Notions in Optimal Transport for Sigmoid Neural Networks

A beginners' analysis of: *"On the Global Convergence of Gradient Descent for Over-Parameterized Models using Optimal Transport"* - Chizat, Bach

Simone Maria Giancola*†

†Bocconi University, Milan

January 8, 2023

Abstract

The following document is an exploration of the results of [CB18], written to understand better the content of the claims. The presentation is at this link. It is not an extension but rather an expansion of some of the elements needed for a less experienced reader. As this production is done in fulfillment of a semester exam for an Optimal Transport course, it does not cover all of the content, and was produced in more or less a month. The focus is on two layer sigmoid neural networks, and all the theoretical results needed to understand them. I also took inspiration from a video presentation of the publication [Ins19] and two blog posts by the authors [Bac20a; Chi20]. Works cited are in line with those of the authors, with some additional resources that I found helpful. Given the breadth of the subject, some of the content is left for future studies, but nothing less than the original publication is presented. I personally see this as a depth project, going very far into theoretical results to see the potential of Theory of neural networks. It is by no means an exposition of skills that I have 100% stored.

Section 1 paves the way for the research work proposed in a broad sense, introducing parametric optimization and the problem that will be studied, as well as a version of it that can be implemented. Section 2 shows how the formalism of Wasserstein Gradient Flows is instrumental to connect the two versions of the problem. Section 3 is the final theoretical contribution, with a characterization of the conditions thanks to which a global optimization is attained with the method considered. Lastly, in Section 4, it is shown that sigmoid neural networks can benefit from the results and be tuned to reach globally optimizing configurations, with satisfactory experimental results.

*simonegiancola09@gmail.com

Contents

1 Introduction

In this section we start by introducing some notation and providing a concise motivation for the subject of the document through brief examples. Moving to the content of the paper, we state the problem, its discretized version, and the assumptions. In order to proceed with the framework of Gradient Flows, it is also shown that it belongs to a wider class of functional optimization tasks by means of a lifting operation. This allows to reconciliate the particle problem presented with a wide collection of results.

Notation

The following is a broad summary of the notation used:

- in $\mathbb{R}^d$ scalar products $\cdot$, norms $|\cdot|$
- in a Hilbert space $\mathcal{F}$ scalar product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$
- norms of non-linear operators $\|\cdot\|$
- differential of f at x as df_x
- $\mathcal{M}(\mathbb{R}^d)$ the set of finite signed Borel measures on $\mathbb{R}^d$
- δ_x a dirac mass at x
- $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures with finite second moments endowed with Wasserstein distance (to see its construction and some properties, refer to Appendix A)

1.1 Parametric Supervised ML Optimization

Setting Assume our data sample is a **finite** size n collection of pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X} \subset \mathbb{R}^{d-2}$ and $y_i \in \mathcal{Y} \subset \mathbb{R}$. The two signals come from an unknown distribution $\rho(x, y)$. We aim to build a prediction function $h : \mathbb{R}^{d-2} \times \mathbb{R}^{d-1} \to \mathbb{R}$ parametrized by $\theta \in \mathbb{R}^{d-1}$. Such function $h(\cdot, \theta)$ is fitted against regularized empirical risk minimization:

$$\theta^* = \underset{\theta\in\mathbb{R}^{d-1}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h(x_i, \theta)) + \lambda\Xi(\theta) \tag{1.1}$$

where:

- $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a loss function
- $\Xi : \mathbb{R}^{d-1} \to \mathbb{R}_+$ is an (optional) regularization function
- λ (optional) is a *Lagrange coefficient* controlling for the impact of regularization

Since we observe a sample $\mathcal{D}$ of the underlying distribution $\rho(x, y)$ what we actually wish to mimic is a minimization of the test error wrt θ:

$$R : \mathcal{F} \to \mathbb{R}_+ \qquad R(h) = \mathbb{E}_{\rho(x,y)}\Big[\ell(y, h(x, \theta))\Big] \tag{1.2}$$

which is **in most reasonable cases** convex by the convexity of ℓ. Here, $\mathcal{F}$ is a Hilbert space (Def. A.15). We call this **expected risk**.

In this primordial analysis, we will focus on **linear** and **non linear** predictors, under which all parametric methods fall. Two trivial examples follow.

Example 1.3 (Advertising)**.** *In practice, we may want to understand which ad maximizes revenues. Datasets in this context have $n > 10^9$. What is commonly implemented is a* ***linear*** *predictor of the form:*

$$h(x, \theta) = \theta^T \Phi(x)$$

where $\Phi(x) \in \{0,1\}^{d-1}$ is a vector that stores information about navigation history and previous ads for a user. Here $d > 10^9$ as well, and $\Phi(\cdot)$ could be non-linear. The importance is ***linearity in the parameters****.*

Example 1.4 (Binary Image classification)**.** *Consider a dataset of images where $\mathcal{Y} = \{-1, 1\} \rightsquigarrow \{dogs, cats\}$. The sizes usually exceed $n, d > 10^6$. A neural network is implemented. It could be described as a* ***non linear*** *predictor with general form:*

$$h(x, \theta) = \theta_l^T \sigma(\theta_{l-1}^T \sigma(\dots \sigma(\theta_2^T \sigma(\theta_1^T x))))$$

Where l denotes the number of layers before the output and σ is a nonlinearity. Observe that the non-linearity is in the parameters in this case.

1.2 Why and What in few lines

A plethora of research questions have been solved when considering linear models of the form $h(x, \theta) = \theta^T \Phi(x)$. Theory and practice meld together beautifully: both worlds are able to interact and draw inspiration from each other. Gradient Descent and faster techinques lead to satisfactory results.
On the contrary, this is not happening in deep learning/non linear parametric optimization, where the optimization is non convex. Gradient descent suffers from many issues, including but not limited to:

- stationary points
- local minima
- plateaux
- bad initialization

In this more complicated setting, there are local guarantees [Jin+18; Lee+], but global efficient convergence is impossible to prove a priori. A line of work that tried to solve this last issue by describing good local minima has led to the establishment of important results given **very strong** assumptions:

- Most local minima are equivalent [Cho+15]
- no spurrious local minima [SJL22]
- other results up to different assumptions [JK17]

Despite the lack of a complete understanding on the theoretical side, neural networks prove to be instrumental for hard tasks where linear models do not perform well, and open the door to higher flexibility in terms of model design. These reasons justify the huge amount of production in the field. A theoretical work on one of the simplest models will be analyzed in the next Sections. In particular, we will see how two-layer sigmoid neural networks of the form $\phi(\theta) = \sigma\left(\sum_{i=1}^{d-2} \theta_i x_i + \theta_{d-1}\right)$ fall under the umbrella of a much broader class of optimization problems which has global optimization guarantees up to conditions to be specified. Such results are achieved thanks to techniques involving Wasserstein Gradient Flows.

1.3 Problem Formulation

We now focus on the space of functional optimization, where we simply decide that our functions of choice are parametrized by θ and live in a Hilbert space $\mathcal{F}$. Instead of minimizing in terms of parameters, we minimize in terms of functions arising from parameters. To do so, the loss function $R : \mathcal{F} \to R_+$ is chosen. A solution will be a combination of elements from the parametric space $\{\phi(\theta)\}_{\theta \in \Theta} \subset \mathcal{F}$. We treat this as a formal but **reasonable** assumption for practical purposes.

Assumption 1.5 (On the form of ϕ)**.** *Assume that ϕ parametrized by $\theta \in \Theta$ lives in the Hilbert space $\mathcal{F}$ and is differentiable.*

Measure perspective The minimization problem of Equation 1.1 could be treated as finding the optimal choice of θ in the $\mathbb{R}^d$ space to minimize the functional loss of a linear combination of functions $\phi(\theta)$. Endowing $\Theta = \mathbb{R}^{d-1}$ with a measure μ in the set $\mathcal{M}(\Theta)$ it is possible to restate the task as:

$$\mu^* = \arg\min_{\mu \in \mathcal{M}(\Theta)} J(\mu) \quad J(\mu) \coloneqq R\left(\int \phi d\mu \right) + G(\mu) \tag{1.6}$$

where:

- $G(\mu) : \mathcal{M}(\Theta) \to \mathbb{R}$ is the regularizer of the functional J, just like $\lambda \Xi(\theta)$ in Eqn. 1.1. Usually, the total variation norm (Def. A.51) is used when we do not want weights to concentrate[1] on specific $\theta \in \Theta$
- $|\Theta| = d - 1$
- $\mathcal{M}(\Theta)$ are all possible measures over the set Θ

In simple words, we look among all possible allocations of choices of the parameters to find the best combination that obtain a function[2] that attains minimal risk/maximum fit. The problem is linear in terms of μ. We give more justification of why we can inspect this form in Section 4.

Optimization practises Aiming to solve Eqn. 1.6 requires searching the ∞ dimensional space of measures on $\mathbb{R}^d$ for a **convex** problem. The latter is **good**, the former is **bad**, and makes the problem hard. The authors discuss references for two choices, which are **difficult to implement in practice**[CB18]:

- Frank-Wolfe Algorithm: greedy approach of adding neurons at every iteration. The decision problem of finding the optimal particle is in general NP-Hard [BP13; Jag13; Bac16]. It has connections with Conditional Gradient and Boosting [BSR15; Wan+15].
- Semidefinite hierarchy: based on expressing the measure in terms of its moments. Despite asymptotic global convergence and inheriting results from the larger class of *generalized moment problems* [Las09], it has its drawbacks. Only specific instances are covered [CDP17] and increasing the dimension growth is exponential.

What is **actually used in practice** is Gradient Descent, allowed by the differentiability of ϕ (Assumption 1.5). In the context of Equation 1.6, the measure μ is **discretized** to a finite set

[1] note that giving more "voting rights" does **not** generalize well

[2] as a linear combination of the ϕ

of *particles* against which backpropagation is performed.

$$\mu = \frac{1}{m}\sum_{i=1}^{m} \underbrace{w_i}_{\text{weight}} \underbrace{\delta_{\theta_i}}_{\text{position}}$$

Thanks to this discretization, a new object comes into play. While positions affect choices in the space of parameters, weights represent **degree of importance** in determining the function to feed into R and G.
The problem is then discretized as:

$$\mu^* = \underset{\mathbf{w}\in\mathbb{R}^m\ \boldsymbol{\theta}\in\boldsymbol{\theta}^m}{\arg\min}\ J_m(\mathbf{w},\boldsymbol{\theta}) \qquad J_m(\mathbf{w},\boldsymbol{\theta}) \coloneqq J\left(\frac{1}{m}\sum_{i=1}^{m} w_i\delta_{\theta_i}\right) \tag{1.7}$$

Observation 1.8 (Comments on Equation 1.7)**.** *Observe that there are m particles (hidden neurons) for which we have weights w_i and positions $\theta_i \in \mathbb{R}^{d-1}$.*
*By using discrete measures, we can weakly approximate any measure, where by **weakly** we mean when measuring an integral with respect to a measure of continuous and bounded functions.*

Despite the absence of problems in implementations, there are **no a priori guarantees** that J_m is convex, implying that convergence is, in most cases, at a local minima. The results shown are mostly centered on two questions:

- evaluating the algorithmic limit as $m \to \infty$, known to be equivalent to a **Wasserstein Gradient Flow** [NS17]
- assessing Global Convergence to the optimal measure μ^*, subject to a *generic ideal dynamics that one can only hope to approximate* [CB18]

Such results are obtained by building an approach that links the discretization with the original convex problem at the divergent limit of the number of particles.

In the context of analyzing Equation 1.7 some related works claimed that Stochastic Gradient Descent finds a global minimizer under very restrictive assumptions [LY17; SH17; VBB20; SJL22]. Interpreting the discretization as a *child* of Equation 1.6 was also present in [NS17] but not explored in search of global optimality conditions.

A non-quantitative condition is characterized for global convergence of Gradient Flows. It is stressed that this is only a starting point, as there is no indication of the criterion for such convergence to take place[3]. The connection between gradient flows and Gradient Descent (its discretized version, see Appendix B.1) is also extended to SGD [KY03](Thm. 2.1) and Accelerated gradient descent [Sci+17].

1.4 A more general problem

In this Subsection it is shown how, up to a certain set of assumptions, it is possible to *lift* the formulation of the easier problem into that of Equation 1.6. This is done thanks to a homogeneity property, which was also used in other optimization works [Jou+10; HV17].

[3] e.g. at what number of particles do we see a ϵ-bound on the error distance?

Lifted Optimization Version Consider the problem over **non negative finite measures** on $\Omega \subset \mathbb{R}^d$ of finding:

$$F^* = \min_{\mu \in \mathcal{M}_+(\Omega)} F(\mu) \qquad F(\mu) = R\bigg(\int \Phi d\mu \bigg) + \int V d\mu \tag{1.9}$$

Notice that we have changed both the normalization and the inner function.

Assumption 1.10. *Require the Hilbert space $\mathcal{F}$ to be separable (Def. A.14) and $\Omega \subset \mathbb{R}^d$ to be the closure of a convex open set. On top of this, establish that:*

1. *(smooth loss) $R : \mathcal{F} \to \mathbb{R}_+$ is differentiable and its differential dR is Lipschitz (Def. A.22) on bounded sets and bounded on sublevel sets (Def. A.10)*
2. *(basic regularity) the function $\Phi : \Omega \to \mathcal{F}$ is Fréchet differentiable, $V : \Omega \to \mathbb{R}_+$ is semi-convex (Def. A.17)*
3. *(sublinear growth and locally Lipschitz derivatives) there exists a sequence $(Q_r)_{r \geq 0}$ of nested non-empty closed convex subsets of Ω:*

$$(Q_r)_{r \geq 0} : Q_r \subset \Omega,\ Q_r \subset Q_{r'}, Q_r \neq \emptyset, convex \quad \forall r, r',\ r' > r$$

such that:

   (a) a kind of matryoshka property

$$\{u \in \Omega\ ;\ dist(u, Q_r) \leq r'\} \subset Q_{r+r'} \quad \forall r, r' > 0$$

   (b) Φ and V are bounded and $d\Phi$ is Lipschitz on each Q_r

   (c) denoting as $\|\partial V(u)\|$ the maximal norm of an element in $\partial V(u)$, the growth of the problem is sublinearly bounded as:

$$\exists C_1, C_2 > 0 \quad : \quad \sup_{u \in Q_r} \bigg\{ \|d\Phi_u\| + \|\partial V(u)\| \bigg\} \leq C_1 + C_2 r \quad \forall r > 0$$

Remark. *Given that Q_r can be unbounded (not necessarily balls of radius r), the result of point 3-(c) is not only that we have local Lipschitzness and sublinear growth, but also serves as a* ***technical requirement*** *for the gradient flow analysis to be stable.*
By convention, we set $F(\mu) = \infty$ if μ is not concentrated on Ω, i.e. if there are non-zero measure sets A such that $A \cap \Omega = \emptyset$. Basically, we force the nature of our measure inside the set we consider. This is done to avoid results in which part of the parameters are assigned outside the region of optimization.
The integral involving Φ is assumed to be a Bochner integral. In simple words, it maps to $\mathcal{F}$ whenever:

- *Φ is measurable*
- *$\int \|\phi\| \, d|\mu| < \infty$*

If these conditions are not verified, it also happens that $F(\mu) = \infty$. We need such a form of integral to allow the result of the operation to map to a Hilbert space.

The authors provide an explanation for pursuing an optimization of Equation 1.9 instead of Equation 1.6. The properties that allow to move from arbitrary measures over $\Theta \subset \mathbb{R}^{d-1}$ to probability[4] measures over $\Omega \subset \mathbb{R}^d$ are not trivial at all.

[4]finite, but we can normalize

Partially 1-homogeneous functions For continuous functions:

$$\phi : \Theta \to \mathcal{F} \qquad \widetilde{V} : \Theta \to \mathbb{R}_+$$

assign $\Omega \coloneqq \mathbb{R} \times \Theta \subset \mathbb{R}^d$, $\Phi(w, \theta) = w \cdot \phi(\theta)$ and $V(w, \theta) = |w|\widetilde{V}(\theta)$. Notice that Φ and V are 1-homogeneous (Def. A.19) in the first entry. We refer to this as **partial 1-homogeneity**.
As a first step, we make use of homogeneity to characterize the minimization over $\mathcal{M}_+(\Omega)$ equivalently in $\mathcal{P}(\Omega)$.

Proposition 1.11 (Normalization of measure)**.**

$$\exists \nu \in \mathcal{P}(\Omega) \quad : \quad F(\nu) = F(\mu) \quad \forall \mu \in \mathcal{M}_+(\Omega)$$

Proof. **(no mass case)** If $|\mu|(\Omega) = 0$ (i.e. recall Def. A.51, this means that there is no negative or positive mass), then $F(\mu) = 0$. For this reason, choose $\nu = \delta_{(0,\theta_0)}$ for arbitrary $\theta_0 \in \Theta$. Clearly $\nu \in \mathcal{P}(\Omega)$ and also $F(\delta_{(0,\theta_0)}) = 0$.
(non zero mass case) Let, $|\mu|(\Omega) > 0$. Define the map and the pushforward:

$$T : (w, \theta) \to (|\mu|(\Omega) \cdot w, \theta) \quad \nu \coloneqq T_{\#}\left(\frac{\mu}{|\mu|(\Omega)}\right) = \frac{1}{|\mu|(\Omega)}\mu \circ T^{-1} \in \mathcal{P}(\Omega)$$

According to Definition A.52. By $|\mu|(\Omega) > 0$ this is allowed and we also get that the new measure is normalized $\nu \in \mathcal{P}(\Omega)$. To conclude, we briefly check that it satisfies the requirement by Proposition A.54. In fact, the hypothesis holds since $\Phi \circ T = (w \cdot \phi) \circ T$ is μ-integrable and $V \circ T = (|\omega|\widetilde{V}) \circ T$ is μ-integrable. The constant $|\mu|(\Omega)$ is taken out of the measure, and we get back the usual μ in the differential. The added terms w and $|w|$ are both integrable and do not impact the operation. □

Making use of Proposition 1.11 we introduce a projection operator specifically crafted for the 1-homogeneous case:

$$h^1 : \mathcal{M}_+(\Omega) \to \mathcal{M}(\Theta) \quad h^1(\mu)(B) = \int_{\mathbb{R}} w\mu(dw, B) \;\; \forall \mu \in \mathcal{P}(\Omega),\; B \subset \Theta \; \textit{measurable} \tag{1.12}$$

Intuitively, we integrate out the $w \in \mathbb{R}$ added part and obtain a measure over Θ.
We can equivalently characterize it as:

$$\int_\Theta \varphi(\theta) dh^1(\mu)(\theta) = \int_{\mathbb{R}\times\Theta} w\varphi(\theta) d\mu(w, \theta) \quad \forall \varphi : \Theta \to \mathbb{R} \; \textit{continuous bounded} \tag{1.13}$$

Which is well defined whenever $(w, \theta) \to w$ is μ-integrable. Usually the latter is easier to check than the former.

Proposition 1.14 (Equivalence under lifting)**.** *we derive three important conclusions:*

1. *the projection h^1 is such that:*

$$\mathcal{M}(\Theta) \subset h^1(\mathcal{P}(\Omega)) = h^1(\mathcal{M}_+(\Omega))$$

2. *Regularizers*[5] *on $\mathcal{M}(\Theta)$ of the form:*

$$G(\nu) = \inf_{\nu \in h^{-1}(\mu)} \int_\Omega V d\mu$$

[5]note that the dependence is on the Ω measure, but we are working on the functional for Θ. This is done to bridge the two formulations

are such that:

$$\inf_{\nu\in\mathcal{M}(\Theta)} J(\nu) = \inf_{\mu\in\mathcal{M}_+(\Omega)} F(\mu)$$

3. *If $G(\cdot)$ attains the infimum and $\nu \in \mathcal{M}(\Theta)$ minimizes J:*

$$\exists \mu \in h^{-1}(\nu) \quad : \quad \mu = \arg\min_{\mathcal{M}_+(\Omega)} F$$

Notice that by Claim #1 we write in Claim #3 that $\mu \in h^{-1}(\nu)$ since h^1 maps to a space that includes $\mathcal{M}(\Theta)$.
To avoid confusion, we recapitulate the operations as follows:

$$\mathcal{P}(\Omega) \ni \mu \overset{h^1(\cdot)}{\rightsquigarrow} \nu \in \mathcal{M}(\Theta)$$

$$\int \phi d\nu \overset{\cdot w}{\rightsquigarrow} \int \Phi d\mu$$

$$G(\nu) = \int \widetilde{V}(\theta) d\nu \overset{\cdot|w|}{\rightsquigarrow} G(\mu) = \int V(w,\theta) d\mu$$

Proof. **(Claim #1)** The equivalence $h^1(\mathcal{P}(\Omega)) = h^1(\mathcal{M}_+(\Omega))$ follows by Proposition 1.11. Taking into account $\nu \in \mathcal{M}(\Theta)$, recognize that any finite measure is decomposed by Proposition A.61:

$$\nu = f\sigma \quad \sigma \in \mathcal{P}(\Theta),\ f : \Theta \to \mathbb{R} \in L^1(\sigma)$$

Namely ν is taken to be a σ-integrable function times a probability measure. For σ, we could take the normalized variation of μ whenever it is positive. Then, the measure obtained by the extension is the pushforward of σ:

$$\mu \coloneqq (f \times \mathtt{id})_{\#}\sigma = \sigma \circ (f \times \mathtt{id})^{-1} \in \mathcal{P}(\Omega) \tag{1.15}$$

Using the projection map characterization (Eqn. 1.13) we can say that for $\varphi : \Theta \to \mathbb{R}$ continuous and bounded:

$$\begin{aligned}
\int_\Theta \varphi(\theta) dh^1(\mu)(\theta) &= \int_\Omega w\varphi(\theta) d\mu(w,\theta) && \text{Eqn. 1.13} \\
&= \int_\Omega \underbrace{w\varphi(\theta)}_{g:\Omega\to\mathbb{R}} d(\sigma \circ (\underbrace{f \times \mathtt{id}}_{T})^{-1})(w,\theta) && \mu = T_{\#}\sigma,\ T(f) = f \times \mathtt{id} \\
&= \int_\Theta \underbrace{[w\varphi(\theta)]}_{g} \circ (\underbrace{f \times \mathtt{id}}_{T}) d\sigma(\theta) && \text{Prop. A.54} \\
&= \int_\Theta \varphi(\theta) f(\theta) d\sigma(\theta) \\
&= \int_\Theta \varphi(\theta) d\nu(\theta) && \nu = f\sigma
\end{aligned}$$

By the arbitrariness of φ, we can state $h^1(\mu) = \nu$, thus proving that for each $\nu \in \mathcal{M}(\Theta)$ in the reduced space there $\exists \mu \in \mathcal{P}(\Omega)$ from the higher dimensional space. This is the very definition of surjectivity, namely $h^1(\mathcal{P}(\Omega)) \supset \mathcal{M}(\Theta)$.
(Claim #2) By the definition of h^1 in Equation 1.13 we have that:

$$\begin{aligned}
\int \Phi d\mu &= \int (w \cdot \phi) d\mu \\
&= \int \phi d\nu
\end{aligned}$$

By the very assignment of G as inf of the regularization in Ω we have the following:

$$F(\mu) \geq J(\nu)$$

But equal at infimums.
(Claim 3) assuming that $G(\nu) = \min_{\nu \in h^{-1}(\mu)} \int_\Omega V d\nu$ and that $\nu = \arg\min_{\mathcal{M}(\Theta)} J$ we have that μ certainly minimizes J as it cannot do better than the min and the Bochner integrals are equal. □

Proposition 1.16 (Total variation is included in regularizers)**.** *Let $V(w, \theta) = |w|$:*

Then:

$$\mu \in \mathcal{M}_+(\Omega) \implies \int V d\mu \geq |h^1(\mu)|(\Theta)$$

With equality if μ is as in 1.15. If this is the case, we satisfy Claims #2, #3 of the previous Proposition.

Proof. Let $\mu \in \mathcal{P}(\Omega)$ and $\nu = h^1(\mu)$. Define:

$$\widetilde{\nu}_+ \coloneqq \int_{\mathbb{R}_+} w\mu(dw, \cdot) \quad \widetilde{\nu}_- \coloneqq -\int_{\mathbb{R}_-} w\mu(dw, \cdot)$$

We have by construction of h^1 (Eqn. 1.12) that $\nu = \widetilde{\nu}_+ - \widetilde{\nu}_-$ and recalling the concept of total variation (Def. A.51):

$$\begin{aligned}
|\nu|(\Theta) &= |\nu_+|(\Theta) + |\nu_-|(\Theta) && \text{Jordan decomposition (Thm. A.49)} \\
&\leq |\widetilde{\nu}_+|(\Theta) + |\widetilde{\nu}_-|(\Theta) && \text{Cor. A.50} \\
&= \int V d\mu && \text{by } V(w, \theta) = |w|
\end{aligned}$$

Using [Coh13] (Cor. 4.1.6) we can further say that equality holds whenever:

$$|\nu|\Big(\{\texttt{spt}\ \widetilde{\nu}_+ \cap \texttt{spt}\ \widetilde{\nu}_-\}\Big) = 0$$

Which is verified in the case of Equation 1.15 [CB18]. □

This last result allows us to reconcile weights of the neurons w_i and positions of the neurons θ inside the parameter space Ω. From now onwards, we mostly work on Ω, where the our two objects of interest can be treated jointly.

