilde{\theta})$ and $V(u) = |w|\beta(\tilde{\theta})$ for all $u = (w, \theta, b) = (w, \tilde{\theta}) \in \mathbb{R} \times \Theta$, where $\phi : \Theta \to L^2(D)$ and $\beta : \Theta \to \mathbb{R}$ are bounded, differentiable with Lipschitz differential. In our case, for all $\tilde{\theta} \in \Theta$, $\phi(\tilde{\theta}) \in L^2(D)$ is defined as*

$$D \ni x \mapsto \sigma(\tilde{\theta} \cdot (x, 1)),$$

where σ is an activation function. Moreover,

(i) (smooth convex loss) The loss R is convex, differentiable with dR Lipschitz on bounded sets and bounded on sublevel sets,

(ii) (Sard-type regularity) For all $f \in L^2(D)$, the set of regular values of $g_f : \Theta \ni \tilde{\theta} \mapsto \langle f, \phi(\tilde{\theta})\rangle + \beta(\tilde{\theta})$ is dense in its range,

(iii) (boundary conditions) The function ϕ behaves nicely at the boundary of the domain : either

(a) $\Theta = \mathbb{R}^{d+1}$ and for all $f \in L^2(D)$, $\mathbb{S}^d \ni \tilde{\theta} \mapsto g_f(r\tilde{\theta})$ converges, uniformly in $C^1(\mathbb{S}^d)$ as $r \to \infty$, to a function satisfying the Sard-type regularity, or

(b) Θ is the closure of a bounded open convex set and for all $f \in L^2(D)$, g_f satisfies Neumann boundary conditions (i.e., for all $\tilde{\theta} \in \partial\Theta$, $d(g_f)_{\tilde{\theta}}(\nu_{\tilde{\theta}}) = 0$ where $\nu_{\tilde{\theta}} \in \mathbb{R}^{d+1}$ is the normal to $\partial\Theta$ at $\tilde{\theta}$).

Choosing the family of nested sets $Q_r := [-r, r] \times \Theta$, with $r > 0$, these assumptions imply the ones made in Proposition 3.2.5. The following theorem mirrors the statement of Theorem 3.3.3 in the partially 1-homogeneous case, but with a different condition on the initialization.

Theorem 3.3.4. *Under Assumptions 2, let $(\mu_t)_{t\geq 0}$ be a Wasserstein gradient flow of $\tilde{F}$ such that, for some $r_0 > 0$, the support of μ_0 is contained in $[-r_0, r_0] \times \Theta$ and separates $\{-r_0\} \times \Theta$ from $\{r_0\} \times \Theta$. If $(\mu_t)_t$ converges to μ_∞ in W_2 when $t \to \infty$, then μ_∞ is a global minimizer of $\tilde{F}$ over $\mathcal{M}_+(\Omega)$. In particular,*

if $(U^{(N)}(t))_{N\in\mathbb{N}}, t \geq 0$ *is a sequence of gradient flows for the functionals* $\tilde{F}_N$*, initialized in* $[-r_0, r_0] \times \Theta$ *such that* $\mu_{N,0}$ *converges to* μ_0 *in* W_2*, then*

$$\lim_{t\to\infty} \lim_{N\to\infty} \tilde{F}(\mu_{N,t}) = \lim_{N\to\infty} \lim_{t\to\infty} \tilde{F}(\mu_{N,t}) = \min_{\mu\in\mathcal{M}_+(\Omega)} \tilde{F}(\mu). \tag{3.33}$$

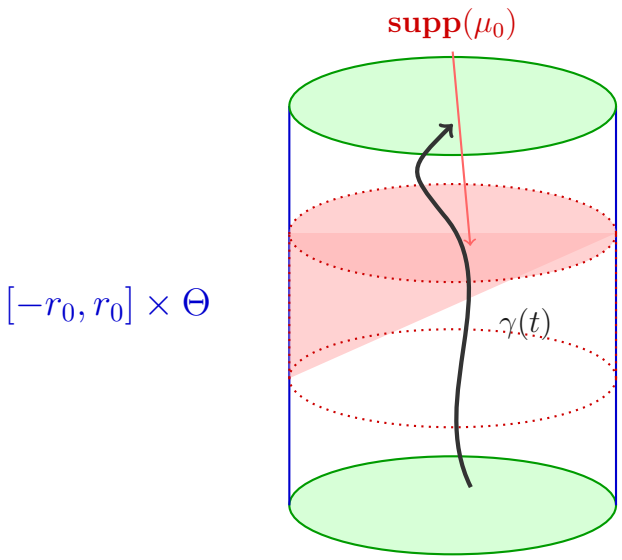

Figure 3.4: Geometric representation of the topological separation property required by Theorem 3.3.4. Any continuous path $\gamma(t)$ connecting the inferior base $\{-r_0\} \times \Theta$ to the superior base $\{r_0\} \times \Theta$ must necessarily intersect the support of the initial measure $\text{supp}(\mu_0)$.

Remark 28. The discussion in Remark 26 also applies to Theorem 3.3.4, by choosing another projection operator $h^1 : \mathcal{M}_+(\mathbb{R} \times \Theta) \to \mathcal{M}(\Theta)$ that is adapted to the partial 1-homogeneity of φ and V. It is characterized, for all continuous and bounded test functions $\psi : \Theta \to \mathbb{R}$ and for all measures $\mu \in \mathcal{M}_+(\mathbb{R} \times \Theta)$, by the relationship

$$\int_\Theta \psi(\tilde{\theta})\, dh^1(\mu)(\tilde{\theta}) = \int_{\mathbb{R}\times\Theta} w\, \psi(\tilde{\theta})\, d\mu(w, \tilde{\theta}).$$

This operator is well defined whenever $(w, \tilde{\theta}) \mapsto w$ is $\mu-$integrable. Furthermore, this partially 1-homogeneous setting naturally includes bounded activation functions that fail to be scale-invariant, such as the sigmoid function introduced in Example 6.