2 Gradient Flows

We move on to describe the evolutionary process of the parameters in terms of a gradient flow equation. Such a formalization is commented on and proved to be well defined. What follows is a presentation of Wasserstein Gradient flows, which describe the evolution of a probability measure over the parameters. The dynamics, viewed from this Optimal Transport perspective, have nicer properties that are outlined. Furthermore, the construction is completely characterized in the context of the main Assumptions. To justify the need to showcase Wasserstein Gradient Flows, a statement about the limit of the particle measure having this form is provided at the end. It is concluded that it is equivalent to work over probability measures, with the possibility of using all the results they come with.

2.1 Particle Gradient Flow

Other authors have studied the so-called many-particle limit of Neural Networks with two layers and quadratic R, focusing on the performance of SGD. For more results in this direction, some references are reported [MMN18; RV19; SS19]. The main difference is that they are based on statistical properties of the problems, while this study is focused on the homogeneous structure chosen.

While in the general framework we would minimize Equation 1.6, by the arguments of Section 1.3 we end up considering a **discretized** version:

$$F_m(\mathbf{u}) \coloneqq F\left(\frac{1}{m}\sum_{i=1}^{m}\delta_{\mathbf{u}_i}\right) = R\left(\frac{1}{m}\sum_{i=1}^{m}\Phi(\mathbf{u}_i)\right) + \frac{1}{m}\sum_{i=1}^{m}V(\mathbf{u}_i) \tag{2.1}$$

For $m \in \mathbb{N}$ and $\mathbf{u}_i = (w_i, \theta_i)\forall i, \mathbf{u} \in \Omega^m$ encoding weights and positions.
Our aim is to build a gradient flow to analyze the extreme dynamics of minimization of such object. This is derived as the limit of a differential step of gradient descent using the Euler technique (see Appendix, Subsection B.1). In our context, given that in Assumption 1.10 we do not have that V is differentiable, we may only conclude that the flow will be a subgradient flow (Def. A.25). Figure 1 is an example of the dynamics over the landscape of a function.

Definition 2.2 (Particle Gradient flow)**.** *A dynamics for F_m:*

$$\boldsymbol{u} : \mathbb{R}_+ \to \Omega^m \quad t \to \boldsymbol{u}(t) \in \Omega^m$$

is a particle gradient flow when the following conditions hold:

1. *absolute continuity (Def. A.70)*
2. *rescaled gradient flow equation $\boldsymbol{u}'(t) = -m\partial F_m(\boldsymbol{u}(t))$ a.e. $t \geq 0$*

Notice that in #2 we have:

- *almost everywhere conditions by the absolute continuity requirement #1*
- *subdifferentials as argued by the potential non-differentiability of V which is only semiconvex*
- *rescaling by m for convenience at divergent size $m \to \infty$. Formally it is the particle gradient flow of a $(\mathbb{R}^d)^m = (\mathbb{R} \times \Theta)^m$ scalar product rescaled with each atom having $\frac{1}{m}$ mass. This does not hurt the dynamics.*

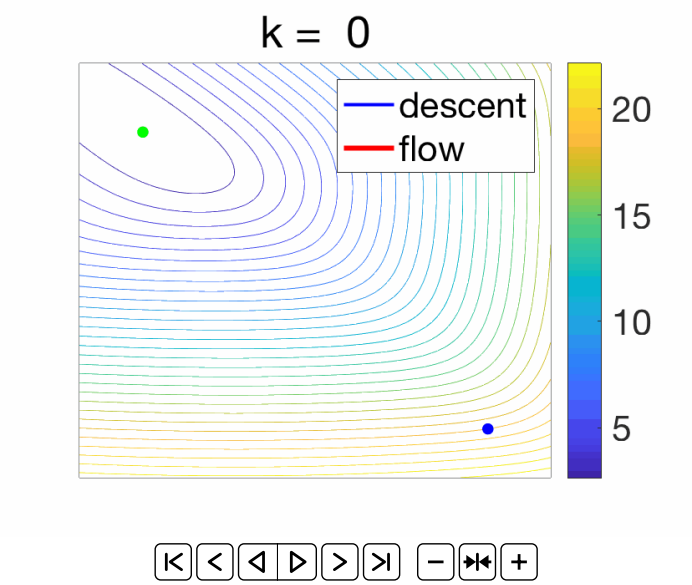


Figure 1: Animated GD vs GF. Source: [Bac20b]
Gradient descent vs. gradient flow on the same time scale for a logistic regression problem.

We first provide a characterization.

Proposition 2.3 (Existance and uniqueness of gradient flow)**.** *There exists a unique gradient flow for any initialization of dynamics in F_m.*

$$\forall \boldsymbol{u}(0) \in \Omega^m \quad \exists!\ \boldsymbol{u} : \mathbb{R}_+ \to \Omega^m$$

Additionally, for almost every $t > 0$ it holds:

$$\frac{d}{ds} F_m(\boldsymbol{u}(s))\Big|_{s=t} = -\frac{1}{m}|\boldsymbol{u}'(t)|^2$$

Namely, the derivative of the evolution of the risk in terms of time is equal to the opposite of the square norm of the gradient flow dynamics.
While considering the velocity of a single particle given by $\boldsymbol{u}_i'(t) = v_t(\boldsymbol{u}_i(t))$, simplifying for $u \in \Omega$ a particle-position pair, and $\mu_{m,t} \coloneqq \frac{1}{m}\sum_{i=1}^m \delta_{\boldsymbol{u}_i(t)}$ we have:

$$v_t(u) = \widetilde{v}_t(u) - \texttt{proj}_{\partial V(u)}(\widetilde{v}_t(u)) \qquad \widetilde{v}_t(u) = -\left[\left\langle R'\left(\int \Phi d\mu_{m,t}\right),\ \partial_j \Phi(u)\right\rangle\right]_{j=1}^d \tag{2.4}$$

Where $R'(f)$ denotes the gradient of R at $f \in \mathcal{F}$ and $\partial_j \Phi \in \mathcal{F}$ is the differential $d\Phi(u)$ applied to the j^{th} vector of the canonical basis of $\mathbb{R}^d$, namely the j^{th} entry of the differential of Φ for $\boldsymbol{u} \in \mathbb{R}^d$.

Proof. **(existance and uniqueness on finite intervals)** By Proposition A.21 F_m is locally semiconvex. Existance and uniqueness of a gradient flow on $[0, T)$ is granted in [San17](Sec. 2.1).
(closed form) the expression for the velocity of single particles is obviously updated over time and requires a projection since it selects subgradients of pointwise minimal norm [San15].

Precisely, for almost every $t \in \mathbb{R}_+, u \in \Omega$ the derivative of $\mathbf{u}$ at each particle (i.e. $v_t(u)$) is minus the subgradient of minimal norm.

$$
\begin{aligned}
v_t(u) &= \arg\min\left\{ |v|^2 \ ; \ \widetilde{v}_t(u) - v \in \partial V(u) \right\} && \text{for } \widetilde{v}_t(u) = -\left[\left\langle R'\Big(\int \Phi d\mu_{m,t} \Big), \ \partial_j \Phi(u) \right\rangle \right]_{j=1}^{d} \\
&= \arg\min\left\{ |v|^2 \ ; \ z \in \partial V(u) \right\} && z = \widetilde{v}_t(u) - v \\
&= \widetilde{v}_t(u) - \arg\min\left\{ |\widetilde{v}_t(u) - z|^2 \ ; \ z \in \partial V(u) \right\} && (2.5) \\
&= \widetilde{v}_t(u) - \texttt{proj}_{\partial V(u)} \widetilde{v}_t(u) && (2.6)
\end{aligned}
$$

where:

- in Equation 2.5 we move the minimization taking out $\widetilde{v}_t(u)$, keeping the same constraint, but updating the norm to be a minimization of the correct remaining part.
- in Equation 2.6 we use the definition of the `proj` operator

(globality) If the function F_m is lower bounded, there are procedures to evaluate $\mathbf{u}(T)$ at $T = \infty$, even if F_m is not globally semiconvex, which is something we do not require. $\forall t > 0$ it holds that:

$$
\begin{aligned}
F_m(\mathbf{u}(0)) - F_m(\mathbf{u}(t)) &= -\int_0^t \frac{d}{ds} F_m(\mathbf{u}(s)) ds \\
&= \frac{1}{m} \int_0^t |\mathbf{u}'(s)|^2 ds \\
&\geq \frac{t}{m} \Big(\int_0^t |\mathbf{u}'(s)| ds \Big)^2 && \text{Jensen's, see below}
\end{aligned}
$$

So F_m is lower bounded, and we get that also the flow's legnth is bounded $\forall [0,t]$ by the difference of F_m at start and at t. By compactness, if $T < \infty$ then $\exists \mathbf{u}(T)$, and we contradict the maximality[6] of the difference above. Hence, the best result is necessarily obtained at $T = \infty$ and the flow is globally defined.
Concerning the application of Jensen's inequality we have:

$$
\begin{aligned}
\int_0^t |\mathbf{u}'(s)|^2 ds &= t^2 \int_0^t \left(\frac{1}{t} |\mathbf{u}'(s)| \right)^2 ds \\
&\geq t^2 \frac{1}{t} \Big(\int_0^t \mathbf{u}'(s) ds \Big)^2 \\
&= t \Big(\int_0^t \mathbf{u}'(s) ds \Big)^2
\end{aligned}
$$

Which follows the unnormalized case of Jensen's application by Hölder's inequality (Prop. A.74). A reference is [CG21](Prop. 2.71, Rem. 2.72). □

Observation 2.7 (Facts about the proposition)**.** *recognize that:*

- $[\widetilde{v}_t(\boldsymbol{u}_i)]_{i=1}^m = -\nabla R \left(\frac{1}{m} \sum_{i=1}^m \Phi(\boldsymbol{u}_i) \right)$*, which is the first term of Equation 2.1*

[6] enforced by the differential equation being always decreasing, and thus always increasing with a $-$ in front

- *recalling the differential vs subdifferential discussion for V:*
  - *V differentiable $\implies$ $v_t(u) = \widetilde{v}_t(u) - \nabla V(u)$, the classical gradient of Equation 2.1*
  - *V non differentiable $\implies$ $v_t(u)$ is the continuous-time version of the forward-backward minimization algorithm ([GBC16] for more information)*

2.2 Wasserstein Gradient Flow (Wgf)

By the result of Proposition 2.3 we are interested in understanding if the same happens with dynamics over probability measures, and if the discrete particle case has ties with its limit.

It is quite easy to recover the differential of F evaluated at a measure $\mu \in \mathcal{M}(\Omega)$:

$$F'(\mu) : \Omega \to \mathbb{R} \quad F'(\mu)(u) \coloneqq \left\langle R'\left(\int \Phi d\mu \right), \Phi(u) \right\rangle + V(u) \tag{2.8}$$

we notice that again by Proposition 2.3 we stated that $v_t(u)$ is a **field** in $-\partial F'(\mu_{m,t})$, since v_t is the derivative of the evolution of the gradient (somehow we could say the subgradient of the derivative of F).

Definition 2.9 (Wasserstein Gradient Flow)**.** *For the functional F and an interval $[0,T)$ a Wasserstein gradient flow is a path $t \to \mu_t$ on $[0,T)$ such that:*

1. *it is absolutely continuous (Def. A.70)*
2. *$(\mu_t)_{t\in[0,T)} \in \mathcal{P}_2(\Omega)$*
3. *for $[0,T) \times \Omega^d$ satisfies (distributionally) the continuity equation:*

$$\partial_t \mu_t = -\texttt{div}(v_t \mu_t) \quad v_t \in \partial F'(\mu_t) \tag{2.10}$$

In the next claim it is shown that Definition 2.9 is a proper generalization of Definition 2.2. Observe that the way we define the former is justified in a distributional sense by the fact that densities are not necessarily smooth. A broader presentation is given in the Appendix A.5 and B.2.

Proposition 2.11 (Link gradient flow and atomic Wasserstein gradient flow)**.** *For a gradient flow $\boldsymbol{u} : \mathbb{R}_+ \to \Omega^m$ of F_m the map:*

$$t \to \mu_{m,t} \coloneqq \frac{1}{m} \sum_{i=1}^{m} \delta_{\boldsymbol{u}_i(t)}$$

is a Wasserstein gradient flow for the non particle version of F_m, denoted as F.

Proof. Assign to $v_t(u)$ the resulting velocity of the vector field $\mathbf{u}$ from Proposition 2.3.
($\triangle$ absolute continuity) relative to W_2 (Def. A.63) the path $t \to \mu_{m,t}$ is absolutely continuous (Def. A.70). This holds intuitively by the fact that we can bound the evolution of parameters in their euclidean norm by the time expired, using the closed form velocity derived in Proposition 2.3.
($\square$ distributional continuity equation) let $\varphi : (0,\infty) \times \mathbb{R}^d \to \mathbb{R}$ be smooth with $\texttt{spt}\varphi$

compact. Then:

$$\begin{aligned}
0 &= \frac{1}{m}\sum_{i=1}^{m}\int_{\mathbb{R}_+}\frac{d}{dt}\varphi_t(\mathbf{u}_i(t))dt \\
&= \frac{1}{m}\sum_{i=1}^{m}\int_{\mathbb{R}_+}\Big(\partial_t\varphi_t(\mathbf{u}_i(t)) + \nabla_u\varphi_t(\mathbf{u}_i(t))\cdot v_t(\mathbf{u}_i)\Big)dt \\
&= \int_{\mathbb{R}_+}\int_{\Omega}\Big(\partial_t\varphi_t(u) + \nabla_u\varphi_t(u)\cdot v_t(u)\Big)d\mu_{m,t}dt
\end{aligned}$$

Where we used smoothness to conclude that it vanishes at the extreme boundaries of $\mathbb{R}_+$ (thus the first zero), the chain rule to derive the second equality, and the definition of discrete measure for the third.
Being equivalent to the definition in distributional sense of the continuity equation (Eqn. 2.10) means that $(\mu_{m,t})_{t\geq 0}$ is a distributional solution to it. □

Observation 2.12 (Comments on the result)**.** *notice that the dynamics are in t at m fixed. Thus, if F does not admit an atomic minimizer with m atoms, $\mu_{m,t}$ converges to a measure that* ***does not*** *minimize F.*

By the result of Proposition 2.11 when referring to the whole $\partial F'$ we call it **Wasserstein subdifferential**. The link of the two different perspectives allows us to think of the dynamics over parameters as a discrete measure. However, we do not have information about the uniqueness and/or existance of a closed form of a Wasserstein Gradient Flow for arbitrary starting measures μ_0. We answer this question in the subsection below.

2.2.1 Properties of the Wasserstein gradient flow

This Subsubsection mostly refers to [AGS05].

Setting The authors start from a collection of **intermediate Wgfs** to replace the lifted problem of Equation 1.9. These are specified by:

$$F^{(r)} : \mathcal{P}_2(\Omega)\to\mathbb{R},\ r>0 \quad F^{(r)}(\mu) = \begin{cases} F(\mu) & \mu(Q_r)=1 \\ \infty & otherwise \end{cases}$$

Recalling that $(Q_r)_{r>0}$ is presented in Assumption 1.10 we further add that the function $\gamma : \mathcal{P}(\Omega\times\Omega)$ is an **admissible transport plan** for $r>0$ if:

- $\pi^1_{\#}\gamma, \pi^2_{\#}\gamma$ are concentrated (Def. A.48) on Q_r
- $\pi^1_{\#}\gamma, \pi^2_{\#}\gamma$ have **finite** second moments[7]

To an admissible transport plan, we assign the **transport cost**:

$$C_p(\gamma) \coloneqq \left(\int |y-x|^p d\gamma(x,y)\right)^{\frac{1}{p}} \quad p\geq 1 \tag{2.13}$$

[7] i.e. they have an expectation and a variance, or equivalently they are in $\mathcal{P}_2(\Omega)$.

to simplify calculations given the **bounded** set[8]

$$\mathcal{F}_r \coloneqq \left\{ \int \Phi d\mu \ ; \ \mu \in \mathcal{P}_2(\Omega), \mu(Q_r) = 1 \right\}$$

we also pre-assign the following symbols:

$$\|d\Phi\|_{\infty,r} = \sup_{u \in Q_r} \|d\Phi_u\| \qquad L_{d\Phi} = \sup \left\{ \frac{\|d\Phi_{\widetilde{u}} - d\Phi_u\|}{|\widetilde{u} - u|} \ ; \ u, \widetilde{u} \in Q_r, \ u \neq \widetilde{u} \right\} \tag{2.14}$$

$$\|dR\|_{\infty,r} = \sup_{f \in \mathcal{F}_r} \|dR_f\| \qquad L_{dR} = \sup \left\{ \frac{\|dR_f - dR_g\|}{\|f - g\|} \ ; \ f, g \in \mathcal{F}_r, \ f \neq g \right\} \tag{2.15}$$

Proposition 2.16 (Finiteness of symbols)**.** *The quantities in Equations 2.14 and 2.15 are finite under Assumptions 1.10.*

Proof. This is a direct implication of the assumptions. □

Observation 2.17 (V is famous in literature)**.** *The quantity V from Equations 1.9 under an integral and alone in 2.8 is well studied in the literature. For a reference of its description, see [AGS05],(Prop. 10.4.2). We set it to $V = 0$ in the next Lemma to focus on the first term. To understand how V and R eventually meld, see Lemma 2.25.*

Lemma 2.18 (Properties of $F^{(r)}$)**.** *Let:*

- *Assumptions 1.10 hold*
- $V = 0$

Then, $\forall r > 0$:

1. *$F^{(r)}$ is proper[9] and continuous on its closed domain $\{\mu \in \mathcal{P}_2(\Omega) \ ; \ \mu(Q_r) = 1\}$*
2. *$\exists \lambda_r > 0$ such that $\forall \gamma$ admissible transport plans the function:*

$$t \to F(\mu_t^\gamma) \qquad \mu_t^\gamma \coloneqq ((1-t)\pi^1 + t\pi^2)_\# \gamma$$

   is differentiable with a derivative that is $\lambda_r C_2^2(\gamma)$-Lipschitz (Def. A.22), where C_2 is the transport cost.
3. *for μ concentrated on Q_r a field $v \in L^2(\mu, \mathbb{R}^d)$ is such that for any admissible transport plan such that $\pi^1_\# \gamma = \mu$ the functional for the second marginal is bounded below by:*

$$F(\pi^2_\# \gamma) \geq F(\mu) + \int v(u) \cdot (\widetilde{u} - u) d\gamma(u, \widetilde{u}) + o(C_2(\gamma)) \iff v(u) \in \partial(F'(\mu) + \iota_{Q_r})(u) \quad a.e.\ u \in \Omega$$

$$\iota_{Q_r}(u) = \begin{cases} 0 & u \in Q_r \\ \infty & u \notin Q_r \end{cases}$$

   Where ι is a convex function.

[8] bounded by the construction of the Q_r. The functions arising as integrals effectively assign parameters from a Q_r subset of the domain. We basically state $\exists M : \|\int \Phi d\mu\| \leq M \forall \int \Phi d\mu \in \mathcal{F}_r$. This holds by classical bounds on the integral and Assumption 1.10 and the construction of $\mathcal{F}_r$.

[9] i.e. finite in at least one point

Proof. **(Claim #1)(♢ properness)** the closed domain $\{\mu \in \mathcal{P}_2(\Omega) \ ; \ \mu(Q_r) = 1\}$ makes F proper since we are restricting it to the ball where all the mass is concentrated $\forall r > 0$. Inside such a ball, $\forall u_0 \in Q_r$ it holds that $F^{(r)}(u_0) = R(\Phi(u_0)) < \infty$.
(□ continuity) Additionally, by the result of Lemma A.69 we have that $F^{(r)}$ is continuous in the closed domain stated.
(Claim #2)(△ differentiability) denoted $h(t) \coloneqq F^{(r)}(\mu_t^\gamma)$, we wish to inspect the dynamics. By Assumption 1.10 #1,#3 dR and $d\Phi$ are Lipschitz, we have that:

$$\begin{aligned} h'(t) = \frac{d}{dt}F^{(r)}(\mu_t^\gamma) &= \frac{d}{dt}\left(R\left(\int \Phi d\mu_t^\gamma\right)\right) \\ &= \frac{d}{dt}\left(R\left(\int \Phi(u) d((1-t)\pi^1 + t\pi^2)_\# \gamma(x,y)\right)\right) \\ &= \frac{d}{dt}\left(R\left(\int \Phi((1-t)x + ty) d\gamma(x,y)\right)\right) \end{aligned}$$

where:

- $d\Phi$ is uniformly bounded (being Lipschitz[10]) in Q_r
- the Bochner integral admits a dominated convergence theorem (i.e. we can integrate bounded functions and bring the differentiation inside to get $d\Phi$) [Coh13](Thm. E6)

which together mean that we can differentiate $h(t)$ to obtain by the chain rule:

$$\begin{aligned} h'(t) &= \left\langle R'\left(\int \Phi d\mu_t^\gamma\right), \frac{d}{dt}\int \Phi((1-t)x+ty)d\gamma(x,y)\right\rangle \\ &= \left\langle R'\left(\int \Phi d\mu_t^\gamma\right), \int \frac{d}{dt}\left(\Phi((1-t)x+ty)\right) d\gamma(x,y)\right\rangle && \text{bring } \frac{d}{dt} \text{ inside} \\ &= \left\langle R'\left(\int \Phi d\mu_t^\gamma\right), \int d\Phi_{(1-t)x+ty}(y-x) d\gamma(x,y)\right\rangle && \text{again chain rule} \end{aligned}$$

(◯ Lipschitz derivative) letting $0 \le t_1 < t_2 < 1$ we have that:

$$\begin{aligned} |h'(t_2) - h'(t_1)| &= \Big| \langle R'(\textstyle\int \Phi d\mu_{t_2}^\gamma), \int d\Phi_{(1-t_2)x+t_2y}(y-x)d\gamma(x,y)\rangle \\ &\quad - \langle R'(\textstyle\int \Phi d\mu_{t_1}^\gamma), \int d\Phi_{(1-t_1)x+t_1y}(y-x)d\gamma(x,y)\rangle \Big| \\ &= \Big| \langle a, b\rangle - \langle c, d\rangle \Big| \\ &= \Big| \langle a, b\rangle - \langle c, b\rangle + \langle c, b\rangle - \langle c, d\rangle \Big| \\ &= \Big| \langle a-c, b\rangle + \langle c, b-d\rangle \Big| \\ &\le \Big| \langle a-c, b\rangle \Big| + \Big| \langle c, b-d\rangle \Big| \\ &=: |(I)| + |(II)| \end{aligned} \tag{2.19}$$

Which is just a manipulation of the inner product using linearity and the triangle inequality. Inspecting the two quantities with the predefined symbols of Equation 2.14, 2.15 in hand, we

[10] trivially Def.A.22 for Lipschitz functions is a precise formulation of uniform boundedness, namely Def. A.34

find that we can decompose the two terms by taking the norms out, obtaining the inequalities:

$$\begin{aligned}
(I) &= \left| \left\langle R'(\int \Phi d\mu_{t_2}) - R'(\int \Phi d\mu_{t_1}), \int d\Phi_{(1-t)x+ty}(y-x)d\gamma(x,y) \right\rangle \right| \\
&\leq \left\| R'(\int \Phi d\mu_{t_2}) - R'(\int \Phi d\mu_{t_1}) \right\| \times \left\| \int d\Phi_{(1-t)x+ty}(y-x)d\gamma(x,y) \right\| \qquad \text{Cauchy-Schwarz} \\
&\leq \left| L_{dR} \left\| d\Phi \right\|_{\infty,r} \underbrace{\int (y-x)d\gamma(x,y)}_{=C_1(\gamma)} |t_2 - t_1| \right| \\
&\times \left\| d\Phi \right\|_{\infty,r} \int (y-x)d\gamma(x,y) \\
&= L_{dR} \left\| d\Phi \right\|_{\infty,r}^2 C_1^2(\gamma)|t_2 - t_1| \\
&\leq L_{dR} \left\| d\Phi \right\|_{\infty,r}^2 C_2^2(\gamma)|t_2 - t_1| \qquad \underline{\text{inequality}}
\end{aligned}$$

Where the inequality is an application of Hölder's inequality for costs ($C_1^2(\gamma) \leq C_2^2(\gamma)$, see Subsec. B.3), and $\times$ is used to highlight the product but is just scalar multiplication. Similarly:

$$\begin{aligned}
(II) &= \left| \left\langle R'(\int \Phi d\mu_{t_1}), \int [d\Phi_{(1-t_2)x+t_2y} - d\Phi_{(1-t_1)x+t_1y}](y-x)d\gamma(x,y) \right\rangle \right| \\
&\leq \left\| R'(\int \Phi d\mu_{t_1}) \right\| \times \left\| [\int d\Phi_{(1-t_2)x+t_2y} - d\Phi_{(1-t_1)x+t_1y}](y-x)d\gamma(x,y) \right\| \\
&\leq \left\| dR \right\|_{\infty,r} C_1(\gamma) \times L_{d\Phi} C_1(\gamma)|t_2 - t_1| \\
&\leq L_{d\Phi} \left\| dR \right\|_{\infty,r} C_2^2(\gamma)|t_2 - t_1|
\end{aligned}$$

And recollecting the results in Equation 2.19 we eventually get that for $0 \leq t_1 < t_2 \leq 1$:

$$\begin{aligned}
|h'(t_2) - h'(t_1)| &\leq (I) + (II) \\
&\leq L_{dR} \left\| d\Phi \right\|_{\infty,r}^2 C_1^2(\gamma)|t_2 - t_1| + L_{d\Phi} \left\| dR \right\|_{\infty,r} C_2^2(\gamma)|t_2 - t_1| \\
&= [L_{dR} \left\| d\Phi \right\|_{\infty,r}^2 + L_{d\Phi} \left\| dR \right\|_{\infty,r}] C_2^2(\gamma)|t_2 - t_1| \\
&= \lambda_r C_2^2(\gamma)|t_2 - t_1| \qquad \lambda_r = L_{dR} \left\| d\Phi \right\|_{\infty,r}^2 + L_{d\Phi} \left\| dR \right\|_{\infty,r}
\end{aligned}$$

which is in accordance with the definition of $\lambda_r C_2^2(\gamma)$-Lipschitzness.

(Claim #3)(♠ locality means Taylor) a velocity field involves that we have to inspect infinitesimal variations over time of our objects. Letting $u, \widetilde{u} \in Q_r$ and $f, g \in \mathcal{F}_r$ for $r > 0$:

$$\Phi(\widetilde{u}) = \Phi(u) + d\Phi_u(\widetilde{u} - u) + M(u, \widetilde{u}) \qquad \text{parameter expansion} \qquad (2.20)$$

$$R(g) = R(f) + \left\langle R'(f), f - g \right\rangle + N(f, g) \qquad \text{loss expansion} \qquad (2.21)$$

Obviously, the remainders are bounded since $L_{d\Phi}$ and L_{dR} bound second derivatives (we prove this in general in Prop. A.23):

$$\begin{aligned}
M(u, \widetilde{u}) &\leq \frac{1}{2} \Phi''_u |\widetilde{u} - u|^2 \\
\implies \left\| M(u, \widetilde{u}) \right\| &\leq \frac{1}{2} \left\| \Phi''_u \right\| |\widetilde{u} - u|^2 \\
&\leq \frac{1}{2} L_{d\Phi} |\widetilde{u} - u|^2 \\
N(f, g) &\leq \frac{1}{2} (f - g)^T R''(f)(f - g) \\
\implies \left\| N(f, g) \right\| &\leq \frac{1}{2} \left\| R''(f) \right\| \left\| f - g \right\|^2 \\
&\leq \frac{1}{2} L_{dR} \left\| f - g \right\|^2
\end{aligned}$$

with these in hand we can eventually expand around F.
(♡ loss and parameter expansion) Denoting $\mu = \pi^1_\# \gamma, \nu = \pi^2_\# \gamma$, up to being both concentrated on Q_r we further specify that in the next expression we will use:

- $f = \int \Phi d\mu$ so that $R(f) = R(\int \phi d\mu)$
- $g = \int \Phi d\nu$ so that $R(g) = R(\int \Phi d\nu)$

With this notation, we first perform a loss expansion and then a parameter expansion, slightly tweaked after noticing that:

$$\Phi(\widetilde{u}) = \Phi(u) + d\Phi_u(\widetilde{u} - u) + M(u, \widetilde{u}) \iff \Phi(\widetilde{u}) - \Phi(u) = d\Phi_u(\widetilde{u} - u) + M(u, \widetilde{u})$$

Which we could express under an integral equivalently as:

$$\int \Phi(\widetilde{u}) - \Phi(u) d\gamma(u, \widetilde{u}) = \int d\Phi_u(\widetilde{u} - u) + M(u, \widetilde{u}) d\gamma(u, \widetilde{u})$$

Performing a Taylor expansion[11] of the end measure ν in terms of the starting measure μ, we have that $\widetilde{u}$ and $\widetilde{u}$ basically decouple, and the integral in $d\gamma$ can be expressed in terms of its expression. Namely:

$$\begin{aligned}
\int \Phi(\widetilde{u}) - \Phi(u) d\gamma(u, \widetilde{u}) &= \int \Phi(\widetilde{u}) d\gamma(u, \widetilde{u}) - \int \Phi(u) d\gamma(u, \widetilde{u}) \\
&= \int \Phi(\widetilde{u}) d\pi^2_\# \gamma(u, \widetilde{u}) - \int \Phi(u) d\pi^1_\# \gamma(u, \widetilde{u}) \\
&= \int \Phi(\widetilde{u}) d\nu - \int \Phi(u) d\mu = \int d\Phi_u(\widetilde{u} - u) + M(u, \widetilde{u}) d\gamma(u, \widetilde{u}) \quad \text{above result}
\end{aligned}$$

Eventually:

$$\begin{aligned}
F^{(r)}(\nu) &= R\left(\underbrace{\int \Phi d\nu}_{=g} \right) \\
&= R\left(\underbrace{\int \Phi d\mu}_{=f} \right) + \left\langle R'\left(\int \Phi d\mu \right), \int \Phi d\nu - \int \Phi d\mu \right\rangle + \underbrace{N\left(\int \Phi d\nu, \int \Phi d\mu \right)}_{:=(I)} \\
&\quad \text{loss expansion} \\
&= F^{(r)}(\mu) + \left\langle R'\left(\int \Phi d\mu \right), \int d\Phi_u(\widetilde{u} - u) + M(u, \widetilde{u}) d\gamma(u, \widetilde{u}) \right\rangle + (I) \\
&\quad \text{parameter expansion} \\
&= F^{(r)}(\mu) + \left\langle R'\left(\int \Phi d\mu \right), \int d\Phi_u(\widetilde{u} - u) d\gamma(u, \widetilde{u}) \right\rangle + \underbrace{\left\langle R'\left(\int \Phi d\mu \right), \int M(u, \widetilde{u}) d\gamma(u, \widetilde{u}) \right\rangle}_{:=(II)} + (I) \\
&= F^{(r)}(\mu) + \left\langle R'\left(\int \Phi d\mu \right), \int d\Phi_u(\widetilde{u} - u) d\gamma(u, \widetilde{u}) \right\rangle + (II) + (I) \qquad (2.22)
\end{aligned}$$

[11] this is very similar to a classical interpolation

(♢ inequalities) Recall that we consider a highly localized viewpoint, so that $\nu \approx \mu$ in some sense. Clearly then $C_2(\gamma) \to 0$, and much faster than $C_1(\gamma)$ for example. In addition to this, we have also proved that h' is Lipschitz in $\bigcirc$. With these facts at disposal:

$$\begin{aligned}
|(I)| &= \left| \left\langle R'(\textstyle\int \Phi d\mu), \int M(u,\widetilde{u}) d\gamma(u,\widetilde{u}) \right\rangle \right| \\
&\leq \left\| R'(\textstyle\int \Phi d\mu) \right\| \left\| \int M(u,\widetilde{u}) d\gamma(u,\widetilde{u}) \right\| \\
&\leq \|dR\|_{\infty,r} \, C_1(\gamma) \int \|M(u,\widetilde{u})\| \, d\gamma(u,\widetilde{u}) \\
&\leq \|dR\|_{\infty,r} \, C_1(\gamma) \int \frac{1}{2} L_{d\Phi} |\widetilde{u} - u|^2 d\gamma(u,\widetilde{u}) \\
&= \frac{1}{2} \|dR\|_{\infty,r} \, C_1(\gamma) L_{d\Phi} \underbrace{\int |\widetilde{u} - u|^2 d\gamma(u,\widetilde{u})}_{\leq C_2^2(\gamma)} \\
&= \mathfrak{o}(C_2(\gamma)) && \text{discarding the rest as } C_2(\gamma) \to 0
\end{aligned}$$

and similarly:

$$\begin{aligned}
|(II)| &= \left| N(\textstyle\int \Phi d\mu, \int \Phi d\nu) \right| \\
&\leq \frac{1}{2} L_{dR} \left\| \textstyle\int \Phi d\nu - \int \Phi d\mu \right\|^2 && \text{last term is } \|g - f\|^2 \\
&= \frac{1}{2} L_{dR} \left\| \textstyle\int d\Phi_u(\widetilde{u} - u) d\gamma(u,\widetilde{u}) + \int M(u,\widetilde{u}) d\gamma(u,\widetilde{u}) \right\|^2 && \text{as before} \\
&\leq \frac{1}{2} L_{dR} \left[\|d\Phi\|_{\infty,r} \, C_1(\gamma) + \frac{1}{2} L_{d\Phi} C_2^2(\gamma) \right]^2 \\
&= \mathfrak{o}(C_2(\gamma)) && \text{discarding the rest}
\end{aligned}$$

Where we reused an implication of Hölder's inequality (again $C_1^2(\gamma) \leq C_2^2(\gamma)$, see Subsec. B.3).

(♣ interior and boundary velocity field) Eventually, with a little trick we take the integral in $d\gamma$ out of the inner product of Equation 2.22 and see that:

$$F^{(r)}(\nu) = F^{(r)}(\mu) + \int \left\langle R' \left(\int \Phi d\mu \right), d\Phi_u(\widetilde{u} - u) \right\rangle d\gamma(u,\widetilde{u}) + \mathfrak{o}(C_2(\gamma)) \tag{2.23}$$

Recalling that in Proposition 2.3 we showed:

$$\nabla F'(\mu) : \Omega \to \mathcal{F} \quad u \to \left[\left\langle R'(\textstyle\int \Phi d\mu), d\Phi_u(e_j) \right\rangle \right]_{j=1}^d$$

The above result means that **in the interior** of Q_r the velocity is characterized by Claim #3. Going to the **boundary** of Q_r, the authors claim that by the fact that $\pi^2_\# \gamma = \nu$ is restricted to be localized in Q_r, a point u in the boundary of Q_r is such that the difference between the two characterizations:

$$v(u) - \nabla F'(\mu)(u)$$

can *live in the normal cone* of Q_r at u. Such normal cone is in our claim $\partial \iota_{Q_r}(u)$.
In a general setting, we can safely say that $v(u) \in \partial(F'(\mu) + \iota_{Q_r})(u)$ where the second element is influential only at the boundary. □

Observation 2.24 (About the lemma)**.** *We specify that:*

- μ_t^γ *is the projection of gamma over an interpolation of the starting set and the arrival set. We call this* ***transport interpolation****. It is such that:*
  - *at* $t = 0$ *the measure* μ_t^γ *is a projection from the starting set (hence the* 1 *as apex of* π*)*
  - *at* $t = 1$ *equivalently it is the result of a projection from the arrival set*
- *the Lipschitz bound of Claim #2* λ_r *depends on the radius* r*:*
  - *may explode if the measure were not concentrated on the ball* Q_r*, thus the technical requirement in Assumption 1.10*
  - *in literature, it is also referred to as* $-\lambda_r$ *geodesical semiconvexity [AGS05]*

Already proved results in the theory of Wasserstein gradient flows guarantee that a gradient flow for Equation 1.9 as in Definition 2.9 is well defined.

Lemma 2.25 (Existance and uniqueness of Wgf for $F^{(r)}$)**.** *Under Assumptions 1.10:*

$$\exists!(\mu_t^{(r)})_{t\geq 0} \quad \forall \mu_0 \in \mathcal{P}_2(\Omega), \textit{continuous}$$

Which is a Wgf for $F^{(r)}$*, thus such that:*

$$\partial_t \mu_t^{(r)} = -\texttt{div}(v_t^{(r)}\mu_t^{(r)}) \qquad \textit{continuity Eqn. 2.10}$$
$$v_t^{(r)}(u) \in \partial(F'(\mu_t^{(r)})(u) + \iota_{Q_r}(u)),\ \forall t > 0, \mu_t^{(r)}\ a.e.\ u \in \Omega \qquad \textit{Lemma 2.18\#3}$$

Proof. **(△ adding V)** by Assumptions 1.10#2 V is semiconvex (say λ_V-semiconvex, according to Def. A.17). Then, it holds that its integral wrt to a measure $\mu \to \int V d\mu$ is λ_V-semiconvex along generalized geodesics [AGS05](Def. 9.2.4, Prop. 10.4.2). By Lemma 2.18, it is rather easy to see that the semiconvexity of the base case and V results in $F^{(r)}$ being $(\lambda_V - \lambda_r)$-semiconvex along generalized geodesics.
(□ Wasserstein subdifferentials) the result of Lemma 2.18#3 holds with an added V, and we get that $F^{(r)}$ has strong Wasserstein subdifferentials as in the claim.
(◯ existance and uniqueness) by the results of △, □ existance and uniqueness of $(\mu_t^{(r)})_{t\geq 0}$ with the properties claimed (i.e. the first is the exact Definition 2.9, the second is □) is guaranteed by [AGS05](Thm. 11.2.1). □

Up to now, we have concentrated on $F^{(r)}$ and proved that there is a unique Wgf under a concentration condition of the flow inside a ball. Namely, if $\exists r_0$ such that $\mu_t^{(2r_0)}$ is concentrated on Q_{r_0} $\forall t \in [0, T]$ then all the $F^{(r)} : r > r_0$ have identic Wgfs. We set a 2 coefficient to make the r of the gradient flow bigger than its existance condition and aim to make the support grow slower, ensuring existance for the whole interval $[0, T)$.

Proposition 2.26 (Existance and uniqueness of Wgf for F)**.** *Under Assumptions 1.10, we have that if* $\mu_0 \in \mathcal{P}_2(\Omega)$ *is concentrated on* $Q_{r_0} \subset \Omega$*:*

$$\exists!(\mu_t)_{t\geq 0} \quad Wgf:$$
$$v_t(u) = \widetilde{v}_t(u) - \texttt{proj}_{\partial V(u)}(\widetilde{v}_t(u)) \qquad \widetilde{v}_t(u) = -\left[\left\langle R'\left(\int \Phi d\mu_t\right),\ \partial_j \Phi(u)\right\rangle\right]_{j=1}^{d}$$

Where the last Equation is that of Proposition 2.3 according to the connection made in Prop. 2.11, ignoring the m *discretization, since we are working in the continuous case to be more general.*

Proof. **(△ the discrete case)** in Proposition 2.3, we worked with an m parametrization. A such, the concentration requirement we impose here, is automatically satisfied by the finite discrete support of $\mu_{m,0}$ $\forall m$ up to large enough r. Thus, we ignore this special case and work without m.
(□ bridging $F^{(r)}$ and F) let $r_0 : \mu_0(Q_{r_0}) = 1$. Using Lemma 2.25 $\forall r > r_0 \exists! (\mu_t^{(r)})_{t\geq 0}$ of $F^{(r)}$. For all larger radii $r > r_0$ the exit time from the ball Q_r denoted as t_r is formalized as:

$$t_r \coloneqq \inf \left\{ t > 0 \; ; \; \mu_t^{(2r)}(Q_r) < 1 \right\}$$

In other words, the first time that increasing the radius the measure is not anymore concentrated in the ball. Notice that in such assignment, the 2 is superfluous and just chosen as reference. Indeed, by Lemma 2.25 we have existance and uniqueness up to the condition $\overline{r} > r_0$ so $\forall \overline{r} > r_0$ it holds $(\mu_t^{(\overline{r})})_{t\geq 0} \equiv (\mu_t^{(2r)})_{t\geq 0}$ on the interval $[0, t_r]$ where the Wgf is in accordance with Definition 2.9.
For this reason, if:

$$\lim_{r\to\infty} t_r = \infty \tag{$\star$}$$

Then the original F, obtained as a limit of the restriction $F^{(r)}$, has a **globally defined** Wasserstein Gradient Flow.
(◯ proving $\star$) In the interval $0 \leq t \leq t_r$ we are not exceeding the ball Q_r. Then, by Lemma 2.18#3 we have that $v_t \in \partial(F'(\mu_t^{(r)}) + \iota_{Q_r})$ in the space $L^2(\mu_t^{(r)}, \mathbb{R}^d)$ but $\iota_{Q_r} = 0$ in Q_r so that $v_t \in \partial F'(\mu_t^{(r)})$.
By Assumption 1.10#3-(c) and dR being bounded on sublevel sets (Ass. 1.10#1) we get from the latter an inequality and from the former a $\forall t$ statement:

$$|v_t^{(r)}(u)| \leq C_1 + C_2 r \quad \forall 0 \leq t \leq t_r \quad \textit{where } C_1, C_2 \perp\!\!\!\perp u, r, t \tag{2.27}$$

Using Gronwall's Lemma (Lem. A.31) on the representation of the Wasserstein Gradient Flow of Lemma A.71 we have a linear bound on the velocity of the particles that translates to an exponential bound of the distance from the starting configuration:

$$\textit{Eqn.2.27} \overset{\substack{\textit{Lem.A.71} \\ \Longrightarrow \\ \textit{Lem.A.31}}}{} \textit{dist}(u, Q_{r_0}) \leq \left(r_0 + \frac{C_1}{C_2} \right) e^{tC_2} \quad \forall 0 \leq t \leq t_r$$

Which is just solving the differential equation on the easy exponential bound for a linear growth function. Thus, $\mu_t^{(r)}$ is concentrated on

$$\left\{ u \in \Omega \; ; \; \textit{dist}(u, Q_{r_0}) \leq \left(r_0 + \frac{C_1}{C_2} \right) e^{tC_2} \right\} \quad \forall 0 \leq t \leq t_r$$

By this fact, $\forall T > 0 \exists r > 0 : t_r > T$. Indeed, as $r \to \infty$ the bounds get more and more relaxed and we can informally reach any point from Q_{r_0} at an arbitrary big T ending time, without losing the existance and uniqueness condition guaranteed by being below the exit time.
We have built a dynamical system that always guarantees a well behaved Wasserstein Gradient flow, and by Proposition 2.11 this brings back to the claimed form of dynamics in terms of the parameters. □

Now that we have information about arbitrary measures and correspondent Wasserstein gradient flows, we eventually explore the situation in which a flow in time t of the parameters of m

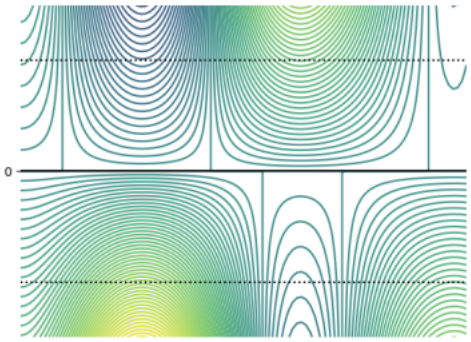


Figure 2: Homogeneous function. Source [CB18]
Partial 1-homogeneous F derivative level sets, where in the vertical axis for $\Omega = \mathbb{R}^2$ makes the landscape of $F'(\mu)$ look like this, subject to a dependence on μ_t itself at each time step. A Wasserstein Gradient Flow would follow down these curves, given that we impose an update of the form $-\partial F'(\mu_t)$. In Section 3 we will see that a minimizing measure is characterized by positivity of $F'(\mu)$ in the domain and nullity over the support of μ. The Homogeneous structure makes vertical directions non-influent for inferring properties. It is sufficient to focus on the behavior at the two dotted lines, which cover the positive and negative weight case.

particles[12] has some convergent (in m) behavior to a *diffuse* measure μ. Does this convergence extend if it is only at $t = 0$? How could we ensure that it holds for any t? In other words, is it possible to take a parameter Gradient Flow convergent at the start, and interpret it as $m \to \infty$ as a Wasserstein gradient flow for a continuous measure? Does the convergence break over time?

2.3 Conciliating particles and diffuse optimization: the many-particle limit

So far we have treated F (Eqn. 1.9) and F_m (Eqn. 2.1) separately, the aim of the next results is to highlight that the limit of the discretized version corresponds to the continuous one. This is a very deep result in practice, since the **feasible architectures** described by such construction can only have finite *particles* (see our final application in Section 4 for an idea). Nevertheless, an equivalence at the limit guarantees that simulations have a nice behavior at large enough m.

Theorem 2.28 (Many-particle limit)**.** *Under Assumptions 1.10, consider a sequence in m of gradient flows (Def. 2.2) for F_m of the form:*

$$(t \to \boldsymbol{u}_m(t))_{m\in\mathbb{N}}$$

which are initialized at $\mu_{m,0}$ concentrated in $Q_{r_0} \subset \Omega$. Then:

$$\lim_{m\to\infty} \|\mu_{m,0} - \mu_0\|_{W_2} = 0 \quad \mu_0 \in \mathcal{P}_2(\Omega) \implies (\mu_{m,t})_{t\geq 0} \overset{m\to\infty}{\underset{W_2}{\rightrightarrows}} (\mu_t)_{t\geq 0}$$

Where $(\mu_t)_{t\geq 0}$ is the unique (and existent) Wgf of F that starts at μ_0. Namely, if our discrete starting point converges in $\mu_0 \in \mathcal{P}_2(\Omega)$, then the whole discrete sequence converges to the continuous version of the same problem.