Remark 29. Recently, the authors in [23] found that one of the technical results [12, Proposition C.4] necessary to prove Theorem 3.3.4 is based on an incorrect claim and provided a correction that preserves its validity.

3.4 Numerical experiments

In this section, we will see some specific examples of the description of neural network training as a Wasserstein gradient flow. The general idea is to choose an objective function (or label function) $y \in L^2(D)$, which represents the training data set. Then we initialize the distribution of the network's parameters with $\mu_0 \in \mathcal{P}_2(\Omega)$ and use the JKO scheme to discretize the unique continuous Wasserstein gradient flow starting from μ_0. We have already seen the JKO scheme in Section 2.2, but, for our experiments, we need to adapt it to the case of neural network training. Given $\tau > 0$, the scheme becomes

$$\mu_\tau^k \in \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\Omega)} \left\{ \tilde{F}(\mu) + \frac{1}{2\tau} W_2^2(\mu, \mu_\tau^{k-1}) \right\}. \tag{3.34}$$

Here, we focus only on the 2-homogeneous case seen in the previous section, choosing a ReLU activation function with L^2-regularization. Hence, we consider $\tilde{F} : \mathcal{P}_2(\Omega) \to \mathbb{R}$ such that, for all $\mu \in \mathcal{P}_2(\Omega)$

$$\begin{aligned} \tilde{F}(\mu) = &\int_D \left| \int_\Omega w \max\{0, \theta \cdot x + b\} \, d\mu(w, \theta, b) - y(x) \right|^2 dx \\ &+ \int_\Omega \frac{\lambda}{2} |(w, \theta, b)|^2 \, d\mu(w, \theta, b), \end{aligned} \tag{3.35}$$

where $\lambda > 0$ is the regularization coefficient. In our experiments, we consider $d = 1$, so we have $(w, \theta, b) \in \mathbb{R}^3$ and $D \subset \mathbb{R}$, to show the evolution in time of both the neural network during its training and of its parameters. Calculating the exact W_2-distance to implement the JKO scheme (3.34) is very expensive in terms of computational cost and lacks the necessary smoothness for gradient-based optimization algorithms. Fortunately, the optimal transport problem can be regularized by an entropic term, using the so-called Sinkhorn algorithm and reducing the computational cost (see [24]). Given $\mu, \nu \in \mathcal{P}_2(\Omega)$, the idea behind this algorithm is to approximate $W_2^2(\mu, \nu)$ by solving the entropically regularized transport problem

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega \times \Omega} |x - y|^2 \, d\gamma(x, y) \; + \; \varepsilon \, H(\gamma \,|\, \mu \otimes \nu), \tag{3.36}$$

where $\varepsilon > 0$ is the "entropic" regularization parameter and $H(\,\cdot\,|\,\mu \otimes \nu)$ denotes the relative entropy with respect to the product measure $\mu \otimes \nu$. More precisely, for any coupling $\gamma \in \Gamma(\mu, \nu)$

$$\begin{aligned} H(\gamma \mid \mu \otimes \nu) &= -\operatorname{KL}(\gamma \parallel \mu \otimes \nu) \\ &= \begin{cases} -\displaystyle\int_{\Omega \times \Omega} \log\left(\frac{d\gamma}{d(\mu \otimes \nu)}(x, y) \right) d\gamma & \text{if } \gamma \ll \mu \otimes \nu, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{3.37}$$

where $\operatorname{KL}(\gamma \parallel \mu \otimes \nu)$ is the Kullback-Leibler divergence between γ and $\mu \otimes \nu$, which quantifies the proximity of two probability distributions (see [26]), and

$\frac{d\gamma}{d(\mu\otimes\nu)}$ denotes the Radon-Nikodym derivative of γ with respect to $\mu\otimes\nu$. Using a probabilistic interpretation, H quantifies how much γ is close to preserving the statistical independence between μ and ν. As $\varepsilon \to 0$, the entropically regularized problem converges to the standard optimal transport problem. Hence, in the Sinkhorn algorithm, we want ε to be small enough to approach the exact value of $W_2^2(\mu,\nu)$. Moreover, we also discretize the particle gradient flow using the batch gradient descent algorithm [30], to compare its dynamics with the Wasserstein gradient flow and highlight relevant differences between them.

3.4.1 Standard approximation

For the first experiment, we choose three different label functions

1. $y(x) = |x|$
2. $y(x) = x^2$
3. $y(x) = \sin(2\pi x) + \frac{1}{2}\cos(6\pi x)$.

Then, we simulate the corresponding Wasserstein gradient flow adding a comparison with the discretized particle gradient flow. For the first two functions, we fix the number of neurons $N = 2000$ and choose their uniform distribution in $\mathbb{S}^2$ as μ_0. This choice is motivated by Theorem 3.3.3, since we have $\text{supp}(\mu_0) \subset \mathbb{S}^2$. Clearly, this support cannot satisfy the separation hypothesis made in Theorem 3.3.3, because we are considering only a finite number of neurons. However, we would approximate the continuous dynamics by choosing N sufficiently large. Moreover, we use Adam optimizer algorithm (see [31]) to solve the minimization problem (3.34) for each JKO step. In the first row of