[12] represented as a measure by Prop. 2.11

Proof. **(classic proof)** The authors point out how there are results in the literature that ensure this through a *discretization in time* [AGS05](Thm. 11.2.1). □

Proof. **(△ *direct* proof approach)** another possibility is performing a *discretization in space.* **(□ exit time lower bound)** the aim is to identify an interval for the discrete paths to be contained in Q_r for $r > r_0$. The exit time in this case is:

$$t_r \coloneqq \inf\left\{ t > 0 \ ; \ \exists m \in \mathbb{N}, \mu_{m,t}(Q_r) < 1 \right\}$$

To have an interval $[0, t_r]$ it must be the case that $t_r > 0$. The evolution of the measure is uniquely determined by the velocities before t_r, that have the form of Proposition 2.3, Equation 2.4. In this setting, we let:

- $L_{V,r}$ be the Lipschitz constant of V in Q_r
- $r - r_0$ the minimum travel distance
- velocities $\widetilde{v}_t(u) = -\left[\langle R'(\int \Phi d\mu_{m,t}), \partial_j \Phi(u) \rangle \right]_{j=1}^{d}$; $\quad v_t(u) = \widetilde{v}_t(u) - \texttt{proj}_{\partial V(u)} \widetilde{v}_t(u)$

Then, the exit time is clearly bounded by $min_{space}/|max_{velocity}|$ which means:

$$\begin{aligned}
t_r &\geq \frac{r - r_0}{|v_t(u)|} \\
&= \frac{r - r_0}{|\widetilde{v}_t(u) - \texttt{proj}_{\partial V(u)} \widetilde{v}_t(u)|} \\
&\geq \frac{r - r_0}{|\widetilde{v}_t(u)| + |-\texttt{proj}_{\partial V(u)} \widetilde{v}_t(u)|} && \text{use Lip. on bounded sets assumption} \\
&= \frac{r - r_0}{\|d\Phi\|_{\infty,r} \|dR\|_{\infty,r} + L_{V,r}} \\
&> 0 && \text{Prop. 2.16, and } L_{V,r} > 0
\end{aligned}$$

(◯ limit curve in interval) by △ we can work on an interval, and aim to show that a discretized flow converges in $\mathcal{P}_2(\Omega)$ to a limit $t \to \mu_t$.
For two subsequent time points $0 \leq t_1 < t_2 \leq t_r$, we have that the distance between the two measures is:

$$\begin{aligned}
W_2(\mu_{m,t_1}, \mu_{m,t_2})^2 &= \min_{\gamma \in \Pi(\mu_{m,t_1}, \mu_{m,t_2})} \int |\widetilde{u} - u|^2 d\gamma(u, \widetilde{u}) && \text{Def. A.63 for } p = 2 \\
&\leq \frac{1}{m} \sum_{i=1}^{m} |\mathbf{u}_{m,i}(t_2) - \mathbf{u}_{m,i}(t_1)|^2 && \text{see below} \\
&= \frac{t_2 - t_1}{m} \sum_{i=1}^{m} \int_{t_1}^{t_2} |\mathbf{u}'_{m,i}(s)|^2 ds && \mathbf{u} \text{ is diff by 2.3 \& Jensen's} \\
&\leq \frac{t_2 - t_1}{m} \int_{t_1}^{t_2} \sum_{i=1}^{m} |\mathbf{u}'_{m,i}(s)|^2 ds && \text{linearity}
\end{aligned}$$

Where the first passage follows by matching each particle at t_1 to its future position at t_2. Clearly, the inequality has a greater since it is *less precise* than the minimization over measures of the same quantity. The application of Jensen's inequality also follows the unnormalized

version explained in Proposition 2.3.
By the result of Proposition 2.3 we can further state that:

$$
\begin{aligned}
W_2(\mu_{m,t_1}, \mu_{m,t_2}) &\leq \sqrt{\frac{t_2 - t_1}{m} \int_{t_1}^{t_2} \sum_{i=1}^{m} |\mathbf{u}'_{m,i}(s)|^2 ds} \\
&= \sqrt{t_2 - t_1} \sqrt{\int_{t_1}^{t_2} -\frac{d}{ds} F(\mu_{m,s}) ds} \qquad \text{Prop. 2.3, } -\frac{1}{m} \sum_{i=1}^{m} |\mathbf{u}'_{m,i}(t)|^2 = \frac{d}{dt} F(\mu_{m,t}) \\
&\leq \sqrt{t_2 - t_1} \sqrt{\sup_m F(\mu_{m,0}) - \inf_{\mu \in \mathcal{P}(\mathbb{R}^d)} F(\mu)}
\end{aligned}
$$

Where the last passage is a very rough bound concerning the supremum possible of F_m[13], and the infimum possible value of F over any measure. We cannot have an improvement greater than this for obvious reasons.
By the sup over m and the dependence on $0 \leq t_1 < t_2 \leq t_r$ we have that the family of flows indexed by m:

$$(t \to \mu_{m,t})_m$$

is equicontinuous (same convergence $\perp\!\!\!\perp m$) in W_2 over $[0, t_r]$ (as $t_2 \to t_1 \implies \|\cdot\|_{W_2} \to 0$).
In addition to this, the collection is constructed inside a W_2 ball, which is weakly precompact (Def. A.39) but not a piori compact [CB18].
Since the weak topology (i.e. the weakest topology) is weaker than the topology induced by W_2 we can apply Arzelà–Ascoli[14] to extract a subsequence $k \to m(k)$ such that:

$$\left((\mu_{m(k),t})_{k \in \mathbb{N}}\right)_{t \geq 0} \underset{k \to \infty}{\overset{w}{\to}} (\mu_t)_{t \geq 0}$$

preserving in the weak topology continuity and concentration on $Q_r \forall t \in [0, t_r]$.
Additionally, by the result of Prop. A.58 we have the convergence in $\|\cdot\|_{BL}$ (Def. A.57) which metrizes[15] weak convergence in $\mathcal{P}_2(\Omega)$ and we can use this.
From now on, we denote the (weakly) converging subsequence as $(\mu_m)_{m \in \mathbb{N}}$
(♢ continuity equation at the limit) as of now, by ◯ we have that discrete measures converge, what about the velocities? Let $(v_t)_{t \geq 0}$ be the limit of $(v_{m,t})_{t \geq 0}$. What we actually need is that the continuity equation 2.10 characterizing Definition 2.9 is satisfied as $m \to \infty$, making the flow a Wgf at divergent number of particles. For this reason, we define *momenta* of the flow as:

$$(E_m)_{m \in \mathbb{N}} \quad E_m : [0, t_r] \times \Omega \to \mathbb{R} \quad E_m \coloneqq v_{m,t} \mu_{m,t} dt$$

Where we know that $\forall r > 0$ such quantities are concentrated on Q_r by construction. What we need to show is that:

$$E_m \underset{m \to \infty}{\overset{w}{\to}} E \coloneqq v_t \mu_t dt$$

Indeed, weak convergence would guarantee that in the limit we will get the unique Wgf of the diffuse case that naturally satisfies the continuity equation (Prop. 2.26). Using weak convergence (Def. A.56) we inspect integrals of bounded and continuous functions of the form:

$$\varphi : [0, t_r] \times \mathbb{R}^d \to \mathbb{R}^d$$

[13] i.e. its *biggest* start
[14] Thm. A.37 for an intuition, Thm. A.40 for slightly more than what we need
[15] namely, it inherits all the metric properties, properly, Proposition A.58

for which we say:

$$\begin{aligned}
\left|\int \varphi dE_m - \int \varphi dE\right| &= \left|\int \varphi \cdot d(E_m - E)\right| \\
&= \left|\int \varphi \cdot d(v_{m,t}\mu_{m,t}dt - v_t\mu_t dt)\right| \\
&= \left|\int \varphi \cdot (v_{m,t}(u) - v_t(u))d\mu_{m,t}dt + \int \varphi \cdot v_t d(\mu_{m,t} - \mu_t)dt\right| \\
&\leq \left|\int \varphi \cdot (v_{m,t}(u) - v_t(u))d\mu_{m,t}dt\right| + \left|\int \varphi \cdot v_t d(\mu_{m,t} - \mu_t)dt\right| \\
&\leq \int \left|\varphi \cdot (v_{m,t}(u) - v_t(u))\right| d\mu_{m,t}dt + \left|\int \varphi \cdot v_t d(\mu_{m,t} - \mu_t)dt\right| \\
&\leq \|\varphi\|_\infty \int \left|v_{m,t}(u) - v_t(u)\right| d\mu_{m,t}dt + \left|\int \varphi \cdot v_t d(\mu_{m,t} - \mu_t)dt\right|
\end{aligned}$$

We show separately that the red and the blue terms are null.
(♣ red is zero) we know that $(\mu_{m,t})_{m,t}$ are concentrated on Q_r, so for the first integral it is sufficient to show that the velocities inside the modulus converge uniformly in m in the time interval , namely:

$$wts \quad (t,u) \to v_{m,t}(u) \overset{m}{\rightrightarrows} v_t(u) \leftarrow (t,u) \quad \forall (t,u) \in [0,t_r] \times \Omega$$

In this setting, by the fact that the projection on a convex ($\partial V(u)$ is by Prop. A.27) set is 1-Lipschitz (Prop. A.24) we have that:

$$\begin{aligned}
|v_{m,t}(u) - v_t(u)| &= |\widetilde{v}_{m,t}(u) - \texttt{proj}_{\partial V(u)}\widetilde{v}_{m,t}(u) - \widetilde{v}_t(u) + \texttt{proj}_{\partial V(u)}\widetilde{v}_t(u)| \\
&\leq |\widetilde{v}_{m,t}(u) - \widetilde{v}_t(u)| + |-\texttt{proj}_{\partial V(u)}\widetilde{v}_{m,t}(u) + \texttt{proj}_{\partial V(u)}\widetilde{v}_t(u)| \\
&\leq |\widetilde{v}_{m,t}(u) - \widetilde{v}_t(u)| + |\widetilde{v}_{m,t}(u) - \widetilde{v}_t(u)| \qquad \text{Prop. A.24} \\
&= 2|\widetilde{v}_{m,t}(u) - \widetilde{v}_t(u)| \\
&= 2|[\langle R'(\textstyle\int \Phi d\mu_{m,t}), \partial_j \Phi(u)\rangle]_{j=1}^d - [\langle R'(\textstyle\int \Phi d\mu_t), \partial_j \Phi(u)\rangle]_{j=1}^d| \\
&= 2|[\langle R'(\textstyle\int \Phi d\mu_{m,t}) - R'(\textstyle\int \Phi d\mu_t), \partial_j \Phi(u)\rangle]_{j=1}^d \\
&\leq 2\,\|d\Phi\|_{\infty,r} \left\|R'(\textstyle\int \Phi d\mu_{m,t}) - R'(\textstyle\int \Phi d\mu_t)\right\|
\end{aligned}$$

Where the second term in the specific interval $[0,t_r]$ derived in $\square$ is such that

$$\begin{aligned}
\left\|R'(\textstyle\int \Phi d\mu_{m,t}) - R'(\textstyle\int \Phi d\mu_t)\right\| &\leq \|dR\|_{\infty,r} \left\|\textstyle\int \Phi d\mu_{m,t} - \int \Phi d\mu_t\right\| \\
&\leq \|dR\|_{\infty,r} \sup_{f \in \mathcal{F}, \|f\| \leq 1} \int \langle f, \Phi(u)\rangle \, d(\mu_{m,t} - \mu_t) \qquad (2.29) \\
&\leq \|dR\|_{\infty,r} \max\left\{ \|\Phi\|_{r,\infty}, \ \|d\Phi\|_{r,\infty} \right\} \|\mu_{m,t} - \mu_t\|_{BL}
\end{aligned}$$

Where in Equation 2.29 we used two tricks:

- the norm of an element in a Hilbert space is bounded by the sup of an inner product (Prop. A.16 with $\leq$ instead)
$$\|z\| \leq \sup_{f \in \mathcal{F}, \|f\| \leq 1} \langle z, f\rangle \quad \forall z \in \mathcal{F}$$

- we make a **slightly informal** switch for an arbitrary measure in the infinite dimensional Hilbert space. If for a moment we think of $\mathcal{F}$ as a k dimensional space, we could write:

$$\langle f, \int \Phi d\sigma \rangle = \begin{bmatrix} \int \Phi_1 d\sigma \\ \vdots \\ \int \Phi_k d\sigma \end{bmatrix} \cdot \begin{bmatrix} f_1 \\ \vdots \\ f_k \end{bmatrix} = \sum_{i=1}^{k} f_k \int \Phi_k d\sigma = \int \sum_{i=1}^{k} f_k \Phi_k d\sigma = \int \langle f, \Phi \rangle \, d\sigma$$

  in the real case of a Hilbert space, this would require the infinite sum to be exchangeable with the integral. In our setting, given that we find a finite bound, it is feasible by Fubini-Tonelli Theorem.

so that:

$$\begin{aligned} \|\int \Phi d\mu_1 - \int \Phi d\mu_2\| &= \|\int \Phi d(\mu_1 - \mu_2)\| \\ &\leq \sup_{f \in \mathcal{F}, \|f\| \leq 1} \left\langle f, \int \Phi d(\mu_1 - \mu_2) \right\rangle \\ &= \sup_{f \in \mathcal{F}, \|f\| \leq 1} \int \langle f, \Phi(u) \rangle \, d(\mu_{m,t} - \mu_t)(u) \end{aligned}$$

Coming back to our inequality:

$$\left\| R'(\int \Phi d\mu_{m,t}) - R'(\int \Phi d\mu_t) \right\| \leq \|dR\|_{\infty,r} \max \left\{ \|\Phi\|_{r,\infty}, \ \|d\Phi\|_{r,\infty} \right\} \|\mu_{m,t} - \mu_t\|_{BL}$$

the first two terms are finite by previous discussions.
By the uniform in BL norm convergence of measures of the form $(t \to \mu_{m,t})_{m \in \mathbb{N}}$ discussed in ◯ we get that the red term is zero at the limit.
(♠ blue is zero) for similar reasons by:

- $\mu_{m,t} \overset{w}{\to} \mu_t$
- the map $(t, u) \to \varphi(t, u) v_t(u)$ being continuous and bounded. To see this, recall that φ is a test function and the discussion of ◯, ♢.

It is quick to realize that the blue term:

$$\left| \int \varphi \cdot v_t d(\mu_{m,t} - \mu_t) dt \right| \overset{m \to \infty}{\to} 0$$

By the results of ♣, ♠ we finally proved ♢ and get $E_m \overset{w}{\to} E$, with the limiting measure being the one that satisfies the continuity equation.
(▽ integrability) A technical missing fact is that of the integrability condition of the continuity equation (namely, Eqn. B.3). This is trivially valid by the boundedness of the sequence $(v_t)_{t \geq 0}$ in a ball Q_r, uniformly in time (see the ◯ part of this proof and the beginning of ♢), ensuring that:

$$\int_0^{t_r} \int_\Omega |v_t(u)|^2 d\mu_t(u) dt < \infty$$

Which, implies absolute continuity in W_2, a requirement of Definition 2.9. For more context (but not too much), see Appendix B.2.
(♡ globality) summarizing, we have convergence up to a subsequence (◯) to a Wgf (♢) in the interval $[0, t_r]$ (□). To ensure global convergence, we are left to show that:

$$\lim_{r \to \infty} t_r = \infty$$

In the interval $[0, t_r]$ we have:

$$\begin{cases} F(\mu_{m,0}) \overset{m\to\infty}{\to} F(\mu_0) \\ F((\mu_{m,t})_{t\in[0,t_r]}) \searrow F(\mu_{m,t_r}) \end{cases}$$

which implies that the dynamics lie in the $F(\mu_{m,0})$ subset of R. Again, by Assumption 1.10#1 dR is bounded. By the very form of the velocity (available in Prop. 2.3) it follows that there is a uniform bound on the velocity **linear in** r. Again as in Proposition 2.26 we have the same form of a bound and can apply Gronwall's inequality (Lem. A.31) to show $\lim_{r\to\infty} t_r = \infty$. Given the unique Wgf by Proposition 2.26, the subsequence forces the sequence to converge to it, and we have proved all of the claims without ambiguity. □

Observation 2.30 (Comments about the theorem)**.** *We have provided two existance proofs of a valid Wgf in Proposition 2.26, and Theorem 2.28.*

As an example, consider a measure $\mu_0 \in \mathcal{P}_2(Q_{r_0})$. If we want to build a sequence converging in W_2 to it, we can simply choose a flow in the parameters governed by the size m:

$$\mathbf{u}_m(0) = (u_1, \dots, u_m) \quad u_i \overset{iid}{\sim} \mu_0 \quad \forall i = 1, \dots, m$$

Namely, parameters picked at random from the diffuse measure μ_0. Then by the CLT the sequence:

$$\mu_{m,0} = \frac{1}{m}\sum_{i=1}^{m} \delta_{u_i} \quad \mu_{m,0} \overset{\text{a.s.-}W_2}{\underset{w}{\to}} \mu_0$$

Where the convergence is almost surely for W_2, the norm of the underlying complete space $\mathcal{P}_2(\Omega)$) (Prop. A.65 states this). For more precise statement, we have that the empirical measure is a.s. weakly convergent [Dud02](Thm. 11.4.1), since the space $(\mathcal{P}_2(\mathbb{R}^d))$ is separable and complete [Bol08]. Apart from the technicalities, what we care about is that we will be able to approximate integrals with respect to μ such as those in F by using its discrete approximation.

3 Convergence to Global Minimizers

In this Section, we conclude the theoretical notions needed to describe the application. The first step is a brief discussion on the difference of measures that are functional-stationary and measures which are also optimal. Having characterized our target, the problem is that a trivial flow of measures would get stuck at local minima. To overcome this issue, a finer set of Assumptions is provided, under which, subject to a condition on the measure given to a specific set A at a point in which the flow is close to a local minima, it is possible to escape from this ϵ-closeness at a **finite time** future point.
The problem is then reformulated as finding a condition on the starting measure such that this weight assignment is verified at any t, allowing the measure to **always** escape local minima. This result is reached through a *topology detour* that returns the final requirement for the starting measure, bound to be preserved throughout the dynamics. We call this separation, and it is a condition entirely satisfied by the support of μ_0. All the properties together under the new Assumptions are the aspects of the final Theorem. A properly setup Wasserstein Gradient Flow will reach the global minimum of F.
As a side note, what remains as a problem is the actual convergence of the flow. Indeed, a global minima is reached assuming that $(\mu_t)_t$ *stops*. In this context, some references for future directions are mentioned at the end.

NB this Section in the original paper [CB18] is very technical. For this reason, I explored it less than the other three, and some parts are very similar to those of the publication[16]. The result we need for partial 1-homogeneity requires conclusions from the 2-homogeneous case, and some passages of the proofs are at a level of expertise which is higher than mine.

3.1 Stationary vs optimal points

A stationary point for a Wgf is clearly a measure $\mu \in \mathcal{P}_2(\Omega)$. The term stationary could be described further in simple terms as a measure that, if encountered along the path $t \to \mu_t$, makes it constant in that configuration. A characterizing condition arising from Definition 2.9 can be derived by noting that a stationary distribution is equivalent to zero velocity. Thus:

$$\mu \textit{ stationary at } t^* \iff v_t(u) = 0 \ \mu\text{-a.e. } u \in \Omega, \ \forall t \geq t^* \iff 0 \in \partial F'(\mu)(u) \quad \mu\text{-a.e. } u \in \Omega \tag{3.1}$$

Namely, to have zero velocity it must be that the Wasserstein subgradient of the dynamics actually includes the option for zero velocity.
Such stationary points are not always global minimizers of Equation 1.9, even adding the assumption of convexity of R. The problem is that even though R is convex, we are building a Wasserstein gradient flow, and not a Gradient Flow based on total variation [Bac20a]. Intuitively, in the finite case a Wgf is simply backpropagation on the problem, which is known **not to guarantee** global convergence a priori. The correct characterization is given below.

Observation 3.2 (About F and stationary vs optimal points)**.** *the authors make two important remarks:*

[16] in particular, Proposition 3.6, Lemma 3.14, Theorem 3.18.

- *a stationary measure μ could be optimal over probabilities that have smaller support [NS17]*
- *Using a Taylor expansion approach for $\mu, \sigma \in \mathcal{M}(\Omega)$ with $F(\mu), F(\sigma) < \infty$, given that Fréchet differentiability holds we can see in the first order the functional derivative is well defined*[17]*, so that $F'(\mu) \in L^1(\sigma)$. Further, recalling that $F'(\mu) : u \to \langle R'(\int \Phi d\mu), \Phi(u) \rangle + V(u)$ we get by the same reasoning that led to Equation 2.23 that:*

$$\begin{aligned}\frac{d}{d\epsilon} F(\mu + \epsilon\sigma)\Big|_{\epsilon=0} &= \frac{d}{d\epsilon}\left\{ F(\mu) + \left(\int F'(\mu) d\sigma \right) \epsilon + \mathcal{O}(\epsilon) \right\} \Big|_{\epsilon=0} \\ &= \int F'(\mu) d\sigma \\ &= \int \left\langle R'(\textstyle\int \Phi d\mu), \Phi \right\rangle + V \, d\sigma\end{aligned}$$

Proposition 3.3 (Minimizers with convexity characterization)**.** *Assume R is convex. Then:*

$$\mu \in \mathcal{M}_+(\Omega),\ F(\mu) < \infty,\ \mu = \underset{\mathcal{M}_+(\Omega)}{\arg\min} F \iff \begin{cases} F'(\mu) \geq 0 \\ F'(\mu)(u) = 0 \quad \textit{for } \mu\textit{-a.e. } u \in \Omega \end{cases}$$

The former condition broadly means that there is no room for decreasing the function, while the latter means that the measure is a stationary point of the PDE we impose (the continuity equation, Eqn. 2.10). An intuition is given in Figure 3.

Proof. **($\triangle$ strategy)** just like classic minimality proofs, we aim to start from an argument of the function (in this case a measure and a functional), perturb it, and see what happens at the neighborhood of it.
($\square$ perturbation decomposition) Let $\mu, \nu \in \mathcal{M}_+(\Omega)$ both satisfying $F(\mu) < \infty, F(\nu) < \infty$ sot that they are valid candidate minimizers. Consider their "distance measure" $\sigma = \mu - \nu$, which by Lebesgue decomposition (Thm. A.62) can be expressed in terms of μ as:

$$\sigma = f\mu + \mu^\perp \quad f \in L^1(\mu),\ \mu^\perp \in \mathcal{M}_+(\Omega),$$

With $f\mu \ll \mu$ and $\mu^\perp \perp \mu$ according to Definition A.60.
($\implies$ direction) this holds by the Taylor expansion we made in the previous paragraph. The hypothesis $\mu = \arg\min_{\mathcal{M}_+(\Omega)} F \implies F'(\mu) \geq 0$ otherwise we could improve it along some direction, with clearly the optimality condition that $F'(\mu)(u) = 0$ for μ-a.e. $u \in \Omega$.
($\impliedby$ direction) The convexity assumption ensures that the Taylor expansion is:

$$\begin{aligned}\frac{d}{dt} F(\mu + t\sigma)\Big|_{t=0} &= \int F'(\mu) d\sigma && \text{by above Obs.} \\ &= 0 && \text{by hyp of } F(\mu) = 0 \text{ for } \mu\text{-a.e. } u \in \Omega \\ &\leq \frac{d}{dt}((1-t)F(\mu) + tF(\nu))\Big|_{t=0} && \text{convexity assumption, } \sigma \text{ as in } \square \\ &= F(\nu) - F(\mu) && t \text{ vanished in } \frac{d}{dt} \text{ so } |_{t=0} \text{ is ignored}\end{aligned}$$

From which we summarize the result we need: $F(\nu) - F(\mu) \geq 0$, namely μ attains a lower value. By the arbitrariness of ν, μ is the minimizer of F in $\mathcal{M}_+(\Omega)$. $\square$

[17] i.e. $\int |F'(\mu)| d\sigma < \infty$

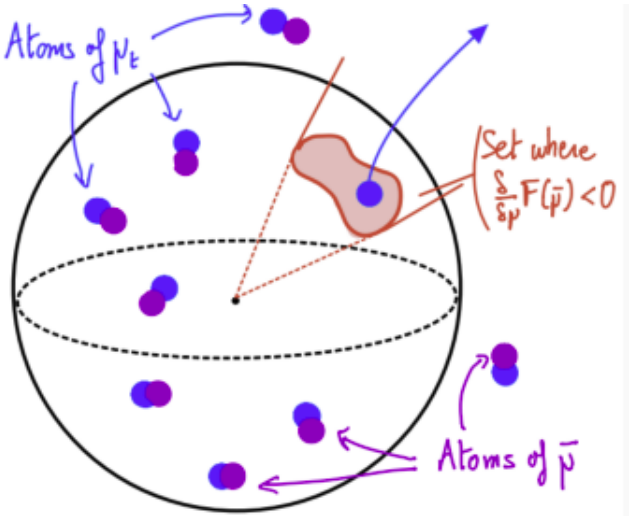


Figure 3: Abstract particle measures space. Source [Chi21]
$\overline{\mu}$ is not a stationary point if it puts weight on the set where $F'(\overline{\mu}) < 0$. We will show that a flow μ_t can escape from $\overline{\mu}$

There is a clear difference between the measure stationarity description in terms of F' (Eqn. 3.1) and the the globally optimal measure description of Proposition 3.3. The discussion of this section makes use of the specific structure we implemented when the problem was lifted (Subsection 1.4), especially its partial 1-homogeneity, and of a proper initialization. In practice, to ensure that Wgfs are allowed to[18] converge to a global minimizer:

- Φ and V need to have a **homogeneity direction**
- $\texttt{spt}\,\mu_0$ for the initial measure of the Wgf has to satisfy a **separation property**, which is **preserved** along the path.

Intuitively, like in the simplest case possible, stationarity is **not a synonym** of global minimality, even with the added assumption of convexity. The dynamics at any point, despite being constructed as to decrease the value of the functional, are guaranteed to stop at the first zero velocity measure. In order to escape such local minima traps, stronger assumptions are needed. Step by step, the **bold** items in the above list will be unrolled and commented.

3.2 Escaping non-optimal stationary points

We briefly introduce a common piece of notation and the technical details of the results that will be proved.

Remark. *When referring to $\mathbb{S}^{d-2}$ we mean the unit sphere of a space $\mathbb{R}^{d-1}$. Namely:*

$$\mathbb{S}^{d-2} := \left\{ \theta \in \mathbb{R}^{d-1} \; ; \; |\theta| = 1 \right\}$$

[18] convergence is not guaranteed, we look for conditions to avoid the issue of the dynamics stopping at local minima, which turns out to be solvable, but it is not the only issed. It could still be the case that the flow does not converge. More discussion is carried out at the end of the Section.