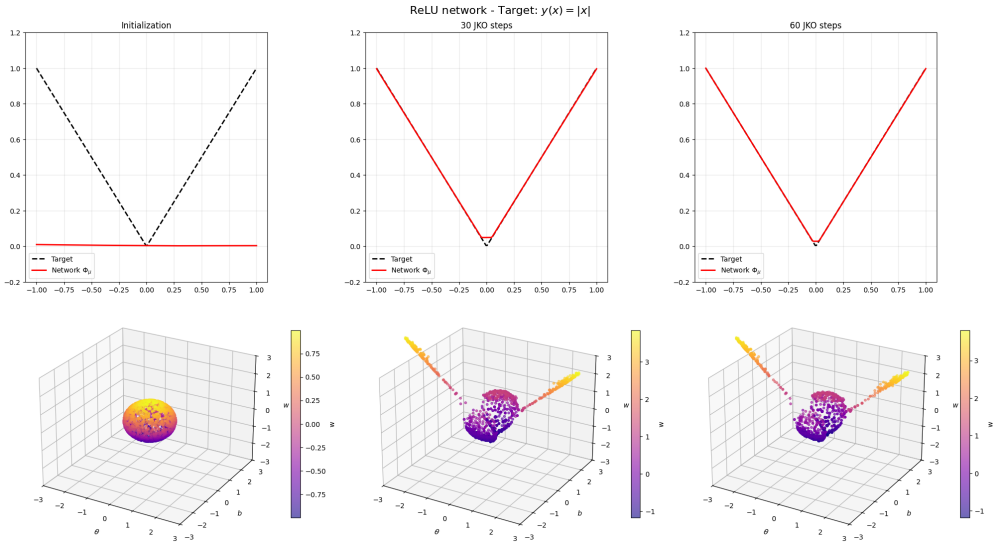

Figure 3.5: Evolution of the Wasserstein gradient flow for a ReLU network approximating the target function $y(x) = |x|$.

Figure 3.5, we can see how easily a ReLU network can approximate $y(x) = |x|$

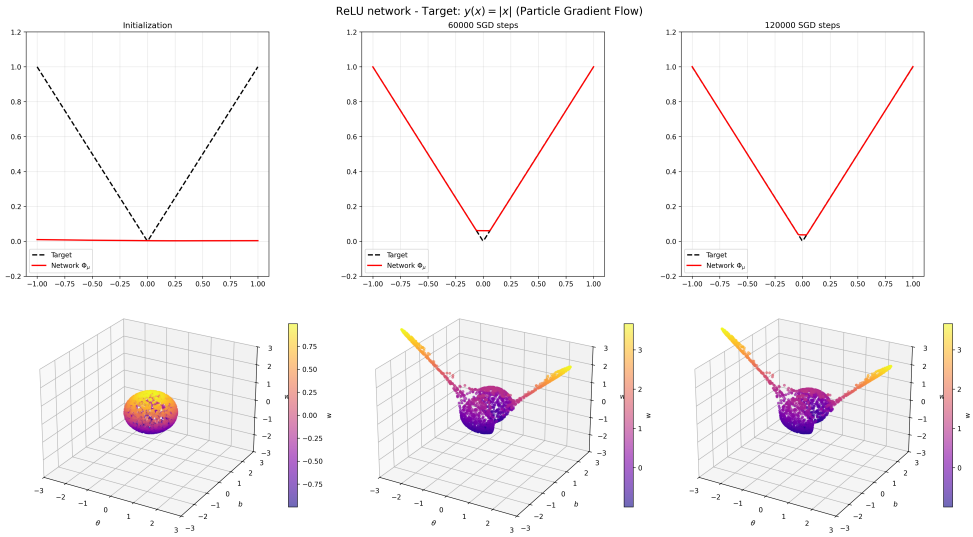

Figure 3.6: Evolution of the Particle gradient flow for a ReLU network approximating the target function $y(x) = |x|$.

in the first 30 JKO iteration steps. In the second row of the same figure, we can observe the evolution of the distribution of neurons during the training. To explain this behavior, consider the second distributional derivative of $\Phi_\mu \in L^2(D)$, calculated knowing that the second distributional derivative of $\max\{x, 0\}$ is the Dirac delta $\delta_0(x)$

$$\Phi''_\mu(x) = \int_\Omega w\,\theta^2 \delta_0(\theta x + b)\, d\mu(w, \theta, b). \tag{3.38}$$

Due to the Dirac delta, a neuron with parameters (w, θ, b) contributes to the curvature of Φ_μ only if $\theta x + b = 0$. Now, considering the target function $y = |x| = 2\,\max\{x, 0\} - x$ we can calculate its second distributional derivative and obtain

$$y''(x) = 2\,\delta_0(x), \tag{3.39}$$

so its curvature changes only if $x = 0$, as we already know. Hence, if we consider neurons with the parameter $\theta \neq 0$, the curvature of Φ_μ changes only in $x = -\frac{b}{\theta} = 0$ and all neurons must have $b = 0$. This forces neurons to be distributed on the plane $b = 0$ during training, but with this constraint alone we have infinitely many parameter configurations in the limit. The decisive role in shaping the distribution is given by the regularization term

$$\tilde{V}(\mu) = \frac{\lambda}{2} \int_\Omega \left[w^2 + \theta^2 + b^2\right]\, d\mu(w, \theta, b), \tag{3.40}$$

which introduces a constraint during the optimization process, forcing the distribution of neurons to the optimal configuration. We can think of $\tilde{V}$ as the potential energy of a harmonic oscillator with elastic constant λ, which guides

the particles to a stable equilibrium configuration. Looking at the second row

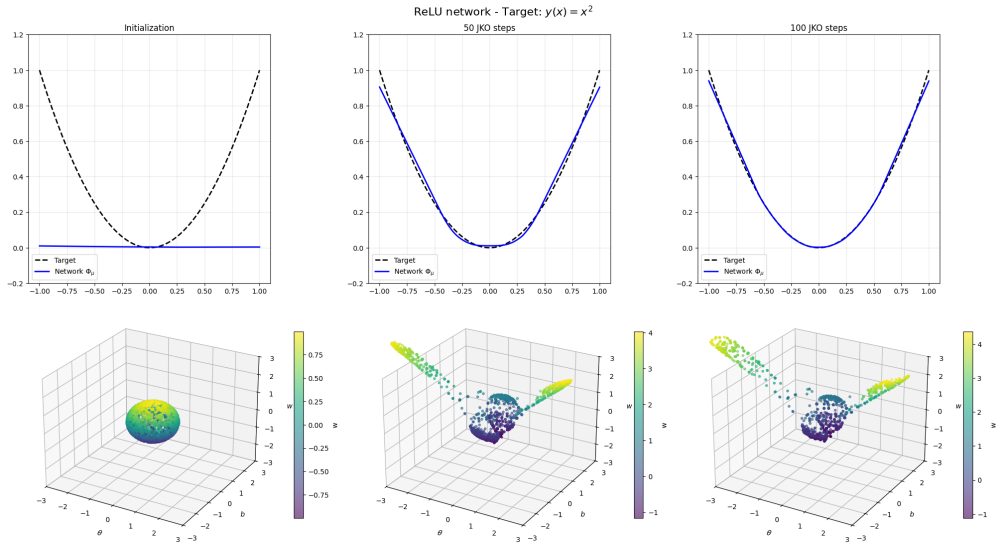

Figure 3.7: Evolution of the Wasserstein gradient flow for a ReLU network approximating the target function $y(x) = x^2$.

of Figure 3.6, we can see that the optimal shape of the support after training is substantially the same as in Figure 3.5, but neurons tend to spread more uniformly. We can imagine that in the infinite-width limit, both flows should be identical, according to Theorem 3.2.11 (Many-particle limit). In the first row of Figure 3.7, we can observe that our network requires more than 50 JKO iteration steps to approximate $y(x) = x^2$ with great precision. Here we have $y''(x) = 2 \neq 0$ for all $x \in D \subset \mathbb{R}$, so we cannot reduce the distribution of neurons in the plane $b = 0$, since the curvature of y is always positive (and constant). In Figure 3.8, we have the same behavior as observed in Figure 3.6. To approximate the third function, we also fix the number of neurons $N = 2000$, but we choose their uniform distribution in $r_0\mathbb{S}^2$, where $r_0 = 25$. This choice is motivated by the failure of the approximation when the radius of the support sphere of initialization is too small, since the network must assign larger values to $|w|$ and $|\theta|$ trying to learn high frequencies. To explain this phenomenon, consider the second derivative of the objective function

$$y''(x) = -4\pi^2 \sin(2\pi x) - 18\pi^2 \cos(6\pi x). \tag{3.41}$$

This function changes sign rapidly and reaches large absolute values. This complicates the learning process for a ReLU neural network, which struggles to capture these rapid changes in curvature. Therefore, moving neurons far from a standard initialization sphere ($\mathrm{supp}(\mu_0) \subset \mathbb{S}^2$) during training is too expensive. Hence, the choice of $r_0 = 25$ is just a tradeoff between the approximation accuracy and the observable motion of the neurons. The results are shown in Figures 3.9 and 3.10, which highlight the struggle of the network to capture high frequencies. Note that we display an early training iteration rather than

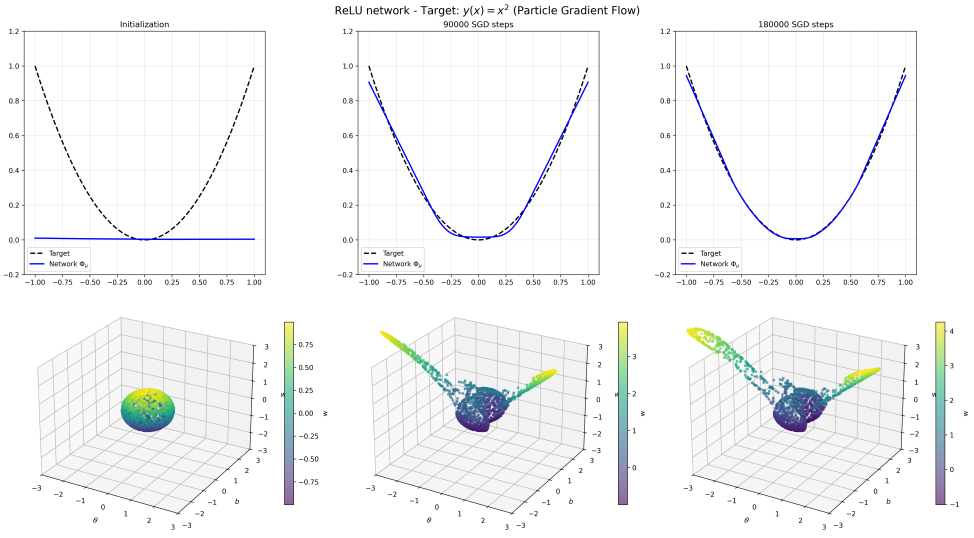

Figure 3.8: Evolution of the Particle gradient flow for a ReLU network approximating the target function $y(x) = x^2$.

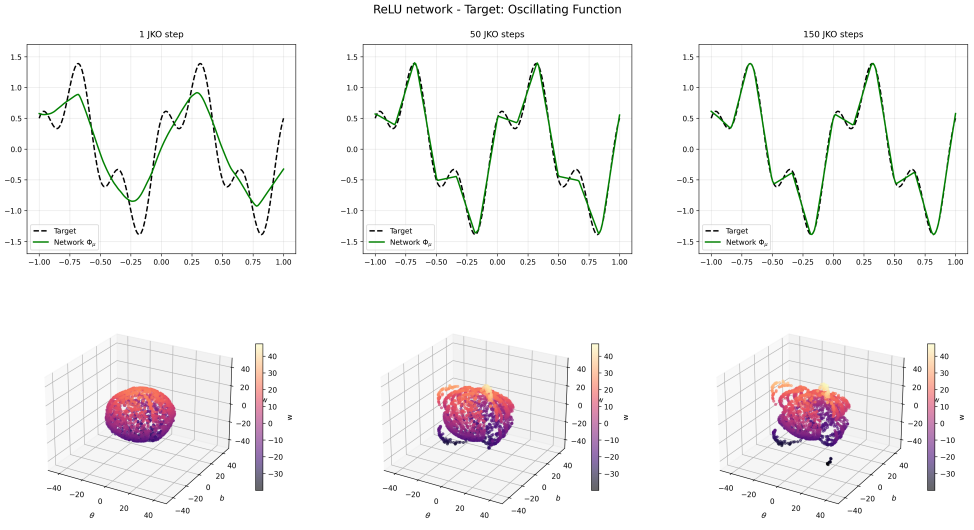

Figure 3.9: Evolution of the Wasserstein gradient flow for a ReLU network approximating the target function $y(x) = \sin(2\pi x) + \frac{1}{2}\cos(6\pi x)$.