Assumption 3.4 (Partial 1-homogeneity suited assumptions)**.** *For a domain $\Omega = \mathbb{R} \times \Theta$ where $\Theta \subset \mathbb{R}^{d-1}$ the functions of Equation 1.9 take the form:*

$$\Phi(w, \theta) = w \cdot \phi(\theta) \quad V(w, \theta) = |w| \cdot \widetilde{V}(\theta)$$

Where both Φ and V are bounded, differentiable, with Lipschitz differential. Moreover:

1. *(smooth convex loss) R is convex, differentiable, dR is Lipschitz on bounded sets and bounded on sublevel sets*
2. *(Sard-type regularity) $\forall f \in \mathcal{F}$ (the domain of R) the set of regular values (Def. A.41) of the function*
$$g_f : \Theta \to \mathbb{R} \quad \theta \to g_f(\theta) = \langle f, \phi(\theta) \rangle + \widetilde{V}(\theta)$$
is dense in the range of g_f itself[19]
3. *(boundary conditions) $\phi : \theta \to \mathcal{F}$ behaves nicely at $\partial\Theta$ (the boundary of Ω. In particular, either of the following are true:*
   (a) *$\Theta = \mathbb{R}^{d-1}$ and $\forall f \in \mathcal{F}, \forall \theta \in \mathbb{S}^{d-2}$ the function g_f of the regular values is such that:*
   $$\lim_{r \to \infty} g_f(r\theta) \overset{C^2(\mathbb{S}^{d-2})}{\rightrightarrows} g_f^* \quad sat.\#2$$
   where it may be useful to precise that $\rightrightarrows$ is uniform convergence (Def. A.36) in the space mentioned and g_f^* is meant to satisfy the Sard-type regularity in $\#2$.
   (b) *Θ is the closure of an (open) bounded convex set and $\forall f \in \mathcal{F}$ the function g_f satisfies the Neumann boundary conditions:*
   $$d(g_f)_\theta(\vec{n}_\theta) = 0 \quad \forall \theta \in \partial\Theta$$
   with $\vec{n}_\theta \in \mathbb{R}^{d-1}$ the normal to $\partial\Theta$ at θ.

Remark. *The assumptions #3 are quite abstract. The former is in a nutshell a refinement of Sard-type regularity in terms of an expanding sphere that goes at the infinite limit of the space $\Theta = \mathbb{R}^{d-1}$, which is enforced to be uniform.*
The latter is even more technical as it achieves the same but in subsets of $\mathbb{R}$, not necessarily finite, but with a boundary that can be described (since they are bounded). For the purpose of this document, it is reported for completeness only, as our application in Section 4 will have $\Theta = \mathbb{R}^{d-1}$, which is #3-(a). At some points, the discussion will be less developed for the same reasons.

Proposition 3.5 (Linking old and new Assumptions)**.** *For nested sets of the form $Q_r \coloneqq [-r, r] \times \Theta$ with $r \in \mathbb{R}_+$ Assumptions 3.4 imply Assumptions 1.10. That is, the previous results still hold in the new setting.*

Proof. We recognize that both statements have a *setting* and 3 points. To distinguish them, we will use the subscripts $_{old}$ and $_{new}$, assuming that $settings_{new}, \#1_{new}, \#2_{new}, \#3_{new}$ hold.
$(settings_{old})$ clearly $\Omega \subset \mathbb{R}^d$ being the closure of a convex open set is cleared by our construction in both #3-(a) and #3-(b) $(\#1_{old})$ in $\#1_{new}$ we have only added convexity of R, which is also a safe assumption.

[19] in other words, the regular values of g_f, a function defined in terms of the f considered, can approximate any value of g_f with arbitrary precision

(#2$_{old}$, intuitively) the density of regular values ensures that the space of functions generated is separable (i.e. it contains a dense countable subset). For the rest of the claims, we refer to the authors statement [CB18]. □

To avoid going back and forth, we rewrite the projection map we used in Subsection 1.4 at Equation 1.12:

$$h^1 : \mathcal{M}_+(\Omega) \to \mathcal{M}(\Theta) \quad h^1(\mu)(B) = \int_{\mathbb{R}} w\mu(dw, B) \;\; \forall \mu \in \mathcal{P}(\Omega),\; B \subset \Theta \; measurable$$

Proposition 3.6 (Criteria to escape local minima)**.** *Let Assumptions 3.4 hold. Then:*

$$\mu \in \mathcal{M}(\Omega) \;:\; F'(\mu) < 0 \implies \exists \epsilon > 0, A \subset \Omega$$

such that if:

$$(\mu_t)_t \; Wgf \; sat \; \left\|h^1(\mu) - h^1(\mu_{t_0})\right\|_{BL} < \epsilon \; for \; t_0 \geq 0, \mu_{t_0}(A) > 0 \implies \exists t_1 > t_0 \;:\; \left\|h^1(\mu) - h^1(\mu_{t_1})\right\|_{BL} \geq \epsilon$$

Namely, for a non global minima to which the Wgf gets ϵ-close, the Assumptions allow the flow to escape at some positive future time point. The set A is:

$$A = (\mathbb{R}_+ \times K^+) \cup (\mathbb{R}_- \times K^-)$$

where:

- K^+ *is the* $-\eta$*-sublevel set of the map* $\theta \to F'(\mu)(1, \theta)$
- K^- *is the* $-\eta$*-sublevel set of the map* $\theta \to F'(\mu)(-1, \theta)$

With $\eta > 0$ *arbitrarily small.*

Proof. **(△ setting)** suppose that $F'(\mu)$ takes a negative value on $\mathbb{R}_+ \times \Theta$, the opposite case is worked out similarly.
Introduce the restriction of $F'(\mu)$ to the domain $\{1\} \times \Theta$. Recalling that $\Phi = w \cdot \phi$ and $V = |w| \cdot \widetilde{V}$ we call this restriction g_μ and write it explicitly:

$$g_\mu : \Theta \to \mathbb{R} \quad g_\mu(\theta) = F'(\mu(1, \theta)) = \left\langle R'\left(\int \Phi d\mu \right), \phi(\theta) \right\rangle + \widetilde{V}(\theta)$$

(□ the regular values) let $-\eta < 0$ be a negative regular value (Def. A.41) of the function g. Such a regular value is guaranteed to exist (Ass. 3.4#2). As per the claim, couple $-\eta$ with $K^+ \subset \Theta$ its sublevel set. By the regular value Theorem (Thm. A.44) the boundary:

$$\partial K^+ = g_\mu^{-1}(-\eta)$$

is a differentiable orientable[20] manifold of dimension $d-1-1 = d-2$, orthogonal to the gradient field (a vector field) of g_μ. Now, we distinguish between the scenarios of Assumption 3.4#3:

- Θ bounded $\implies$ ∂K^+ compact [21] $\implies$ $\exists \beta > 0$: $\inf_{\theta \in \partial K^+} |dg_\mu(\theta)| \geq \beta$, since on a compact set we have an inf by Weierstrass Thm. and such inf is necessarily > 0 by the definition of regular values that we work on.

[20] Differentiability comes from smoothness. We do not care much about the orientability, but trivially any Euclidean space is orientable since we can define coordinates on it with a homeomorphism to the space itself.

[21] K^+ is a sublevel set, so it is bounded since the function F' is bounded on sublevel sets by Assumption. The boundary is always closed (Prop. A.4, and it is additionally bounded since it is the boundary of a closed set.

- $\Theta = \mathbb{R}^{d-1}$ and the $(-\eta)$-sublevel set of K^+ is unbounded, we implement the construction of Assumption 3.4#3-(a). To do so, we force the η to be a regular value of the function g_μ^* on the limiting sphere such that $g_\mu \rightrightarrows g_\mu^*$. By imposing such further requirement:

$$\exists \beta > 0 \ : \ \inf_{\theta \in \partial K^+} |dg_\mu(\theta)| \geq \beta \implies \begin{cases} g_\mu \leq -\eta & \theta \in K^+ \\ \nabla g_\mu(\theta) \cdot \vec{n}_\theta \leq -\beta & \theta \in \partial K^+ \end{cases}$$

  where the first implication is actually the requirement for sublevel sets, and the second is a bound on the normal growth.

this is a Lemma included in the flow of the proof to avoid losing the flow of the exposition. Its purpose is that of showing that the properties of the set K^+ are true also for the g_ν of a measure close enough to μ. For simplicity, we denote:

$$\|f\|_{C^1} \coloneqq \max \left\{ \|f\|_\infty , \|\nabla f\|_\infty \right\}$$

i.e. the max of the sup-norm of function and gradient.

Lemma 3.7 (A bound on the norm of regular values for close measures)**.** $\forall C_0 > 0 \exists \alpha > 0$ *such that* $\forall \mu, \nu \in \mathcal{M}_+(\Omega)$ *with* $\left\|h^1(\mu)\right\|_{BL} < C_0, \left\|h^1(\nu)\right\|_{BL} < C_0$ *it holds:*

$$\|g_\nu - g_\mu\|_{C^1} \leq \alpha \|\phi\|_{C^1}^2 \cdot \left\|h^1(\mu) - h^1(\nu)\right\|_{BL}$$

Proof. The set $\{\int \Phi d\mu \ ; \ \mu \in \mathcal{P}(\mathbb{R}^d) \ ; \ h^1(\mu) < C_0\}$ is bounded in $\mathcal{F}$ by the fact that we impose a bound on the projection. Such boundedness, by Assumption 3.4 means that dR is Lipschitz, with a positive constant that we call α. Then:

$$\begin{aligned}
\|g_\nu - g_\mu\|_{C^1} &= \left\| \langle R'(\textstyle\int \Phi d\nu), \phi(\theta) \rangle + \widetilde{V}(\theta) - \langle R'(\int \Phi d\mu), \phi(\theta) \rangle - \widetilde{V}(\theta) \right\| \\
&= \left\| \langle R'(\textstyle\int \Phi d\nu) - R'(\int \Phi d\mu), \phi(\theta) \rangle \right\| \\
&\leq \left\| R'(\textstyle\int \Phi d\nu) - R'(\int \Phi d\mu) \right\| \|\phi\| \\
&\leq \alpha \left\| \textstyle\int \Phi d\nu - \int \Phi d\mu \right\| \|\phi\| && \text{Lipschitz } R \\
&\leq \alpha \left\| \textstyle\int \Phi d\nu - \int \Phi d\mu \right\| \|\phi\|_{C^1} \\
&\leq \alpha \left\| \textstyle\int \phi dh^1(\nu) - \int \phi dh^1(\mu) \right\| \|\phi\|_{C^1} && \text{Prop. A.54} \\
&\leq \alpha \|\phi\|_{C^1} \sup_{f \in \mathcal{F}, \|f\| \leq 1} \int \langle f, \phi \rangle \, d(h^1(\nu) - h^1(\mu)) && \text{Like for Eqn. 2.29} \\
&\leq \alpha \|\phi\|_{C^1} \|\phi\|_{C^1} \left\| h^1(\nu) - h^1(\mu) \right\|_{BL}
\end{aligned}$$

Where the last passage holds since the map $u \to \langle f, \phi(u) \rangle$ is $\|\phi\|_{C^1}$-Lipschitz and upper bounded in norm by $\|\phi\|_{C^1}$ whenever $\|f\| \leq 1$ as the sup claimed □

(◯ the updated setting) With the result of the above box, fix $C_0 > 0$. For $\nu : \left\|h^1(\nu)\right\|_{BL} \leq C_0$. Set

$$\epsilon = \frac{\min\{\eta, \beta\}}{4\alpha M^2} \quad \text{M is unclear here}$$

Where the α is from Lemma 3.7, β, η are from □. If $\left\|h^1(\nu) - h^1(\mu)\right\|_{BL} < \epsilon$ then:

- $g_\nu \leq -\frac{\eta}{2}$ on K^+

- $\nabla g_\nu \cdot \vec{n}_\theta \leq -\frac{\beta}{2}$ on ∂K^+

Which is just a workout of the updated bounds given those of $g_\mu, \nabla g_\mu$ considering that by Lemma 3.7 ϵ-close measures in their projection satisfy an inequality.
($\triangledown$ Wgfs and time flows) for the function F consider a Wgf $(\mu_t)_t$ with starting measure μ_0 concentrated on $[-r_0, r_0] \times \Theta$, additionally, the BL distance with the measure μ presented in $\triangle$ is $\left\|h^1(\mu_0) - h^1(\mu)\right\|_{BL} < \epsilon$, where μ is from $\triangle$. We are basically setting the start point ϵ-close in projection to a stationary measure. We are in the context of Lemma 3.7 and can also state that $\left\|h^1(\mu_t)\right\|_{BL} < C_0$. We denote t_1 as the first, possibly divergent time, at which this last condition does not hold. Again, with the representation X of Lemma A.71:

- by construction of K^+ a flow $t \to X(t, (w_t, \theta_t))$ starting in $(w_0, \theta_0) \in \mathbb{R}_+ \times K^+$ stays inside for $t \leq t_1$
- by the homogeneity of $F'(\mu_t)$ which is 0-homogeneous (Prop. A.20) the velocity field component of w is lower bounded by $\frac{\eta}{2}$ so that:

$$w_t \geq w_0 + t\frac{\eta}{2}$$

- additionally, by the fact that no path enters $\mathbb{R}_- \times \Theta$ and that $F'(-1, \cdot) \geq F'(1, \cdot)$ we have:

$$w_t \geq w_0 + t\frac{\eta}{2} \quad in\ \mathbb{R}_- \times K^+$$

- for such interval of time with $0 \leq t < t_1$:

$$h^1(\mu_t)(K^+) \geq \left(t\frac{\eta}{2}\right) \cdot \mu_0(\mathbb{R}_+ \times K^+) + \min\left\{0, t\frac{\eta}{2} - r_0\right\} \cdot \mu_0(\mathbb{R}_- \times K^+) \tag{3.8}$$

For a starting measure that places positive weight on the target set (i.e. $\mu_0(\mathbb{R}_+ \times K^+) > 0$ the growth of the projection is **at least linear**.
($\diamondsuit$ cases distinction) again, we have to distinguish between two cases for Θ.
If $\Theta = K^+$, then we can choose $f \equiv 1$ in Def. A.57 for $\|\cdot\|_{BL}$ and get that the time t_1 is $< \infty$ in Equation 3.8.
Contarily, if $\Theta \subset K^+$, it is sufficient to check that the growth is still bounded at ∂K^+. For this purpose, denote a new sublevel set of g_μ with regular value in the interval $\widetilde{\eta} \in (-\eta, 0)$ and symbol $\widetilde{K}^+$, where $\widetilde{K}^+ \subset \Theta$ in this case.
We know g_μ is Lipschitz, so there exists $\Delta \in [0, 1]$ such that:

- the distance between K^+ and $\Theta \setminus \widetilde{K}^+$ is bounded above by Δ
- for $\epsilon > 0$ a small radius, it holds that either $t_1 < \frac{2r_0}{\widetilde{\eta}}$ or $\exists \widetilde{t} > t_0$ such that $h^1(\mu_t) \geq 0 \forall t \in [\widetilde{t}, t_1)$ on K^+

Now, a test function of Def. A.57 for the BL norm such as the distance to the set $\Theta \setminus \widetilde{K}$ clipped to 1 namely satisfying [CB18]:

$$\varphi : \mathbb{R}^d \to \mathbb{R},\ Lip(\varphi) \leq 1,\ \|\varphi\|_\infty \leq 1$$

is such that:

$$\left\|h^1(\mu_t)\right\|_{BL} \geq \Delta h^1(\mu_t)(K^+) \quad \forall t \in [\widetilde{t}, t_1)$$

Which also grows linearly in t, making $h^1(\mu_t)$ leave any $\|\cdot\|_{BL}$-ball in finite time, meaning that t_1 is necessarily finite. □

Lemma 3.9 (General property in projection of stationary points, nullity at convergence)**.** *Under Assumptions 3.4, consider a Wgf $(\mu_t)_t$ for F. Then:*

$$h^1(\mu_t) \overset{w}{\to} \nu \in \mathcal{M}_+(\Theta) \implies F'(\nu) = 0 \quad \nu\text{-a.e.}$$

Namely, if we converge in the projection, the derivative of the functional is stationary, and no direction of improvement is available.

Proof. For $u = (1, \theta) \in \Omega$ we can recover the velocity field with the 2-Lipschitz map (as per Prop. 2.26) through the map g_μ defined in Proposition 3.6, for the current μ_t. Namely:

$$(\mathtt{id} - \mathtt{proj}_{\partial V((1,\theta))})(\widetilde{v}_t(u)) = (\mathtt{id} - \mathtt{proj}_{\partial V((1,\theta))})\Big[\underbrace{g_{\mu_t}(\theta)}_{weights \in \mathbb{R}} \Big| \underbrace{\nabla g_{\mu_t}(\theta)}_{positions \in \mathbb{R}^{d-1}} \Big]$$

Thus, by Lemma 3.7 it holds that $v_{\mu_t} \rightrightarrows v_\nu$ in the space $\{1\} \times \Theta$. The arguments similar to those of Proposition 3.6, we briefly outline the key points of the proof:

- use uniform convergence of $g_{\mu_t} \rightrightarrows g_\nu$
- if $g_\nu(\theta_0) > 0$ for $\theta_0 \in \Theta$ build a set $\mathbb{R}_+ \times K$ with $\theta_0 \in \mathtt{int}\, K$ such that:
  - for some $t_0 > 0$ the dynamics of X_t never enters such a set for $t > t_0$
  - the velocity of the weights w, namely, $g_{\mu_t} \leq -\frac{g_\nu(\theta)}{2} \leq \frac{\eta}{2}$
- by μ_{t_0} concentrated on Q_{r_0} this implies that $\mu_{t_0}(\mathbb{R}_+^* \times K)$ vanishes in finite time and, in particular, $\nu(K) = 0$.
- thus, we have shown that $F'(\nu)$ is nonpositive ν-a.e.
- Also it can be deduced by Proposition 3.6, that $F'(\nu)$ is nonnegative ν-a.e. So $F'(\nu)$ vanishes ν -a.e.

□

We will use this property in the main result, to establish a contradiction, subject to the escaping criterion being satisfied throughout the dynamics. This last matter is solved in the following Subsection.

3.3 Stability

This Subsection is mostly based on *topological degree theory*, a tool that allows the treatment of cases in which V is non differentiable[22]. Allowing a differentiable structure for V, these facts follow by μ_t being the pushforward of μ_0 by a homeomorphism (last implication of the representation of Lemma A.71). These properties are crucial for the development of the main Theorems presented later.

Definition 3.10 (Topological degree)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be continuous, $A \subset \mathbb{R}^d$ bounded and open, $y \notin f(\partial A)$. The topological degree is denoted as $\mathtt{deg}(f, A, y)$. It satisfies:*

1. $\mathtt{deg}(f, A, y) \neq 0 \implies \exists x \in A \ : \ f(x) = y$
   $y \in A \implies \mathtt{deg}(id, A, y) = 1$

[22] recall that we are working with subgradients, Wasserstein subdifferentials and semiconvex V. One could also think of non-differentiable regularizers, as the simple modulus $|w|$ of the weights in the basic regularized Optimization setting.

2. A_1, A_2 *open*, $A_1 \cap A_2 = \emptyset$, $y \notin f(\overline{A} \setminus (A_1 \cup A_2)) \implies \texttt{deg}(f, A, y) = \texttt{deg}(f, A_1, y) + \texttt{deg}(f, A_2, y)$
3. $X : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ *continuous,* $y : [0,1] \to \mathbb{R}^d$ *a continuous curve s.t.* $y(t) \notin X_t(\partial A)\ \forall t \in [0,1] \implies \texttt{deg}(X_t, A, y_t) \equiv kost\ \forall t \in [0,1]$

Such Definition is a full characterization of a map from triplets (f, A, y) to $\mathbb{Z}$ [Bro83](Thm. 1-2). It serves as a signed **algebraic** counter of solutions $f(x) = y$ for $x \in A$. The term algebraic refers to the fact that a solution x counts as $+1$ if f preserves its orientation, and -1 otherwise, the sign is of no use for us, but ideally indicates the sign of the determinant of the derivative at the inverse map. In this context, the definition is just an adaptation of the classical one.

Lemma 3.11 (A property of continuous maps and measures)**.**

$$f : \mathbb{R}^d \to \mathbb{R}^d \text{ continuous}, \quad \mu \in \mathcal{M}_+(\mathbb{R}^d) \implies \texttt{spt}(f_{\#}\mu) = \overline{f(\texttt{spt}\,\mu)}$$

Namely, the support of a measure and the pushforward of the support by a continuous map are ***almost commutative****, with almost meaning that they commute if the map is closed (Def. A.29).*

Proof. **(△ first inclusion)** let $y \in f(\texttt{spt}\,\mu)$. Consider a neighborhood of y, denoted as $\mathcal{N}$. By (topological) continuity of f, we have that $f^{-1}(\mathcal{N})$ is a neighborhood of $\texttt{spt}\,\mu$. Being inside the support, we can say that it has positive measure, so that:

$$\mu(f^{-1}(\mathcal{N})) > 0 \quad \mu(f^{-1}(\mathcal{N})) = f_{\#}\mu(\mathcal{N})$$

Where the second statement is just the definition of pushforward (Def. A.52). From this we can state that $y \in \texttt{spt}(f_{\#}\mu)$ since we are stating that its neighborhood $\mathcal{N}$ has positive measure. By the arbitrariness of y:

$$f(\texttt{spt}\,\mu) \subset \texttt{spt}(f_{\#}\mu)$$

(□ second inclusion) let $y \in (\overline{f(\texttt{spt}\,\mu)})^c$, let $\mathcal{N}$ be a neighborhood of y such that $\mathcal{N} \cap \overline{f(\texttt{spt}\,\mu)} = \emptyset$, we can do so by continuity. Again, by (topological) continuity $f^{-1}(\mathcal{N}) \subset (\texttt{spt}\,\mu)^c$, which can be derived by basic reasoning on closures and inverses. Hence:

$$f_{\#}\mu(\mathcal{N}) = \mu(f^{-1}(\mathcal{N})) \leq \mu((\texttt{spt}\,\mu)^c) = 0$$

Where the first equality is again the definition of pushforward, the inequality follows from monotonicity of measures, and the nullity of the last equation is clear. In conclusion, we derive that $y \in (\texttt{spt}\,f_{\#}\mu)^c$ since it is part of a neighborhood with zero measure. By the arbitrariness of y:

$$(\overline{f(\texttt{spt}\,\mu)})^c \subset (\texttt{spt}\,f_{\#}\mu)^c \implies \texttt{spt}\,f_{\#}\mu \subset \overline{f(\texttt{spt}\,\mu)}$$

Which trivially holds by negating the complements.
Combining the results of △, □, we have:

$$A \subset B \quad B \subset \overline{A}$$

Hence, if B is closed, we are done, as we basically prove $B \supset \overline{A}$ for the right quantities. In our context $B = \texttt{spt}(f_{\#}\,\mu)$, which is closed by Definition of support (Def. A.47). □

Definition 3.12 (Separation of sets induced by a set)**.** *A set C in an ambient space Ω separates two sets B, A if any continuous path in Ω with endpoints in B, A has a point in C.*

We state the separation property in the partially 1-homogeneous setting considered. It will be shown that as in the abstract setting, it comes with nice properties for the problem we are considering.

A a closed set $K \subset [-r, r] \times \Theta$ satisfies the separation property if it separates $\{-r\} \times \Theta$ and $\{r\} \times \Theta$ for some $r > 0$.

Proposition 3.13 (Set separation, boxes, abstract)**.** *Let $\Theta \subset \mathbb{R}^d$ be the closure of a bounded, connected, open set. For some $T > 0$, let $X : [0, T] \times (\mathbb{R} \times \Theta) \to \mathbb{R} \times \Theta$ (recall the representation of Lemma A.71) be a continuous map such that:*

- $X(0, \cdot) = id$
- $X_t(\mathbb{R} \times \partial\Theta) \subset \mathbb{R} \times \partial\Theta \quad \forall t \in [0, T]$

If K satifies the separation property, then $X_t(K)$ does so $\forall t \in [0, T]$

Proof. Let $0 < \epsilon < \alpha < \beta$ be such that:

- $X_t(K) \subset (-\alpha - \epsilon, \alpha + \epsilon) \times \Theta \quad \forall t \in [0, T]$
- $[-\alpha, \alpha] \times \Theta \subset X_t((-\beta - \epsilon, \beta + \epsilon) \times \Theta) \quad \forall t \in [0, T]$

Intuitively, these conditions make sense since we allow arbitrary variables to include or be included in the sets (they are just a construction). Let A be the intersection of:

- $(-\beta, \beta) \times \Theta$
- the connected component (Def. A.5) of $\{\alpha\} \times \Theta$ in $(\mathbb{R} \times \Theta) \setminus K$, unique by the connectedness of Θ and $\mathbb{R} \times \Theta$ itself

A is bounded and open in $\mathbb{R} \times \mathbb{R}^{d-1}$. Since $\beta > \alpha$ and we are imposing an intersection of bounded sets.
Consider the function $\widetilde{X} : (t, x) \to (t, X_t(x))$ and the compact set $S = \widetilde{X}([0, T] \times \partial A)$ of $[0, T] \times (\mathbb{R} \times \Theta)$. S is compact since it is the result of a closed and bounded times the boundary of a bounded set that is bounded and closed (Prop. A.4).
For S^c, connected components are path-connected (Obs. A.9), so that the map:

$$(t, (w, \theta)) \to \texttt{deg}(X_t, A, (w, \theta))$$

is constant on each connected component of S^c (Def. 3.10#3). Diving further by the separation induced by K:

- $\texttt{deg}(X_t, A, (w, \theta)) = 1$ for $[0, T] \times (\{\alpha\} \times \Theta)$
- $\texttt{deg}(X_t, A, (w, \theta)) = 0$ for $[0, T] \times (\{-\alpha\} \times \Theta)$

For fixed $t \in [0, T]$ any path joining $\{-\alpha\} \times \Theta$ and $\{\alpha\} \times \Theta$ must intersect $X_t(\partial A)$. Paths that are contained in $[-\alpha, \alpha] \times \texttt{int}\,\Theta$ are of this kind.
Eventually, combining the assumption on X with the fact that:

$$\partial A \subset K \cup (\mathbb{R} \times \partial\Theta) \cup (\{\beta\} \times \Theta)$$

we have that:

$$X_t(\partial A) \cap ([-\alpha, \alpha] \times \texttt{int}\,\Theta) \subset X_t(K)$$

Which proves that X_t separates $\{-\alpha\} \times \Theta$ from $\{\alpha\} \times \Theta$ in the space $\mathbb{R} \times \texttt{int}\,\Theta$.
By $X_t(K)$ being closed, the last claim holds also in $\mathbb{R} \times \Theta$. □

The following is a Lemma that transfers the above result to the field of Wasserstein Gradient flows.