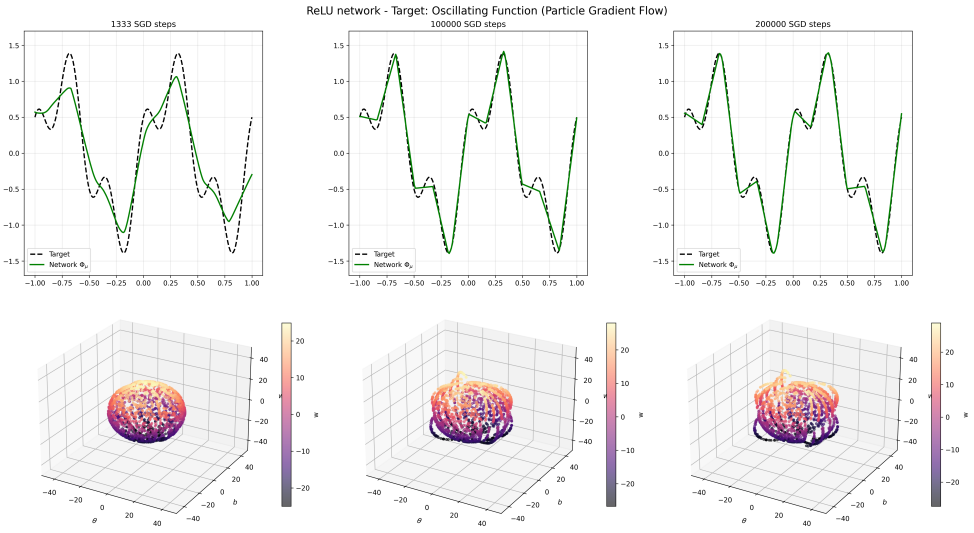

Figure 3.10: Evolution of the particle gradient flow for a ReLU network approximating the target function $y(x) = \sin(2\pi x) + \frac{1}{2}\cos(6\pi x)$.

the exact initialization since, due to the large initialization radius, the initial graph of the network would fall entirely out of view.

3.4.2 Approximation with noisy data

In the next experiment, we consider a quadratic label function with a Gaussian noise

$$y(x) = x^2 + \sigma Z, \tag{3.42}$$

where $Z \sim \mathcal{N}(0,1)$ and σ is the standard deviation of the noise. This choice reflects some applications, since it is common to collect noisy data. Here, we fix again $N = 2000$ and simulate both the approximation and the dynamics of the neurons. As we can observe in the first row of Figure 3.11, a ReLU network can fit the function correctly even with the extra noise and using less than 100 JKO iteration steps. In the second row, we can see that the distribution of the parameters is similar to what is shown in Figure 3.7, but with a visible dispersion between the particles along the two emerging directions after a sufficient number of JKO iterations steps. We can physically interpret this effect as a thermal agitation of the particle system (the neurons) caused by adding noise to the data set.

3.4.3 Transport dynamics with two neurons

For the final experiment, we want to observe the dynamics of a small network with only two neurons trying to approximate a ReLU target function. More precisely, we choose $y(x) = \max\{x-1, 0\}$ and we perform just 48 JKO iteration steps to simulate the Wasserstein gradient flow, since the target is very easy to approximate using a ReLU network. Figure 3.13 shows the evolution in time

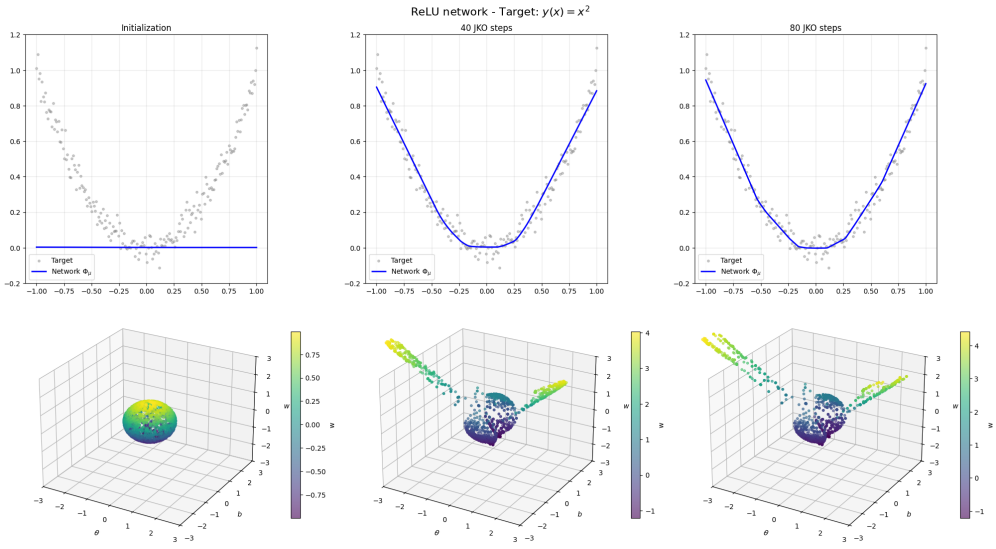

Figure 3.11: Evolution of the Wasserstein gradient flow for a ReLU network approximating a noisy target function $y(x) = x^2 + 0.05 \cdot Z$, where $Z \sim \mathcal{N}(0, 1)$.

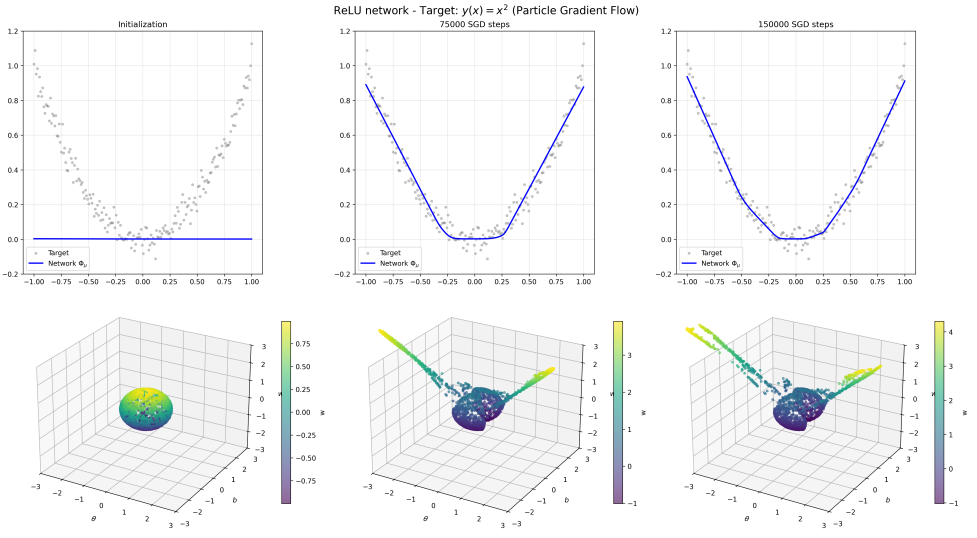

Figure 3.12: Evolution of the Particle gradient flow for a ReLU network approximating a noisy target function $y(x) = x^2 + 0.05 \cdot Z$, where $Z \sim \mathcal{N}(0, 1)$.

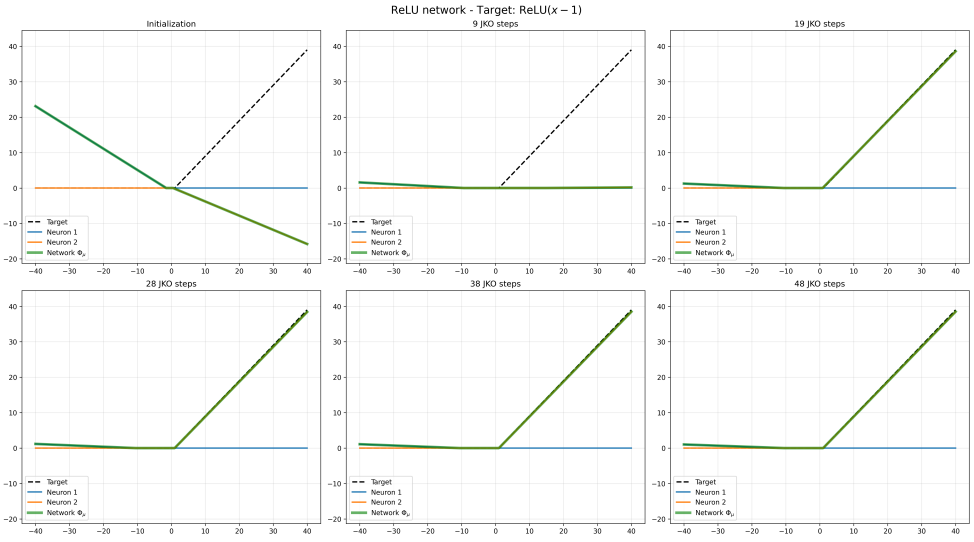

Figure 3.13: Evolution of a ReLU network with two neurons approximating the target function $y(x) = \max\{x - 1, 0\}$ using the JKO scheme.

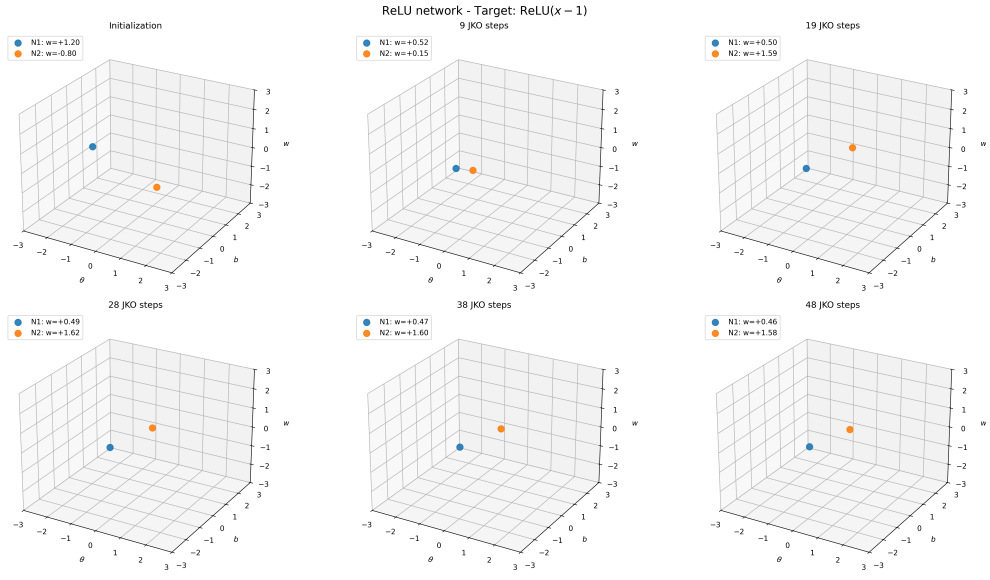

Figure 3.14: Evolution of the parameters of a ReLU network with two neurons approximating the target function $y(x) = \max\{x - 1, 0\}$ using the JKO scheme.