Lemma 3.14 (Stability of the separation property)**.** *Under Assumptions 3.4, let $(\mu_t)_t$ be a Wgf for F. If $\mathtt{spt}\,\mu_0$ satisfies the separation property, then $\mathtt{spt}\,\mu_t$ does $\forall t > 0$.*

Proof. **($\triangle$ construction)** Recall the representation of the velocities of a Wgf of Lemma A.71. It is continuous and satisfies the pushforward property that $\mu_t = (X_t)_{\#}\mu_0$.
$X_0 = id$ is clear, and X_t is coercive (Def. A.28) and closed by Proposition A.30.
We are left to check the two cases of Assumption 3.4#3.
(Θ bounded case) By Lemma 3.11, we can just check the assumptions of Proposition 3.13. In the context of $\triangle$, it suffices to check that:

$$X_t(\mathbb{R} \times \partial\Omega) \subset \mathbb{R} \times \partial\Omega$$

Which is guaranteed by the Neumann boundary conditions of Assumption 3.4#3-(b) [CB18].
($\Theta = \mathbb{R}^d$ case) the aim is recapturing the problem under the bounded case by means of the diffeomorphism:

$$\psi : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R} \times \mathtt{int}\, B(0,1) \quad \psi(w,\theta) = \begin{cases} \left(w, \dfrac{\theta}{|\theta|} \cdot \tanh|\theta|\right) & \theta \neq 0 \\ (w, 0) & \theta = 0 \end{cases}$$

Let $Y_t = \psi \circ X_t \circ \psi^{-1}$. Such a map moves the positions θ to the open unit ball centered at zero $\mathtt{int}\, B(0,1)$. It is also the flow of the velocity field:

$$\widetilde{v}_t(y) = d\psi_{\psi^{-1}(y)}(v_t \circ \psi^{-1}(y)) \quad on \quad \mathbb{R} \times \mathtt{int}\, B(0,1)$$

Which can be extendend by continuity to $\mathbb{R} \times \mathbb{S}^{d-2}$ by means of the Sard-regular limit function $g_\infty(\theta)$ of Assumption 3.4#3-(a) as[23]:

$$(g_\infty(\theta) \cdot \mathtt{sgn}\, w, 0) \quad on \quad \mathbb{R} \times \mathbb{S}^{d-2}$$

The velocity flow Y_t satisfies the requirements of Proposition 3.13, so the claim holds for:

$$\psi(\mathtt{spt}\,\mu_t) = \psi \circ X_t(\mathtt{spt}\,\mu_0)$$

By the fact that ψ is a diffeomorphism, it keeps topological properties such as connectedness invariant, making the claim true (i.e., we can remove ψ with its inverse). □

Having found a condition on the support of a measure to be satisfied for any time point of the flow, we could simply design a support that is both **stable** in the sense of this Subsection and **able to escape** in the sense of the previous one. Apart from further technical conditions, this is the spirit of the results we will present next.

[23] the convention is $\mathtt{sgn}\, 0 = 0$

3.4 Main result of the Section, and a generalization

The first fact outlined, apart from being instrumental for the main Theorem, is also individually interesting, as it draws a connection between convergence of a Wgf and asymptotic properties of Gradient flows.
The Lemma could have been stated before, as we declare the previous set of Assumptions. For clarity, it is reported here, but it is more general. We will use it to establish the convergence of the particles to a global optima.

Lemma 3.15 (Limit order is not important)**.** *Under Assumptions 1.10, assume that there is:*

- $(\mu_t)_t$ *a Wgf for* F *such that* μ_0 *is concentrated on* Q_{r_0} *and* $F(\mu_t) \overset{t\to\infty}{\to} F^*$
- $(\mu_{0,m})_m$ *concentrated on* Q_{r_0} *and such that* $\mu_{m,0} \underset{m\to\infty}{\overset{W_2}{\to}} \mu_0$

Then, the limits can be exchanged in the process to find the optimal risk configuration:

$$F^* = \lim_{t\to\infty} \lim_{m\to\infty} F(\mu_{m,t}) = \lim_{m\to\infty} \lim_{t\to\infty} F(\mu_{m,t})$$

Proof. **(*m* first)** By Theorem 2.28, we have $\mu_{m,0} \to \mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ the unique Wgf starting from μ_0 of F (Prop. 2.26). By Lemma A.69, we have continuity of F, so that $F(\mu_m) \to F(\mu)$ and the limit makes sense. Imposing the subdifferential equation, we have necessarily that the dynamics will stop at least at a local minima F.
(*t* first) Recall that again, by the construction of the dynamics, $F(\mu_t)$ is monotonic along Wgfs, whatever the m. We have the following:

$$\forall \epsilon > 0 \quad \exists t_0 \in \mathbb{R}_+ \mid F(\mu_{t_0}) < F^* + \frac{\epsilon}{2} \tag{3.16}$$

Now, by Theorem 2.28:

$$\exists m_0 \in \mathbb{N} \mid \forall m \geq m_0 \quad F(\mu_{m,t_0}) < F(\mu_{t_0}) + \frac{\epsilon}{2} \tag{3.17}$$

because we tend to that μ_{t_0} at divergent m.
Eventually, combining the two, by the map $t \to F(\mu_{m,t})$ being decreasing and lower bounded independently of m (i.e. $\forall m$) we have:

$$\begin{aligned} \forall m \geq m_0 \quad \lim_{t\to\infty} F(\mu_{m,t}) &\leq F(\mu_{m,t_0}) \\ &< F(\mu_{t_0}) + \frac{\epsilon}{2} && \text{Eqn. 3.17} \\ &< F^* + \frac{\epsilon}{2} + \frac{\epsilon}{2} \\ &= F^* + \epsilon \end{aligned}$$

Notice that this is the Definition of limit for m of the quantity $\lim_{t\to\infty} F(\mu_{m,t})$. □

Lastly, the tools derived allow to state a general fact about convergence to a global minima, followed by a second statement, which is a Corollary, but presented as the main result of the Section by the authors for clarity.

Theorem 3.18 (Global minimization of projection, general)**.** *Under Assumptions 3.4, for some* $r_0 > 0$ *let:*

- *(concentration)* $\mathtt{spt}\,\mu_0 \subset [-r_0, r_0] \times \Theta$.
- *(separation)* $(\mu_t)_t$ *be a Wgf of* F *such that* $\mathtt{spt}\,\mu_0$ *separates* $\{-r_0\} \times \Theta$ *and* $\{r_0\} \times \Theta$

Then:

$$h^1(\mu_t) \overset{w}{\to} \nu \implies F(\mu_t) \overset{t\to\infty}{\to} F^* = \min_{\mathcal{M}_+(\Omega)} F$$
$$\lim_{t\to\infty} F(\mu_t) = F^*$$

Proof. Let $h^1(\mu_t) \overset{w}{\to} \nu \in \mathcal{M}(\Theta)$. A nice perspective is interpreting such limit as a measure on $\{1\} \times \Theta$, where the one does not matter in terms of computations. By Lemma 3.9 we have:

$$h^1(\mu_t) \overset{w}{\to} \nu \implies F'(\nu) = 0 \quad \nu\text{-a.e.}$$

(△ strategy) we proceed by contradiction, assuming that ν is not a minimizer of F over $\mathcal{M}_+(\Omega)$, interpreting it as a measure in the enlarged space, which by the characterization of Proposition 3.3 means that we assume $F(\nu)$ to be **not nonnegative**.
(□ highlights of the proof) the authors make use of results from the 2-homogeneous case, which are not covered in this document. For this reason, we reroute the reader to the original publication [CB18]. We also need an application of [CB18](Lem. C.18). □

Theorem 3.19 (Global minimization, use case)**.** *Under Assumptions 3.4 add that* $(\mu_t)_t$ *is a Wgf of* F *which for some* $r_0 > 0$ *satisfies*

- *(concentration)* $\mathtt{spt}\,\mu_0 \subset [-r_0, r_0] \times \Theta$.
- *(separation)* $(\mu_t)_t$ *be a Wgf of* F *such that* $\mathtt{spt}\,\mu_0$ *separates* $\{-r_0\} \times \Theta$ *and* $\{r_0\} \times \Theta$

Then:

1. $(\mu_t)_t \overset{W_2}{\to} \mu_\infty \implies F(\mu_t) \overset{t\to\infty}{\to} F^* = \arg\min_{\mathcal{M}_+(\Omega)} F$
2. *for a given (parameter) classical Gradient flow* $(\boldsymbol{u}_m(t))_{m\in\mathbb{N}, t\in\mathbb{R}_+}$ *which is initialized at its Wgf in* $[-r_0, r_0] \times \Theta$*:*

$$\mu_{m,0} \overset{W_2}{\underset{m\to\infty}{\to}} \mu_0 \implies \lim_{t,m\to\infty} F(\mu_{m,t}) = \min_{\mu\in\mathcal{M}_+(\Omega)} F(\mu)$$

Proof. **(Claim #1)** Direct implication of Thm. 3.18, after noting that the convergence in W_2 means that the projected measures converge weakly (Prop. A.66) and the conditions are all satisfied. Also, notice that the hypothesis of the general case is thus stronger.
(Claim #2) we showed the link gradient flow-Wgf in Proposition 2.11. Such Wgf is unique and the correspondence is non ambiguous (Prop. 2.26). Convergence of the particles is granted by Theorem 2.28. We can exchange the limits by Lemma 3.15, thus writing t, m in the subscript.
All of these facts, together with Claim #1, allow for the final statement.
Some comments are reported in the description of Figure 4. □

Observation 3.20 (Differences use case VS general case)**.** *The hypothesis, as pointed out in the proof of Claim #1 above, is weaker in the general case, as only a projection of the measures needs to converge (weakly). The latter is presented since the former might be hard to check. The former is presented since the latter might be hard to hold. This is also stressed in the final comments as it is important, where more discussion on convergence in* W_2 *is carried out.*

Observation 3.21 (On the limit exchange)**.** *Being able to exchange limits is fundamental. The divergent indices m, t do not influence each other in the convergence to F^*.*

Observation 3.22 (On the Assumptions)**.** *The authors stress the fact that while the structural homogeneity and initialization assumptions are instrumental, Sard-type regularity is purely technical [CB18]. Nevertheless, it was not possible to remove it due to its hardness in principle (see some more discussion in Subsection A.2.2). A mentioned artificial counter example of this regularity condition is the Cantor function [Whi35].*

Observation 3.23 (Final comments)**.** *Below are some* ***important observations*** *made by the authors [CB18].*

- *Theorem 3.19 assumes that the Wasserstein gradient flow converges in W_2. The authors stress that this is not always guaranteed and, in most of the cases, requires:*
  1. *compactness of trajectories*
  2. *a Łojasiewicz inequality*[24]
- *compactness in W_2 is a strong assumption, the topology of convergence was relaxed to the weak setting in Theorem 3.18*
- *even if compactness holds, there is no guarantee of convergence of gradient flows, a counter-example is found in [AMA05]. It is acknowledged that no general result exists for non-geodesically convex R, i.e. R interpreted with the Wasserstein distance instead of total variation. Some improvements in this directions are explored in [BSR15; HM19].*
- *Proposition 3.6 is qualitative, it does not provide bounds of convergence, but only states that a sufficient condition to escape local minima is having a particle belonging to the 0-sublevel set of $F'(\mu)$. One could think of ways other than using the many-particle limit to approach the measure, or calculating the size of such sublevel sets, to extract a measure of complexity of reaching global minima via particle gradient descent. In [MMM19] a quantification is given.*

[24] Informally: a bound on the extent to which a function is flat around its critical points. The original publication is *S. Łojasiewicz, "Sur les trajectoires du gradient d'une fonction analytique", Seminari di Geometria, Bologna (1982/83), Universita' degli Studi di Bologna, Bologna (1984), pp. 115–117*

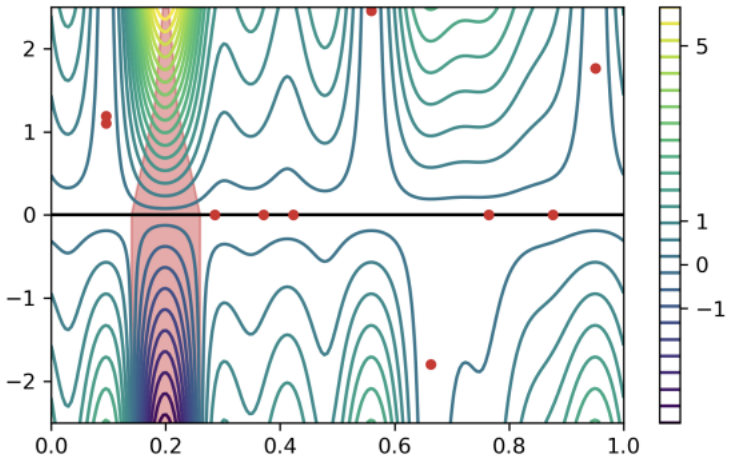


Figure 4: Homogeneous landscape strucure, [CB18]

Again, plotting a view of $F'(\mu)$, we are now in the position to comment further. The red dots identify the support of a measure ν which is a non optimal stationary point. It is non optimal since $F'(\nu)$ is negative at some points, meaning that it does not satisfy Proposition 3.3. We could then imagine a Wgf (μ_t) which gets ϵ-close in BL-norm to it. To escape getting trapped, it should give positive weight to the red region. The part below the horizontal line has $F'(\nu)$ negative. There Proposition 3.6 can be used. On the contrary, the part above where $F'(\nu)$ is positive is required as well but needs more technical conditions. These are largely discussed by the authors [CB18](Lem. C.18). Theorem 3.18 uses both the technical result and the fact that the separation property satisfies the criteria to escape throughout the dynamics and conclude that a global minima will be reached.

4 One layer Neural Networks

Having outlined the general framework, we now apply them to a very important setting. The general properties of the loss function are inspected at the functional level. We then connect the discussion of Subsection 1.1, recalling that we lifted the problem to a more general class of problems in Subsection 1.4. The focus is on sigmoid neural networks, for which sufficient conditions are given for the theoretical results to be applied. Lastly, a Theorem resumes all of the work done in order, with some brief practical considerations on the implementation and the empirical results.

4.1 Loss

We first give necessary conditions on the loss structure to be in line with our results. The Hilbert space we consider is $\mathcal{F} = L^2(\rho)$ for $\rho : \mathcal{X} \to \mathbb{R}$ a probability measure with $X \subset \mathbb{R}^d$. A functional loss takes the form:

$$R(f) = \int r(x, f(x))d\rho(x) \quad r : \mathcal{X} \times \mathbb{R} \to \mathbb{R}_+$$

Where the probability distribution is not to be seen in the sense of dependent and independent variables yet. We are just inspecting the properties of the functional with respect to a measure.

Lemma 4.1 (Sufficient conditions for functional loss)**.** *The following hold:*

1. *r convex in the second variable $\implies$ R convex*
2. *$\exists \partial_2 r$ Lipschitz uniformly in the first variable $\implies$ $\exists dR$ and dR is Lipschitz*
3. *$\partial_2 r \leq C_1 r + C_2$ $C_1, C_2 > 0$ $\implies$ dR bounded on sublevel sets*

Proof. **(Claim #1)** We have by hypothesis:

$$\forall x \in \mathcal{X}, \quad \forall f, h \in \mathcal{L}^2(\rho), \forall \alpha \in [0,1] \quad r(x, \alpha f(x) + (1-\alpha)h(x)) \leq \alpha r(x, f(x)) + (1-\alpha) r(x, h(x))$$

so that trivially $\forall f, h \in L^2(\rho), \forall \alpha \in [0,1]$:

$$\begin{aligned} R(\alpha f + (1-\alpha)h) &= \int r(x, \alpha f(x) + (1-\alpha)h(x))d\rho(x) \\ &\leq \int \alpha r(x, f(x)) + (1-\alpha) r(x, h(x))d\rho(x) \qquad \text{by } \leq \forall x \text{ convexity of } r \\ &= \alpha \int r(x, f(x))d\rho(x) + (1-\alpha) \int r(x, h(x))d\rho(x) \\ &= \alpha R(f) + (1-\alpha) R(h) \end{aligned}$$

(Claim #2)($\triangle$ differentiability) To show differentiability, we aim to find a linear operator with a Taylor expansion for which we have a $\mathcal{O}(\|\|_{L^2(\rho)})$ bound, being that we are in the Hilbert space $L^2(\rho)$. Knowing that r is differentiable we can expand it inside R. By the uniformity in the first variable assumption, we avoid writing $f(x), h(x)$ every time.

$$\begin{aligned} \forall f, h \in \mathcal{F} \quad R(f+h) &= \int r(x, f+h)d\rho \\ &= \int r(x, f) + r'(x, f) \cdot h + \mathcal{O}(|h|)d\rho \end{aligned}$$

Which plugged into the outer expansion for R suggests that a good candidate is:

$$dR_f : \mathcal{F} \to \mathbb{R} \quad h \to \int r'(x, f(x))h(x)d\rho(x) \quad \forall h \in \mathcal{F}$$

Indeed for L the uniform Lipschitz constant of r:

$$\begin{aligned}
\left|R(f+h) - R(f) - dR_f\right| &= \left|R(f) + dR_f + \int \mathfrak{o}(|h|^2)d\rho(x) - R(f) - dR_f\right| \\
&= \int \mathfrak{o}(|h|^2)d\rho(x) && \mathfrak{o}(|h|^2) = \frac{1}{2}r''(f)\mathfrak{o}(|h|) \\
&\leq \int \frac{1}{2}L|h|^2 d\rho && \text{Prop. A.23} \\
&= \frac{L}{2}\|h\| \\
&= \mathfrak{o}(\|h\|)
\end{aligned}$$

(□ Lipschitzness) we aim to show:

$$\exists C \in \mathbb{R} \quad : \quad \|dR_f - dR_g\| \leq C\|f - g\| \quad \forall f, g \in \mathcal{F}$$

With a similar application of Proposition A.23:

$$\begin{aligned}
\|dR_f - dR_g\| &= \left\|\int r'(x, f)d\rho - \int r'(x, g)d\rho\right\| \\
&= \left\|\int r'(x, f) - r'(x, g)d\rho\right\| \\
&\leq C\|f - g\|
\end{aligned}$$

(Claim #3) We have:

$$\begin{aligned}
\|dR_f\|^2 &= \int |\partial_2 r(x, f(x))|^2 d\rho(x) \\
&\leq \int C_1 r(x, f(x)) + C_2 d\rho(x) \\
&= C_1 \underbrace{\int r(x, f(x))d\rho}_{=R(f)} + C_2 \underbrace{\int d\rho(x)}_{=1} \\
&= C_1 R(f) + C_2 \\
\implies \|dR_f\| &= \sqrt{C_1 R(f) + C_2}
\end{aligned}$$

So that in the sublevel sets dR is bounded. □

4.2 A Machine Learning Application

Before getting to Neural Networks, we formulate the task in the classical way.

4.3 From an Optimization Problem to a Learning Problem

Consider a distribution of labels and features $\rho \in \mathcal{P}(\mathbb{R}^{d-2} \times \mathbb{R})$ where $\rho_x \in \mathcal{P}(\mathbb{R}^{d-2})$ is the marginal of the features. We can treat such problem via a conditional probability expression

[AGS05](Thm. 5.3.1):

$$\rho(dx \otimes dy) = \rho(dy|x)\rho_x(dx) \quad (\rho(\cdot|x))_{x\in\mathcal{X}} = \{p.m.\ on\ \mathcal{Y}\}$$

This disintegration is required as in principle we do not have access to the joint law over the space $\mathcal{X}\times\mathcal{Y}$, and hypothesize that there is a hierarchical relation such that x somehow influences y through something we would like to estimate up to reasonable precision. As loss, we use the expected risk:

$$R : L^2(\rho) \to \mathbb{R} \quad R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y) d\rho(x, y)$$

In the previous Subsection, we analyzed such functional loss **in terms of** x only. To reconcile these two perspectives and exploit the conclusions of Lemma 4.1, we choose:

- $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$ a convex loss function, either:
  - square loss
  $$\ell(f(x), y) = (f(x) - y)^2$$
  - logistic loss
  $$\ell(f(x), y) = [f(x)]^y[1 - f(x)]^{1-y}$$
- as **separable** Hilbert space $\mathcal{F} = L^2(\rho_x)$
- as r function[25]:
  $$r(x, p) = \int_{\mathbb{R}} \ell(p, y)\rho(dy|x) \quad p : \mathcal{X} \to \mathbb{R}$$
  where p stands for "predictor" and we are **integrating out** $y \in \mathcal{Y}$.

This reparametrization splits the integrals of the functional loss:

$$R : L^2(\rho_x) \to \mathbb{R} \quad R(f) = \int_{\mathcal{X}}\int_{\mathbb{R}} \ell(f(x), y)\rho(dy|x)\rho_x(dx)$$

For ℓ as stated, the function r coupled with the optional $\widetilde{V} = 1$ satisfies the requirements for Lemma 4.1 to apply, namely:

- r is convex in the second variable
- r is differentiable in the second variable
- $\partial_2 r$ is Lipschitz uniformly in the first variable
- $|\partial_2 r|^2 \le C_1 r + C_2 \quad C_1, C_2 > 0$

Now, we have features in $\mathbb{R}^{d-2}$ but wish to add a bias term, so the positions θ will be in $\mathbb{R}^{d-1} = \Theta$. To simplify calculations, we denote:

$$z = (x, 1) \in \mathbb{R}^{d-1} \sim \rho_z = \rho_x \times id$$

which is just an extension.

4.4 Sigmoid Neural Networks

Neural Networks are a context in which our theory works well. Two very common choices for nonlinear activation functions are sigmoid and `ReLu` [Hay99; GBC16]. The specific results

[25] notice that the order of inputs is slightly misleading

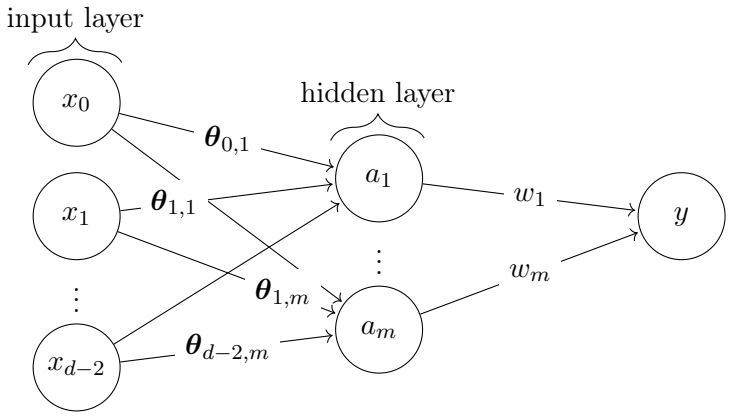


Figure 5: The diagram shows an intuitive representation of a two-layer neural network. The inputs are $d-2$ dimensional, with an added bias. They are passed to activations a_i of the form $a_i(x) = \sigma(\boldsymbol{\theta}(\cdot, i)^T x)$. The final output is then determined by a weighted sum of activations.

reported here are sufficient to examine the former architecture. The latter is also analyzed by the authors extensively, together with *sparse spikes deconvolution* [CB18]. The main difference between the three cases is the domain Θ and the type of homogeneity they generate.

We focus on Neural Networks with one hidden layer. Simplifying the dependence on $u = (w, \theta)$ which is implicitly present:

$$h(x) = \mathbf{w}^T \sigma(\boldsymbol{\theta}^T x)) = \sum_{i=1}^{m} w_i \cdot \sigma(\boldsymbol{\theta}(\cdot, i)^T x) \tag{4.2}$$

Where m is the number of hidden neurons, w_i is the outgoing weight of the i^{th} neuron, $\boldsymbol{\theta}(\cdot, i)$ are the ingoing weights of the i^{th} neuron.
The single hidden layer structure allows for the formulation of Eqn. 4.2 which is quite peculiar since:

- there is total independence of contributions for the hidden layers to the output. More than one hidden layer would have led to interactions.
- it is a linear combination of hidden neurons

The functions of the first original optimization problem, (expressed in terms of the inputs x, with specified positions θ) are then:

$$\phi(\theta) : \mathcal{X} \to \mathbb{R} \quad x \to \sigma(z \cdot \theta) = \sigma\Big(\sum_{i=1}^{d-2} \theta_i x_i + \underbrace{\theta_{d-1}}_{bias} \Big) \quad \widetilde{V} = 1$$

Where σ is a **sigmoid**[26].

With this in mind, we show that the settings of Assumption 3.4 are verified. Notice that if we *see* ϕ in two different interplaying ways:

- as a simple application of the $\theta \cdot z$ product for the realization of the functional loss R

$$\phi(\theta) : \mathcal{X} \to \mathbb{R}$$

[26] the classical sigmoid $\sigma(s) = \dfrac{1}{1 + e^{-s}}$

- as a function reproducing functions[27]. **To be tuned correctly in terms of** θ to be in line with the assumptions and make R behave as wanted:

$$\phi : \Theta \to \mathcal{F} \quad \Theta = \mathbb{R}^{d-1}$$

Lemma 4.3 (ϕ for Sigmoid NN)**.** *Order* 4 *finite moments of the features distribution* ρ_x*:*

$$x : \mathbb{E}[|x|^4] < \infty \iff \int |x|^4 d\rho_x(x) < \infty$$

imply that:

1. $\phi : \Theta \to \mathcal{F}$ *is differentiable*
2. *the differential of* ϕ *is:*

$$d\phi_\theta(h) : x \to (h \cdot z)\sigma'(z \cdot \theta) \quad z = (x, 1) \in \Theta$$

3. $d\phi_\theta$ *is Lipschitz*

Proof. As a side note, recognize that:

- finite 4^{th} moment for ρ_x means that also ρ_z has finite 4^{th} moment.
- by the fact that $z = (x, 1)$ we can safely say $\mathbb{E}[|x|^n] = \mathbb{E}[|z|^n] \quad \forall n \leq 4$
- the sigmoid function has Lipschitz derivative. Denote as L its constant.
- the sigmoid function has a finite sup norm $\exists \, \|\sigma'\|_\infty$

(Claim #1, #2) we prove together that there is a linearization bounded in the norm. Firstly, we precompute the Taylor expansion of ϕ in the weight direction h:

$$\phi(\theta + h) = \sigma(z \cdot (\theta + h)) = \sigma(z \cdot \theta + z \cdot h) = \sigma(z \cdot \theta) + \sigma'(z \cdot \theta)(z \cdot h) + \frac{1}{2}\sigma''(z \cdot h)(z \cdot h)^2 + \mathcal{O}(|h|^2)$$

To ease out the process, check directly the differential form claimed:

$$\begin{aligned}
\Delta(h)^2 &:= \|\phi(\theta + h) - \phi(\theta) - d\phi_\theta(h)\|^2 && \text{square norm} \\
&= \int_{\mathcal{X}} \left| \mathcal{O}\left(\frac{1}{2}\sigma''(z \cdot \theta)(z \cdot h)^2\right) \right|^2 d\rho_z(z) \\
&\leq \frac{L^2}{4} \int_{\mathcal{X}} \left|(h \cdot z)\right|^4 d\rho_z(z) && \text{Prop. A.23} \\
\implies \Delta(h) &\leq \frac{L\sqrt{\mathbb{E}[|z^4|]}}{2}|h|^2 = \mathcal{O}(|h|)
\end{aligned}$$

[27] wordy

(Claim #3) we have a differential, now we want a uniform Lipschitz bound.
($\triangle$ boundedness)observe that:

$$\begin{aligned}
\|d\phi_\theta\|^2 &\leq \sup_{x\in\mathcal{X},\|x\|\leq 1,z=(x,1)} \frac{\|(h\cdot z)\sigma'(z\cdot\theta)\|^2}{|h|^2} \\
&= \sup_{x\in\mathcal{X},\|x\|\leq 1,z=(x,1)} \frac{\int_{\mathcal{X}} |(h\cdot z)\sigma'(z\cdot\theta)|^2 d\rho_z(z)}{|h|^2} && \text{where } \exists \left\|\sigma'\right\|_\infty \\
&= \left\|\sigma'\right\|_\infty^2 \sup_{x\in\mathcal{X},\|x\|\leq 1,z=(x,1)} \frac{\int_{\mathcal{X}} |(h\cdot z)|^2 d\rho_z(z)}{|h|^2} \\
&= \left\|\sigma'\right\|_\infty^2 \sup_{x\in\mathcal{X},\|x\|\leq 1,z=(x,1)} \int_{\mathcal{X}} |z|^2 d\rho_z(z) \\
&= \left\|\sigma'\right\|_\infty^2 \mathbb{E}[|z|^2] \\
\implies \|d\phi_\theta\| &\leq \left\|\sigma'\right\|_\infty \sqrt{\mathbb{E}[|z|^2]} < \infty
\end{aligned}$$

($\square$ Lipschitzness) we work on the displacement for $\theta, \widetilde{\theta} \in \mathbb{R}^{d-1}$:

$$\begin{aligned}
\left\|d\phi_\theta - d\phi_{\widetilde{\theta}}\right\|^2 &\leq \sup_{x\in\mathcal{X},\|x\|\leq 1,z=(x,1)} \frac{\left\|(h\cdot z)\sigma'(z\cdot\theta) - (h\cdot z)\sigma'(z\cdot\widetilde{\theta})\right\|^2}{|h|^2} \\
&= \sup_{x\in\mathcal{X},\|x\|\leq 1,z=(x,1)} \frac{\left\|(h\cdot z)\overbrace{(\sigma'(z\cdot\theta) - \sigma'(z\cdot\widetilde{\theta}))}^{\leq L|z(\theta-\widetilde{\theta})|}\right\|^2}{|h|^2} && \sigma' \text{ is L-Lipschitz} \\
&= L^2|\theta-\widetilde{\theta}|^2 \left\|z^2\right\|^2 && \text{cancel } h \\
&\leq L^2|\theta-\widetilde{\theta}|^2 \mathbb{E}[|z|^4] \\
\implies \left\|d\phi_\theta - d\phi_{\widetilde{\theta}}\right\| &\leq L\sqrt{\mathbb{E}[|z|^4]}|\theta-\widetilde{\theta}|
\end{aligned}$$

which means that $d\phi$ is Lipschitz in Θ. $\square$

The next Lemma is used to inspect regularity properties of our construction.

Lemma 4.4 (Sigmoid Sard-type regularity)**.** *In our functional space $\mathcal{F}$ it holds that the function:*

$$\forall f \in \mathcal{F} \quad \theta \to \langle f, \phi(\theta)\rangle = \int_{\mathcal{X}} f(x)\sigma((x,1)\cdot\theta) d\rho_x(x)$$

Has regular values (Def. A.41) that are dense in the range of the map $\theta \to \langle f, \phi(\theta)\rangle$ if it has bounded moments up to order $2d-2$.

Proof. **($\triangle$ base case)** if the function is constantly $c \in \mathbb{R}$ we have that:

- the range is $\{c\}$
- any point of the domain is a critical point
- regular values are dense in the range since we can choose any point which is not $\{c\}$ and get the empty inverse image, which is by convention a regular value.

(□ general case) if the function is not identically constant, the range is a subset of the real line $\mathbb{R}$. We report below a calculation which is useful but should not break the course of the reasoning. Since we want to use Morse-Sard (Lem. A.45), we need the map to be smooth in $\mathbb{R}^d$, which means $d-1$ differentiable.
(◯ moments and differentiability connection) denote the map $\theta \to \langle f, \phi(\theta)\rangle$ as $g_f(\theta)$. Then, differentiability can be checked:

$$\begin{aligned}\|g_f(\theta+h)-g_f(\theta)\|^2 &= \left\|\int_{\mathcal{X}} f(x)\sigma((\theta+h)\cdot z)d\rho_x(z) - \int_{\mathcal{X}} f(x)\sigma(\theta\cdot z)d\rho_x(x)\right\|^2 \\ &\leq \int_{\mathcal{X}}\left|f(x)[\sigma((\theta+h)\cdot z)-\sigma(\theta\cdot z)]\right|^2 d\rho_x(x) \\ &= \int_{\mathcal{X}}\left|f(x)(\sigma'(\theta\cdot z)(z\cdot h)+\mathfrak{o}(|h|))\right|^2 d\rho_x(x) \qquad \text{Taylor } \sigma\end{aligned}$$

So, if we chose as differential $f(x)\sigma'(\theta\cdot z)(z\cdot h)$ we could have had a $\mathfrak{o}(|h|)$ (i.e. apply again Prop. A.23) term where the norm is in $L^2(\rho_x)$ and thus requires bounded second moments for the first derivative to exist.
It can be shown that this procedure iterated becomes:

$$\mathbb{E}[|x|^{2d-2}] < \infty \implies \exists d^{d-1} g_f(\theta)$$

(▽ back to regular values) with this fact in mind, we use ◯ to state that the function will be smooth ($d-1$ differentiable) whenever moments are bounded up to order $2d-2$. Recalling the notions of Subsection A.2.2, such smoothness ensures by Morse-Sard (Lem. A.45) that the critical points have measure zero in their realization, so that the regular values are dense in the range of the function. □

We have now proved the *settings* and #1, #2 of Assumption 3.4. To check the regularity at the boundary of #3-(a)(b), the discussion is carried out without much detail in the Appendix B.4. In the context of the our final result, it will be taken for granted.
On top of this, we reparametrize the equation to highlight the dependence on single hidden neurons $w_1, \dots, w_m$ via a family of functions:

$$h = \frac{1}{m}\sum_{i=1}^{m}\Phi(u_i) \quad \Phi(u_i)(x) = m w_i \cdot \sigma(\boldsymbol{\theta}(\cdot,i)^T x) \quad u_i = (w_i, \boldsymbol{\theta}(\cdot,i)) \in \mathbb{R}^d$$

So, h is now an average of the predictions of each of the hidden neurons. This formulation allows to highlight nice properties, especially at the divergent limit of m. Important results in this setting suggest that an overparametrized model ($m > nd$) with infinite hidden neurons ($m \to \infty$) is approximately equal to evaluating the integral of a dirac measure on the neurons [Bar93](with further studies in [KS01; Ben+05; Ros+07]):

$$h \overset{m\to\infty}{\approx} \int_{\mathcal{U}} \Phi(u)d\mu(u) \quad \mu(w) = \frac{1}{m}\sum_{i=1}^{m}\delta_{u_i}$$

Which is **linear** in $d\mu$. Namely, in the limit we replace the empirical measure with the integral of the empirical measure of neurons. This is exactly the 1-homogeneous lifting procedure that we performed in Section 1.4.

$$\Phi(w,\theta) = w\phi(\theta) \quad x \to w\sigma\left(\sum_{i=1}^{d-2}\theta_i x_i + \theta_{d-1}\right) \quad V(w,\theta) = |w|\widetilde{V} = |w|$$

where the regularization term is a simple weight penalization. Namely, weight of the m particles times a sigmoid of the positions plus a bias term. The construction is partially 1-homogeneous in the sense of Definition A.19.

The final result is rather trivial, but is reported here for the sake of summarizing what we found.

Theorem 4.5 (Sigmoid Neural Network global minima convergence)**.** *The information is a data sample, consisting of a collection of tuples $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X} \subset \mathbb{R}^{d-2}$ and $y_i \in \mathcal{Y} \subset \mathbb{R}$ with unknown distribution $\rho(x, y)$. A structural assumption suggests finding the best model that expresses y as a function of x. The choice of the function is a sigmoid neural network.*
To estimate it, we are given a general minimization problem of the form of Equation 1.6, namely:

$$\mu^* = \underset{\mu \in \mathcal{M}(\Theta)}{\arg\min} J(\mu) \quad J(\mu) \coloneqq R\bigg(\int \phi d\mu\bigg) + G(\mu)$$

Where:

- $\Theta = \mathbb{R}^{d-1}$
- $\phi(\theta) : \mathcal{X} \to \mathbb{R} \quad x \to \sigma(\sum_{i=1}^{d-2} x_i\theta_i + \theta_{d-1})$
- *R is the risk of the quadratic or logistic loss, with the sufficient conditions of Lem. 4.1*
- *G is the total variation norm $G(\mu) = |\mu|(\Theta)$*

Basically, our objective is to find the best possible allocation of positions in the space $\Theta = \mathbb{R}^{d-1}$ of a measure so that the error with respect to the real data distribution is minimized.
Assume:

- *$\rho_x \in \mathcal{P}(\mathbb{R}^{d-2})$ has moments that are finite up to $\max\{4, 2d-2\}$*
- *$\mathtt{spt}\,\mu_0 = \{0\} \times \Theta$*
- *the condition of Assumption 3.4#3-(a) is verified*

Then a Wgf for the problem $(\mu_t)_{t \in \mathbb{R}_+}$ is such that:

$$\mu_t \overset{W_2}{\to} \mu_\infty \implies \mu_\infty = \arg\min F$$

Where we can easily recover the measure $\nu \in \mathcal{M}(\Theta)$ corresponding to $\mu \in \mathcal{P}(\Omega)$ Additionally, such Wgf can be obtained by performing particle gradient descent on a discretized version of our functional optimization problem that can be performed in practice.

Proof. We go by the order of exposition of the various discussions we had, except that we start from the results of this Section, which guarantees that the assumptions are met. The results allow to:

- recollect facts that prove how the setting of sigmoid neural networks is in line with the results of the previous Sections (§Sec. 4)
- lift the optimization problem to a higher dimensional space, including the weights w inside the parameters of optimization, and work with measures on $\mathcal{P}_2(\Omega) \quad \Omega = \mathbb{R}^d$ (§Sec. 1, Prop. 1.14, projection map Eqn. 1.12)
- formulate the discretized minimization problem and examine the gradient flow of its parameters (§Sec. 2, Prop. 2.3)

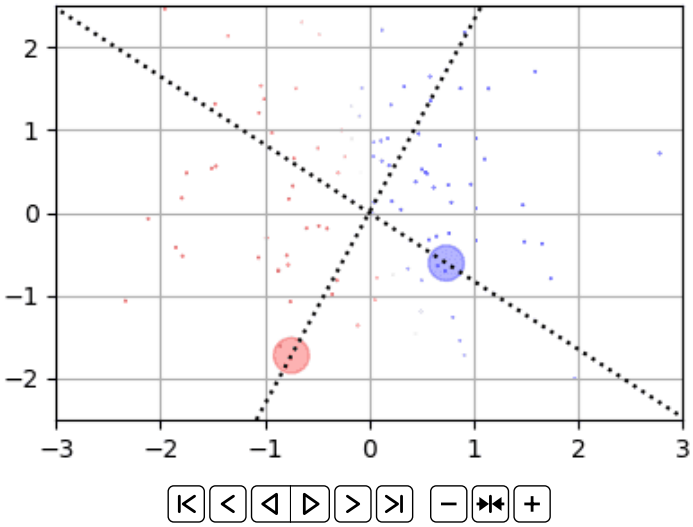


Figure 6: Animated Sigmoid NN particle dynamics. Source `https://lchizat.github.io/PGF.html`

Weights $w_i(t)$ are represented by the size of the particle. Colors are red for positive weights and blue for negative weights. The generators of the labels (i.e. the ground truth) are the big disks.

- reconcile it with the notion of Wasserstein gradient flow at the particle limit $m \to \infty$ (§Sec. 2, Prop. 2.11)
- move on to the actual convergence analysis, where up to the initialization condition, it is guaranteed that the local minima will be escaped (§Sec. 3, Prop. 3.6, Subsec. 3.3)
- given guaranteed convergence at $m \to \infty$, simply choose a large enough m^* so that the behavior is met (§Sec. 3, especially Thm. 3.19, Lem. 3.15)

We can eventually say that the dynamics:

$$\lim_{m,t\to\infty} J(\mu_{m,t}) = J^* \quad \mu_{m,t} = \frac{1}{m}\sum_{i=1}^{m} w_i^{(m)}(t)\delta_{\theta_i^{(m)}(t)}$$

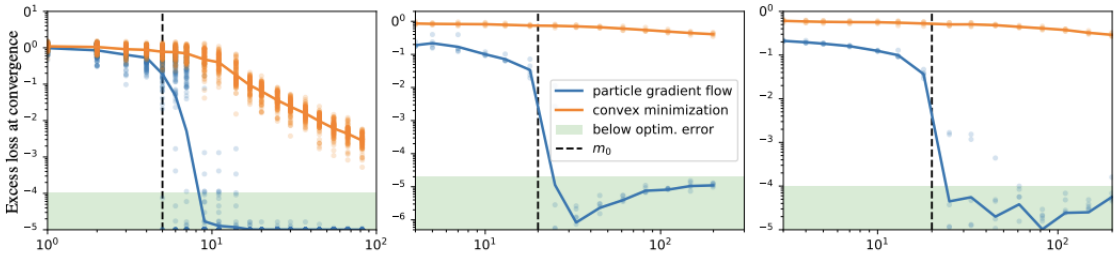
are guaranteed to converge at some non-identified m^* to the global minima of J. The convergence is independent of the order of m, t, and we could simply increase the number of particles and let them flow in t until convergence (this is by Lemma 3.15). □

4.5 A brief note on experimental results

In our final discussion, the global minimizer is always constructed to be zero.

Fixed number of particles dynamics

We work in dimension $d = 2$, and the directions on the plane at which the minimizer is attained are dotted in Figure 6. To obtain the separation condition on the support of the starting measure, the choice is using SGD with no regularization coupled with an initialization at finite m, and



(a) Sparse deconvolution ($d = 1$) (b) ReLU activation ($d = 100$) (c) Sigmoid activation ($d = 100$)

Figure 7: Empirical particle complexity of naive optimization vs particle gradient flow. Source [CB18]

Here m_0 is the minimum number of particles to have a minimizer.

θ gaussian distributed over Θ. The authors claim that this satisfies the separation condition as $m \to \infty$ [CB18]. Intuitively, adding more and more neurons around $\{0\} \times \Theta$ we will eventually separate the sets needed to escape local minima. It is also stressed that it is the *de facto* choice in practice [Bac20a]. Figure 6 is an animation of the particles moving towards the minimizing directions.

Performance

The results of this production are non-quantitative in the overparametrization required to attain reasonable performances. For this reason, the authors showed how simpler models lead to less satisfactory results and how the number of particles required is slightly above the overparametrization regime. The naïve optimization method used for comparison is based on the convex problem of the following rough routine:

- given m, sample m positions in Θ
- optimize $\{w_i\}_{i=1}^m$ over the randomly chosen candidates

Fixing the dimension at $d = 100$, the data is distributed on a sphere, and the labels are generated by a NN with 20 neurons and random normal weights. The performance in excess loss (i.e. comparison with the optimal Bayes regressor) is reasonably better with particle gradient descent[28], especially at $m > m_0$, where we identify the overparametrization regime. The lines plotted are the geometric averages over different runs at the same number of particles. For more context, see the original publication [CB18].

4.6 Summary, Weaknesses and further directions

In this document, further context was given to the very interesting piece of research titled:

On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport - Chizat, Bach (2018)

[28] again SGD and no regularization

Through the use of Wasserstein Gradient Flows, it was shown how a precise interpretation of the dynamics of two-layer neural networks can be tuned to reach global optimality. The approach is analogous to the mean-field limit of a family of functions. Experimental results are more favourable than easier optimization routines and give further support to the motivation to expand this direction of study.
Problems open in this context are quantifying the convergence rate and using the same formalism for larger networks.

Additional References

After having explored this work, I plan to strengthen my theoretical knowledge ([AGS05] is a good starting point, but there are other works) and read other Neural Networks Theory works widely cited in related literature such as [MMN18]. A quantitative result is proved in [MMM19]. The authors have continued on this line of research with a blog post and a publication [Chi20; COB20], but I am sure there is more.

References

[CB18] Lenaic Chizat and Francis Bach. *On the Global Convergence of Gradient Descent for Over-parameterized Models Using Optimal Transport*. Oct. 29, 2018. DOI: 10.48550/arXiv.1805.09545. arXiv: 1805.09545 [cs, math, stat]. URL: http://arxiv.org/abs/1805.09545 (visited on 11/20/2022).

[Ins19] Institut Henri Poincaré, director. *On the Global Convergence of Gradient Descent for (...) - Bach - Workshop 3 - CEB T1 2019*. May 10, 2019. URL: https://www.youtube.com/watch?v=rQM5uh2EsHA (visited on 12/16/2022).

[Bac20a] Francis Bach. *Gradient Descent for Wide Two-Layer Neural Networks – I : Global Convergence – Machine Learning Research Blog*. 2020. URL: https://francisbach.com/gradient-descent-neural-networks-global-convergence/ (visited on 12/16/2022).

[Chi20] Lenaic Chizat. *Gradient Descent for Wide Two-Layer Neural Networks – II: Generalization and Implicit Bias – Machine Learning Research Blog*. 2020. URL: https://francisbach.com/gradient-descent-for-wide-two-layer-neural-networks-implicit-bias/ (visited on 12/16/2022).

[Jin+18] Chi Jin et al. "On the Local Minima of the Empirical Risk". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/hash/da4902cb0bc38210839714ebdcf0efc3-Abstract.html (visited on 11/21/2022).

[Lee+] Jason D Lee et al. "Gradient Descent Only Converges to Minimizers". In: (), p. 12.

[Cho+15] Anna Choromanska et al. *The Loss Surfaces of Multilayer Networks*. Jan. 21, 2015. DOI: 10.48550/arXiv.1412.0233. arXiv: 1412.0233 [cs]. URL: http://arxiv.org/abs/1412.0233 (visited on 11/21/2022).

[SJL22] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. *Theoretical Insights into the Optimization Landscape of Over-Parameterized Shallow Neural Networks*. Aug. 23, 2022. DOI: 10.48550/arXiv.1707.04926. arXiv: 1707.04926 [cs, math, stat]. URL: http://arxiv.org/abs/1707.04926 (visited on 11/20/2022).

[JK17] Prateek Jain and Purushottam Kar. "Non-Convex Optimization for Machine Learning". In: *Foundations and Trends® in Machine Learning* 10.3-4 (2017), pp. 142–336. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/2200000058. arXiv: 1712.07897 [cs, math, stat]. URL: http://arxiv.org/abs/1712.07897 (visited on 11/21/2022).

[BP13] Kristian Bredies and Hanna Katriina Pikkarainen. "Inverse Problems in Spaces of Measures". In: *ESAIM: Control, Optimisation and Calculus of Variations* 19.1 (2013), pp. 190–218. ISSN: 1262-3377. DOI: 10.1051/cocv/2011205. URL: http://www.numdam.org/item/COCV_2013__19_1_190_0/ (visited on 11/19/2022).

[Jag13] Martin Jaggi. "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization". In: *Proceedings of the 30th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, Feb. 13, 2013, pp. 427–435. URL: https://proceedings.mlr.press/v28/jaggi13.html (visited on 11/19/2022).

[Bac16] Francis Bach. *Breaking the Curse of Dimensionality with Convex Neural Networks*. Oct. 31, 2016. DOI: 10.48550/arXiv.1412.8690. arXiv: 1412.8690 [cs, math, stat]. URL: http://arxiv.org/abs/1412.8690 (visited on 11/19/2022).

[BSR15] Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. *The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems*. July 6, 2015. arXiv: 1507.01562 [math]. URL: http://arxiv.org/abs/1507.01562 (visited on 11/19/2022).

[Wan+15] Chu Wang et al. *Functional Frank-Wolfe Boosting for General Loss Functions*. Oct. 8, 2015. DOI: 10.48550/arXiv.1510.02558. arXiv: 1510.02558 [cs, stat]. URL: http://arxiv.org/abs/1510.02558 (visited on 11/20/2022).

[Las09] Jean Bernard Lasserre. *Moments, Positive Polynomials and Their Applications*. Vol. 1. Series on Optimization and Its Applications. IMPERIAL COLLEGE PRESS, Oct. 2009. ISBN: 978-1-84816-445-1 978-1-84816-446-8. DOI: 10.1142/p665. URL: https://www.worldscientific.com/worldscibooks/10.1142/p665 (visited on 11/20/2022).

[CDP17] Paul Catala, Vincent Duval, and Gabriel Peyré. "A Low-Rank Approach to Off-The-Grid Sparse Deconvolution". In: *Journal of Physics: Conference Series* 904 (Oct. 2017), p. 012015. ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/904/1/012015. arXiv: 1712.08800 [cs, math]. URL: http://arxiv.org/abs/1712.08800 (visited on 11/19/2022).