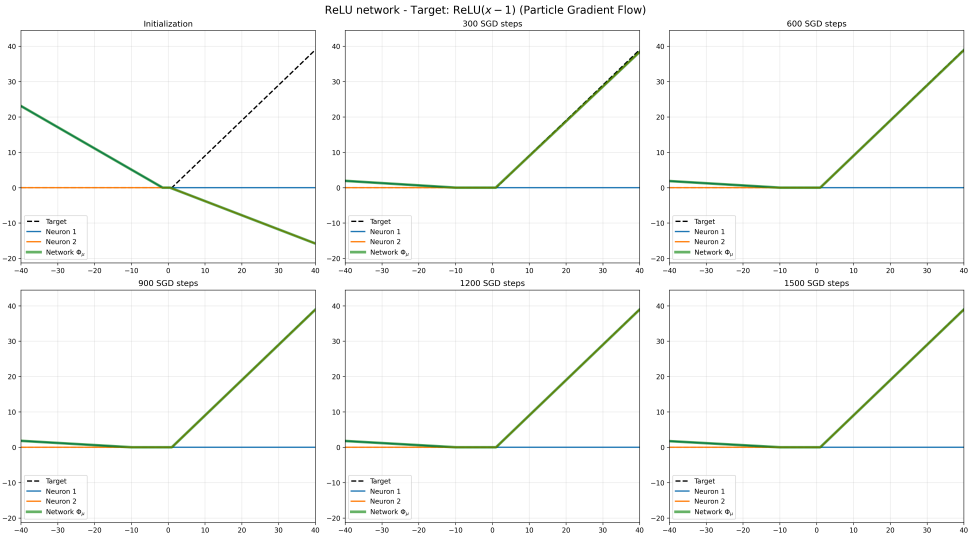

Figure 3.15: Evolution of a ReLU network with two neurons approximating the target function $y(x) = \max\{x - 1, 0\}$ using the batch gradient descent.

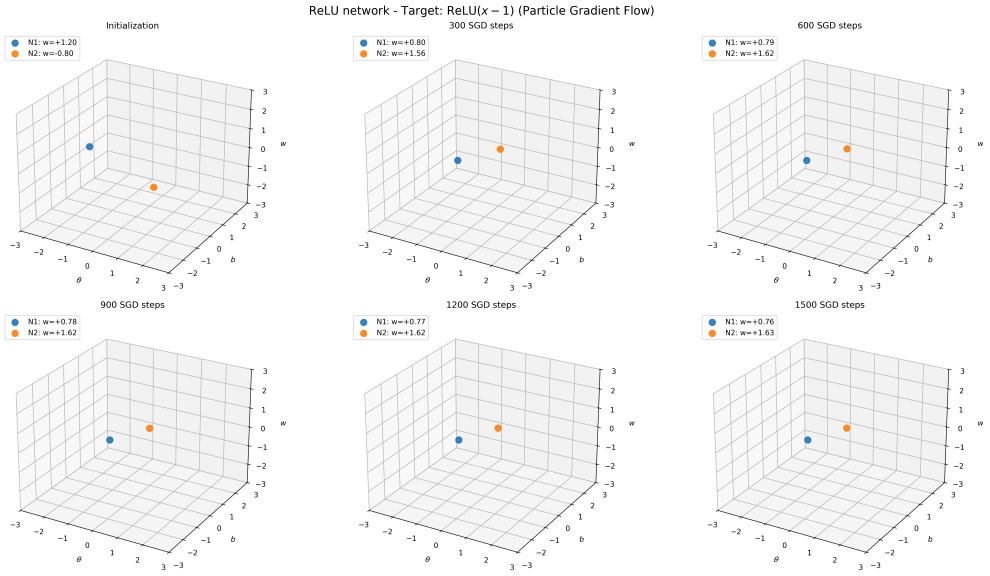

Figure 3.16: Evolution of the parameters of a ReLU network with two neurons approximating the target function $y(x) = \max\{x-1, 0\}$ using the batch gradient descent.

of the network and of individual neurons during the approximation. Moreover, Figure 3.14 illustrates the time evolution of the parameters of the two neurons. In Figures 3.15 and 3.16 we can observe the dynamics of training using gradient descent, which highlights differences in the trajectories of neurons during optimization, but this could be fully appreciated with an animation showing the paths of the particles after each step. More specifically, using gradient descent, we can see that neurons follow continuous trajectories in Euclidean space. In contrast, the JKO scheme solves an optimization problem in $\mathcal{P}_2(\Omega)$ with respect to the W_2 distance. Therefore, sometimes neurons can rapidly change their position in Ω between one iteration step and another, since W_2 is a non-Euclidean metric. [1]

[1] The source code and animations for all the experiments discussed in this thesis are available at `https://github.com/edward-f-buccieri/master-thesis-math-2026-buccieri`.

Conclusions

In conclusion, our main goal was to describe the dynamics of neural network training as a Wasserstein gradient flow. We started by introducing the theory of gradient flows in the Euclidean space $\mathbb{R}^N$, showing the results necessary to understand the generalization to the space of probability measures. Then, we gave an overview of the theory of optimal transport, focusing on the Wasserstein spaces and introducing the JKO scheme. The last chapter was entirely dedicated to the training of shallow neural networks, starting from its classical formulation as a gradient flow in $\mathbb{R}^N$ and reinterpreting it using probability measures. The key physical interpretation was thinking of the collection of neurons as a system of interacting particles driven by a gradient flow in Euclidean space. Moreover, we proved the existence and uniqueness of a Wasserstein gradient flow, starting from an initial distribution of the parameters of neurons μ_0. Then, we showed that the "particle gradient flow" converges to the unique Wasserstein gradient flow starting from μ_0, when the number of neurons goes to infinity, using the setting of [12]. Therefore, this result covered the main goal of our discussion, but we tried to understand when the Wasserstein gradient flow converges to a global minimizer of the functional, recalling the global convergence results shown in [12]. Then, we conducted our numerical experiments to show the dynamics of the training using the JKO scheme and the assumptions made on the support μ_0 in the global convergence results. In the experiments, we considered only the ReLU activation function, which is a standard choice in many applications and it is covered by one of the global convergence results (Theorem 3.3.3). To calculate the W_2-distance in the JKO scheme, reducing the computational cost, we used the Sinkhorn algorithm and saw how the distribution of the parameters evolves during training in different choices of the data set. We also discretized the particle gradient flow using the batch gradient descent algorithm to compare its dynamics with the Wasserstein gradient flow and pinpoint the relevant differences between them, observing that in the case of particle gradient flow, neurons tend to spread more uniformly during training. These experiments concluded our discussion and it is important to remember that we covered only the training of shallow neural networks. Using optimal transport to describe the training of deep neural networks is a natural extension of this thesis, but it remains an open problem in general, since the composition of layers means that the output of a neuron becomes the input of another neuron in the next layer. Even in the simplified case of deep linear networks, where this composition becomes a product of matrices, establishing infinite-width limits requires tools different from optimal transport [29]. Recent

works such as [27] and [28] make progress in this direction by considering architectures (infinitely deep ResNets and Transformers, respectively) whose specific structure makes them more tractable with optimal transport tools.

Bibliography

[1] Endre Süli, *Numerical Solution of Ordinary Differential Equations*, Mathematical Institute, University of Oxford, (2014).

[2] Aris Daniilidis, *Gradient Dynamical Systems, Tame Optimization and Application*, Spring School on Variational Analysis, Paseky nad Jizerou, Czech Republic (April 20-24, 2009).

[3] Ralph Tyrrell Rockafellar, *Characterization of the Subdifferentials of Convex Functions*, Pacific Journal of Mathematics, (1966).

[4] Bruck, R., *Asymptotic convergence of nonlinear contraction semigroups in Hilbert space*, J. Funct. Anal. **18** (1975), 15-26.

[5] Baillon, J.-B., *Un exemple concernant le comportemement asymptotique de la solution du problème* $\frac{du}{dt} + \partial\varphi(u) \ni 0$, J. Funct. Anal. **28** (1978), 369-376.

[6] Daniilidis, A., Ley, O. & Sabourau, S., *Asymptotic behaviour of self-contracted planar curves and gradient orbits of convex functions*, preprint 19 p., (UAB 31/08)

[7] Palis, J. & De Melo, W., *Geometric theory of dynamical systems. An introduction*, (Translated from the Portuguese by A. K. Manning), Springer-Verlag, New York-Berlin, 1982.

[8] A. Figalli, F. Glaudio, *An invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*, EMS Press, 2021.

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

[10] Philipp Peterson and Jakob Zech, *Mathematical Theory of Deep Learning*, Philipp Christian Petersen's Lab. (2024)

[11] F. Bach, *Learning Theory from First Principles* (Section 12.3), The MIT Press, 2024

[12] F. Bach, L. Chizat, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, Advances in Neural Information Processing Systems, 3036-3046 (2018)

[13] F. Bach, L. Chizat, *Gradient Descent on Infinitely Wide Neural Networks: Global Convergence and Generalization*, arXiv preprint arXiv:2110.08084 (2021)

[14] Filippo Santambrogio. {Euclidean, Metric, and Wasserstein} Gradient Flows : an overview. *Bulletin of Mathematical Sciences*, 7(1):87-154, 2017

[15] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in mertric spaces and in the space of probability measures.* Springer Science & Business Media, 2008.

[16] Donald L. Cohn. *Measure Theory*, volume 165. Springer, 1980.

[17] Atsushi Nitanda and Taiji Suzuki. *Stochastic particle gradient descent for infinite ensembles.* arXiv preprint arXiv:1712.05438v1, 2017.

[18] H. Cartan, *Differential Calculus*, Hermann, Paris, 1971.

[19] Ralph Abraham and Joel Robbin. *Transversal mappings and flows. WA Benjamin New York*, 1967.

[20] Stephan Wojtowytsch. *On the Convergence of Gradient Descent Training for Two-layer ReLU-networks in the Mean Field Regime.* arXiv preprint arXiv: 2005.13530v1, 2020

[21] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, Michael I. Jordan. *How to Escape Saddle Points Efficiently.* Proceedings of the 34th International Conference on Machine Learning (ICML), 2017.

[22] Wenqing Hu, Chris Junchi Li, Lei Li, Jian-Guo Liu. *On the diffusion approximation of nonconvex stochastic gradient descent.* Annals of Mathematical Sciences and Applications, 4(1), 3-32, 2019.

[23] Romain Petit, Clarice Poon, Gabriel Peyré. *On the global convergence of gradient descent for wide shallow models with bounded nonlinearities.* arXiv preprint arXiv:2605.10775, 2026.

[24] Marco Cuturi, *Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances.* arXiv preprint arXiv:1306.0895, 2013.

[25] Marcel Nutz, *Introduction to Entropic Optimal Transport*, Lecture notes, Columbia University, 2022.

[26] S. Kullback, R.A. Leibler, *On Information and Sufficiency*, The Annals of Mathematical Statistics, 22(1), pp. 79–86, 1951.

[27] R. Barboni, G. Peyré, F.-X. Vialard, *Understanding the training of infinitely deep and wide ResNets with Conditional Optimal Transport*, Communications on Pure and Applied Mathematics, 2025. arXiv: 2403.12887v2.

[28] G. Bruno, A. Agazzi, F. Pasqualotto, *A multiscale analysis of mean-field transformers in the moderate interaction regime*, 39th Conference on Neural Information Processing Systems (NeurIPS 2025), 2025. arXiv: 2509.25040.

[29] L. Chizat, M. Colombo, X. Fernández-Real, A. Figalli, *Infinite-width limit of deep linear neural networks*, Communications on Pure and Applied Mathematics, 77, 3958–4007, 2024 (arXiv:2211.16980)

[30] S. Ruder, *An overview of gradient descent optimization algorithms*, arXiv preprint, 2017 (arXiv:1609.04747)

[31] D. P. Kingma, J. Ba, *Adam: A Method for Stochastic Optimization*, arXiv preprint, 2014 (arXiv:1412.6980)