[NS17] Atsushi Nitanda and Taiji Suzuki. *Stochastic Particle Gradient Descent for Infinite Ensembles*. Dec. 14, 2017. DOI: 10.48550/arXiv.1712.05438. arXiv: 1712.05438 [cs, math, stat]. URL: http://arxiv.org/abs/1712.05438 (visited on 11/20/2022).

[LY17] Yuanzhi Li and Yang Yuan. *Convergence Analysis of Two-layer Neural Networks with ReLU Activation*. Nov. 1, 2017. DOI: 10.48550/arXiv.1705.09886. arXiv: 1705.09886 [cs]. URL: http://arxiv.org/abs/1705.09886 (visited on 11/20/2022).

[SH17] Daniel Soudry and Elad Hoffer. *Exponentially Vanishing Sub-Optimal Local Minima in Multilayer Neural Networks*. Oct. 28, 2017. DOI: 10.48550/arXiv.1702.05777. arXiv: 1702.05777 [stat]. URL: http://arxiv.org/abs/1702.05777 (visited on 11/20/2022).

[VBB20] Luca Venturi, Afonso S. Bandeira, and Joan Bruna. *Spurious Valleys in Two-layer Neural Network Optimization Landscapes*. June 16, 2020. DOI: 10.48550/arXiv.1802.06384. arXiv: 1802.06384 [cs, math, stat]. URL: http://arxiv.org/abs/1802.06384 (visited on 11/20/2022).

[KY03] Harold Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Vol. 35. Stochastic Modelling and Applied Probability. New York: Springer-Verlag, 2003. ISBN: 978-0-387-00894-3. DOI: 10.1007/b97441. URL: http://link.springer.com/10.1007/b97441 (visited on 11/20/2022).

[Sci+17] Damien Scieur et al. *Integration Methods and Accelerated Optimization Algorithms*. Feb. 22, 2017. DOI: 10.48550/arXiv.1702.06751. arXiv: 1702.06751 [math]. URL: http://arxiv.org/abs/1702.06751 (visited on 11/20/2022).

[Jou+10] M. Journée et al. “Low-Rank Optimization for Semidefinite Convex Problems”. In: *SIAM Journal on Optimization* 20.5 (Jan. 2010), pp. 2327–2351. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/080731359. arXiv: 0807.4423 [math]. URL: http://arxiv.org/abs/0807.4423 (visited on 11/19/2022).

[HV17] Benjamin D. Haeffele and Rene Vidal. “Global Optimality in Neural Network Training”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, July 2017, pp. 4390–4398. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.467. URL: http://ieeexplore.ieee.org/document/8099950/ (visited on 11/19/2022).

[Coh13] Donald L. Cohn. *Measure Theory*. Birkhäuser Advanced Texts Basler Lehrbücher. New York, NY: Springer, 2013. ISBN: 978-1-4614-6955-1 978-1-4614-6956-8. DOI: 10.1007/978-1-4614-6956-8. URL: http://link.springer.com/10.1007/978-1-4614-6956-8 (visited on 11/19/2022).

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. “A Mean Field View of the Landscape of Two-Layer Neural Networks”. In: *Proceedings of the National Academy of Sciences* 115.33 (Aug. 14, 2018), E7665–E7671. DOI: 10.1073/pnas.1806579115. URL: https://www.pnas.org/doi/10.1073/pnas.1806579115 (visited on 11/20/2022).

[RV19] Grant M. Rotskoff and Eric Vanden-Eijnden. *Trainability and Accuracy of Neural Networks: An Interacting Particle System Approach*. July 30, 2019. DOI: 10.48550/arXiv.1805.00915. arXiv: 1805.00915 [cond-mat, stat]. URL: http://arxiv.org/abs/1805.00915 (visited on 11/20/2022).

[SS19] Justin Sirignano and Konstantinos Spiliopoulos. *Mean Field Analysis of Neural Networks: A Law of Large Numbers*. Nov. 11, 2019. DOI: 10.48550/arXiv.1805.01053. arXiv: 1805.01053 [math]. URL: http://arxiv.org/abs/1805.01053 (visited on 11/20/2022).

[Bac20b] Francis Bach. *Effortless Optimization through Gradient Flows – Machine Learning Research Blog*. 2020. URL: https://francisbach.com/gradient-flows/ (visited on 12/26/2022).

[San17] Filippo Santambrogio. “{Euclidean, Metric, and Wasserstein} Gradient Flows: An Overview”. In: *Bulletin of Mathematical Sciences* 7.1 (Apr. 2017), pp. 87–154. ISSN: 1664-3607, 1664-3615. DOI: 10.1007/s13373-017-0101-1. URL: http://link.springer.com/10.1007/s13373-017-0101-1 (visited on 11/20/2022).

[San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. 1st ed. 2015. Progress in Nonlinear Differential Equations and Their Applications 87. Cham: Springer International Publishing : Imprint: Birkhäuser, 2015. 1 p. ISBN: 978-3-319-20828-2. DOI: 10.1007/978-3-319-20828-2.

[CG21] Piermarco Cannarsa and Filippo Gazzola. *Dynamic Optimization for Beginners: With Prerequisites and Applications*. 1st ed. EMS Press, Oct. 11, 2021. ISBN: 978-3-9854701-2-9 978-3-9854751-2-4. DOI: 10.4171/etb/23. URL: https://ems.press/doi/10.4171/etb/23 (visited on 12/28/2022).

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: The MIT Press, 2016. 775 pp. ISBN: 978-0-262-03561-3.

[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savarè. *Gradient Flows*. Lectures in Mathematics ETH Zürich. Basel: Birkhäuser-Verlag, 2005. ISBN: 978-3-7643-2428-5. DOI: 10.1007/b137080. URL: http://link.springer.com/10.1007/b137080 (visited on 11/20/2022).

[Dud02] R. M. Dudley. *Real Analysis and Probability*. 2nd ed. Cambridge University Press, Oct. 14, 2002. ISBN: 978-0-521-80972-6 978-0-521-00754-2 978-0-511-75534-7. DOI: 10.1017/CBO9780511755347. URL: https://www.cambridge.org/core/product/identifier/9780511755347/type/book (visited on 01/07/2023).

[Bol08] F. Bolley. "Separability and Completeness for the Wasserstein Distance". In: *Séminaire de Probabilités XLI*. Ed. by Catherine Donati-Martin et al. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer, 2008, pp. 371–377. ISBN: 978-3-540-77913-1. DOI: 10.1007/978-3-540-77913-1_17. URL: https://doi.org/10.1007/978-3-540-77913-1_17 (visited on 01/05/2023).

[Chi21] Lénaïc Chizat. "Analysis of Gradient Descent on Wide Two-Layer ReLU Neural Networks". In: (2021).

[Bro83] Felix E. Browder. "Fixed Point Theory and Nonlinear Problems". In: *Bulletin (New Series) of the American Mathematical Society* 9.1 (July 1983), pp. 1–39. ISSN: 0273-0979, 1088-9485. URL: https://projecteuclid.org/journals/bulletin-of-the-american-mathematical-society-new-series/volume-9/issue-1/Fixed-point-theory-and-nonlinear-problems/bams/1183550974.full (visited on 11/19/2022).

[Whi35] Hassler Whitney. "A Function Not Constant on a Connected Set of Critical Points". In: *Duke Mathematical Journal* 1.4 (Dec. 1935), pp. 514–517. ISSN: 0012-7094, 1547-7398. DOI: 10.1215/S0012-7094-35-00138-7. URL: https://projecteuclid.org/journals/duke-mathematical-journal/volume-1/issue-4/A-function-not-constant-on-a-connected-set-of-critical/10.1215/S0012-7094-35-00138-7.full (visited on 11/20/2022).

[AMA05] P. A. Absil, R. Mahony, and B. Andrews. "Convergence of the Iterates of Descent Methods for Analytic Cost Functions". In: *SIAM Journal on Optimization* 16.2 (Jan. 2005), pp. 531–547. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/040605266. URL: http://epubs.siam.org/doi/10.1137/040605266 (visited on 11/19/2022).

[HM19] Daniel Hauer and José Mazon. *Kurdyka-Lojasiewicz-Simon Inequality for Gradient Flows in Metric Spaces*. Jan. 24, 2019. arXiv: 1707.03129 [math]. URL: http://arxiv.org/abs/1707.03129 (visited on 11/19/2022).

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Mean-Field Theory of Two-Layers Neural Networks: Dimension-Free Bounds and Kernel Limit". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Conference on Learning Theory. PMLR, June 25, 2019, pp. 2388–2464. URL: https://proceedings.mlr.press/v99/mei19a.html (visited on 01/07/2023).

[Hay99] Simon S. Haykin. *Neural Networks: A Comprehensive Foundation*. 2nd ed. Upper Saddle River, N.J: Prentice Hall, 1999. 842 pp. ISBN: 978-0-13-273350-2.

[Bar93] A.R. Barron. "Universal Approximation Bounds for Superpositions of a Sigmoidal Function". In: *IEEE Transactions on Information Theory* 39.3 (May 1993), pp. 930–945. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/18.256500. URL: https://ieeexplore.ieee.org/document/256500/ (visited on 11/21/2022).

[KS01] V. Kurkova and M. Sanguineti. "Bounds on Rates of Variable-Basis and Neural-Network Approximation". In: *IEEE Transactions on Information Theory* 47.6 (Sept./2001), pp. 2659–2665. ISSN: 00189448. DOI: 10.1109/18.945285. URL: http://ieeexplore.ieee.org/document/945285/ (visited on 11/21/2022).

[Ben+05] Yoshua Bengio et al. "Convex Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 18. MIT Press, 2005. URL: https://papers.nips.cc/paper/2005/hash/0fc170ecbb8ff1afb2c6de48ea5343e7-Abstract.html (visited on 11/21/2022).

[Ros+07] Saharon Rosset et al. "$\ell 1$ Regularization in Infinite Dimensional Feature Spaces". In: *Learning Theory*. Ed. by Nader H. Bshouty and Claudio Gentile. Vol. 4539. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 544–558. ISBN: 978-3-540-72925-9. DOI: 10.1007/978-3-540-72927-3_39. URL: http://link.springer.com/10.1007/978-3-540-72927-3_39 (visited on 11/21/2022).

[COB20] Lenaic Chizat, Edouard Oyallon, and Francis Bach. *On Lazy Training in Differentiable Programming*. Jan. 7, 2020. DOI: 10.48550/arXiv.1812.07956. arXiv: 1812.07956 [cs, math]. URL: http://arxiv.org/abs/1812.07956 (visited on 11/21/2022).

[SS05] Wilson A. Sutherland and Wilson Alexander Sutherland. *Introduction to Metric and Topological Spaces*. Reprint. Oxford Science Publications. Oxford: Clarendon Press, 2005. 181 pp. ISBN: 978-0-19-853161-6.

[Roc70] Ralph Tyrell Rockafellar. *Convex Analysis:* Princeton University Press, Dec. 31, 1970. ISBN: 978-1-4008-7317-3. DOI: 10.1515/9781400873173. URL: https://www.degruyter.com/document/doi/10.1515/9781400873173/html (visited on 11/20/2022).

[Bou98] Nicolas Bourbaki. *Elements of Mathematics. Chapters 5/10: 3. General Topology*. Softcover ed., [Nachdr.] Berlin Heidelberg: Springer, 1998. 363 pp. ISBN: 978-3-540-64563-4 978-3-540-19372-2.

[Du16] Lingyu Du. "Sard's Theorem and Applications". In: (2016).

[Çin11] Erhan Çinlar. *Probability and Stochastics*. Vol. 261. Graduate Texts in Mathematics. New York, NY: Springer, 2011. ISBN: 978-0-387-87858-4 978-0-387-87859-1. DOI: 10.1007/978-0-387-87859-1. URL: http://link.springer.com/10.1007/978-0-387-87859-1 (visited on 11/07/2022).

[Fis12] Tom Fischer. *Existence, Uniqueness, and Minimality of the Jordan Measure Decomposition*. June 27, 2012. arXiv: 1206.5449 [math, stat]. URL: http://arxiv.org/abs/1206.5449 (visited on 12/22/2022).

[Bog07] Vladimir I. Bogachev. *Measure Theory*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. ISBN: 978-3-540-34513-8 978-3-540-34514-5. DOI: 10.1007/978-3-540-34514-5. URL: http://link.springer.com/10.1007/978-3-540-34514-5 (visited on 12/23/2022).

[Vil09] Cédric Villani. *Optimal Transport*. Red. by M. Berger et al. Vol. 338. Grundlehren Der Mathematischen Wissenschaften. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. ISBN: 978-3-540-71049-3 978-3-540-71050-9. DOI: 10.1007/978-3-540-71050-9. URL: http://link.springer.com/10.1007/978-3-540-71050-9 (visited on 12/26/2022).

[WZ77] Richard L. Wheeden and Antoni Zygmund. *Measure and Integral: An Introduction to Real Analysis*. Monographs and Textbooks in Pure and Applied Mathematics ; 43. New York: M. Dekker, 1977. 274 pp. ISBN: 978-0-8247-6499-9.

[Bre11] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. New York, NY: Springer New York, 2011. ISBN: 978-0-387-70913-0 978-0-387-70914-7. DOI: 10.1007/978-0-387-70914-7. URL: https://link.springer.com/10.1007/978-0-387-70914-7 (visited on 12/29/2022).

[Rud13] Walter Rudin. *Real and Complex Analysis*. 3. ed., internat. ed., [Nachdr.] McGraw-Hill International Editions Mathematics Series. New York, NY: McGraw-Hill, 2013. 416 pp. ISBN: 978-0-07-100276-9 978-0-07-054234-1.

[Fan+22] Jiaojiao Fan et al. "Variational Wasserstein Gradient Flow". In: *Proceedings of the 39th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, June 28, 2022, pp. 6185–6215. URL: https://proceedings.mlr.press/v162/fan22d.html (visited on 01/05/2023).

[Rot20] Grant M. Rotskoff. *Wasserstein Gradient Flows and the Fokker Planck Equation (Part I)*. Rotskoff Group. May 26, 2020. URL: https://statmech.stanford.edu/post/gradient_flows_00/ (visited on 01/05/2023).

[Ans20] Abdul Fatir Ansari. *Introduction to Gradient Flows in the 2-Wasserstein Space*. 2020. URL: https://abdulfatir.com/blog/2020/Gradient-Flows/ (visited on 01/05/2023).

[Amb03] Luigi Ambrosio. "Lecture Notes on Optimal Transport Problems". In: Luigi Ambrosio et al. *Mathematical Aspects of Evolving Interfaces*. Vol. 1812. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 1–52. ISBN: 978-3-540-14033-7 978-3-540-39189-0. DOI: 10.1007/978-3-540-39189-0_1. URL: http://link.springer.com/10.1007/978-3-540-39189-0_1 (visited on 12/29/2022).

[AG13] Luigi Ambrosio and Nicola Gigli. "A User's Guide to Optimal Transport". In: Luigi Ambrosio et al. *Modelling and Optimisation of Flows on Networks*. Vol. 2062. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 1–155. ISBN: 978-3-642-32159-7 978-3-642-32160-3. DOI: 10.1007/978-3-642-32160-3_1. URL: http://link.springer.com/10.1007/978-3-642-32160-3_1 (visited on 12/29/2022).

Appendix

We report in the Appendix some facts not directly related to the results but needed for their understanding. In Appendix A the tools used throughout the literature are mentioned. In B more paper-specific notions are explained. As a disclaimer, the derivation of the continuity equation in the distributional sense is not self-contained, but hopefully the references will help.

A Required Notions

A.1 Set Theory

There are plenty of equivalent definitions for the next three objects, which are used throughout the paper extensively.

Definition A.1 (Closure of a set $\overline{A}$)**.** *The closure of a set A, denoted as $\overline{A}$ is the smallest closed set containing A.*

Definition A.2 (Interior of a set $\mathtt{int}\,(A)$)**.** *The interior of a set A, denoted as $\mathtt{int}\,A$ is the largest open set contained in A.*

Definition A.3 (Boundary of a set ∂A)**.** *We use the notation $\partial A \coloneqq \overline{A} \setminus \mathtt{int}\,(A)$.*

Proposition A.4 (Boundary is a closed set)**.** *The boundary ∂A is closed.*

Proof. Trivially $\partial A = \overline{A} \setminus \mathtt{int}\,(A) = \overline{A} \cap (\mathtt{int}\,(A))^c$ where $\mathtt{int}\,(A)$ is open and has a closed complement. The intersection of closed sets is closed. □

Definition A.5 (Connectedness)**.** *A topological space $(\mathcal{X}, \mathcal{T})$ is connected if it cannot be represented as the union of two disjoint nonempty open sets. Similarly, we could have subspaces which are connected.*
We say a subset of X is a connected component if making it bigger implies losing connectedness.

Proposition A.6 (Closure of connected set is connected)**.**

$$A \text{ connected} \implies \overline{A} \text{ connected}$$

Proof. Assume by contradiction that A is connected but $\overline{A}$ is not. Then, $\overline{A}$ is the union of (at least two) disjoint nonempty open sets. By being open, it holds that:

$$\overline{A} = B \cup C,\ B \cap C = \emptyset \implies B \subset A, C \subset A \implies A = (A \cap B) \cup (A \cap C)$$

which contradicts the assumption that A is connected. □

Proposition A.7 (Connected components are closed)**.** *For a topological space $(\mathcal{X}, \mathcal{T})$, if $A \subset \mathcal{X}$ is a connected component, A is closed.*

Proof. We have for free that $A \subset \overline{A}$, and need to show the opposite. By Proposition A.6, for A connected $\overline{A}$ is. By A being a connected component, it is the largest possible, so that $A \supset \overline{A}$. Eventually $A = \overline{A}$ and A is closed. □

Definition A.8 (Path-connectedness). *Two points x, x' in a topological space $(\mathcal{X}, \mathcal{T})$ are path-connected if there exists a curve:*

$$\exists \gamma : [0,1] \to \mathcal{X} \text{ continuous} \quad \gamma(0) = x, \quad \gamma(1) = x'$$

Clearly, such formalism allows to split the topological space into equivalence classes (i.e. collections of path connected points).

Observation A.9 (Connectedness notions). *The existance of a connection curve is not equivalent to the concept of connected component. Path connectedness implies connectedness, this is easily proved by contradiction. The opposite is generally not true, but will hold in our application since we are in $\mathbb{R}^n$. This is proved in [SS05](Prop. 6.4.2).*

A.2 Analysis

Definition A.10 (Sublevel sets). *For a real valued function $f : \mathbb{R}^d \to \mathbb{R}$ the sublevel sets for a given value c are clearly:*

$$\left\{ x \in \mathbb{R}^d \; ; \; f(x) \leq c \right\}$$

Definition A.11 (Supremum norm $\|\cdot\|_\infty$). *For functions defined on a set $\mathcal{X}$ the supremum norm is:*

$$\|f\|_\infty = \sup \left\{ |f(x)| : x \in \mathcal{X} \right\}$$

Definition A.12 (Cauchy sequence). *A sequence (f_n) in a metric space $(\mathcal{F}, \|\cdot\|)$ is Cauchy when:*

$$\forall n, m > N \; \exists \epsilon \quad \|f_n - f_m\| < \epsilon$$

Definition A.13 (Complete space). *A space $\mathcal{F}$ is complete with respect to a distance $\|\cdot\|$ if each Cauchy sequence is convergent to an element of $\mathcal{F}$.*

Definition A.14 (Separable space). *A topological space is separable if it contains a countably dense subset. Namely, a sequence such that every non-empty open set of the topology contains at least one element of the sequence.*

Definition A.15 (Hilbert space). *A $\mathcal{F}$ is a Hilbert space when it is paired with a valid inner product, complete with respect to the distance induced by the inner product.*
In the context of this paper, $\mathcal{F}$ is endowed with an inner product $\langle \cdot, \cdot \rangle$ and a norm $\|\cdot\|$ by Assumptions 1.10.

Proposition A.16 (Norm is sup of inner products). *For a normed space $(\mathcal{X}, \|\cdot\|)$ with an inner product it holds:*

$$\|x\| = \sup_{x' \in \mathcal{X}, \|x'\| = 1} \left\langle x, x' \right\rangle$$

And enlarging the space of functions over which we take the sup *we get the trivial inequality:*

$$\|x\| \leq \sup_{x' \in \mathcal{X}, \|x'\| \leq 1} \left\langle x, x' \right\rangle$$

Proof. By Cauchy Schwarz and $\|x'\| = 1$:

$$\left\|\langle x, x'\rangle\right\| \leq \|x\| \left\|x'\right\| = \|x\|$$

For the reverse inequality observe that:

$$\begin{aligned} \|x\|^2 = \langle x, x\rangle = \|x\| \left\langle x, \frac{x}{\|x\|} \right\rangle & \qquad \left\| \frac{x}{\|x\|} \right\| = 1 \\ \leq \|x\| \sup\left\{\langle x, x'\rangle : \left\|x'\right\| = 1\right\} & \\ \implies \|x\| \leq \sup\left\{\langle x, x'\rangle : \left\|x'\right\| = 1\right\} & \end{aligned}$$

□

Definition A.17 (Semiconvexity or λ convexity)**.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ such that for some $\lambda \in \mathbb{R}$ the function $f + \lambda|\cdot|^2$ is convex*

Proposition A.18 (Semiconvexity of smooth functions over compact domain)**.**

$$f : \mathcal{X} \text{ compact} \to \mathbb{R} \text{ smooth} \implies f \text{ semiconvex}$$

Definition A.19 (Positive p-homogeneity)**.** *for vector spaces $\mathcal{X}, \mathcal{Y}$ a function $f : \mathcal{X} \to \mathcal{Y}$ is positively p-homogeneous whenever:*

$$f(\lambda x) = \lambda^p f(x) \quad \forall \lambda > 0, \forall x > 0$$

Proposition A.20 (Properties of homogeneous functions)**.** *Consider a positively p-homogeneous function f. Then:*

1. *the (sub)derivative is positive $(p-1)$-homogeneous*
2. *for f differentiable (neglect $u = 0$) it holds:*

$$u \cdot \nabla f(u) = p f(u) \quad u \neq 0$$

Proposition A.21 (Local semiconvexity via sum of differentiable homogeneous and semiconvex)**.** *Consider two functions f and g. For f continuous, differentiable, 1-homogeneous, and g semiconvex (Def. A.17) it holds that:*

$$h = f + g \quad \text{semiconvex } (\textit{locally})$$

meaning that $\forall x_0$ there is a neighborhood that is semiconvex.

Proof. Without loss of generality we work in $\mathbb{R}$.
From the hypothesis, f is differentiable and continuous and h is such that for $\lambda \in \mathbb{R}$ the map $x \to g(x) + \lambda|x|^2$ is convex, namely for all $\alpha \in (0,1)$ and x, x':

$$g(\alpha x + (1-\alpha)x') + \lambda|\alpha x + (1-\alpha)x')|^2 \leq \alpha\Big(g(x) + \lambda|x|^2\Big) + (1-\alpha)\Big(g(x') + \lambda|x'|^2\Big)$$

Consider h evaluated at a point x_0. Then interpolating with a factor α and another arbitrary point x we conclude that for the same coefficient $\lambda \in \mathbb{R}$:

$$\begin{aligned} h(\alpha x_0 + (1-\alpha)x) + \lambda|\alpha x_0 + (1-\alpha)x|^2 &= f(\alpha x_0 + (1-\alpha)x) + g(\alpha x_0 + (1-\alpha x)) \\ &\quad + \lambda|\alpha x_0 + (1-\alpha)x|^2 \\ &\leq f(\alpha x_0 + (1-\alpha)x) + \alpha\Big(g(x_0) + \lambda|x_0|^2\Big) \\ &\quad + (1-\alpha)\Big(g(x) + \lambda|x|^2\Big) \end{aligned}$$

Where we used the semiconvexity of g. We now concentrate on the f part and use continuity and differentiability to perform a Taylor expansion around αx_0 to obtain an affine function. Doing so, since we aim to prove local subconvexity, we impose that the convex combination belongs to a neighborhood of x_0, in the sense:

$$\alpha x_0 + (1-\alpha)x \approx x \quad : \quad \alpha \to 1$$

$$\begin{aligned} f(\alpha x_0 + (1-\alpha)x) &= f(\alpha x_0) + \nabla f\Big|_{\alpha x_0} (1-\alpha)x + \mathcal{O}(|(1-\alpha)x|) && \text{Taylor} \\ &= \alpha f(x_0) + \nabla f\Big|_{\alpha x_0} (1-\alpha)x + \mathcal{O}(|(1-\alpha)x|) && \text{homogeneity of } f \\ &= \alpha f(x_0)\alpha^0 \nabla f\Big|_{x_0} (1-\alpha)x + \mathcal{O}(|(1-\alpha)x|) && \text{Prop. A.20\#1} \end{aligned}$$

Where by Proposition A.20#1 we also derive that ∇f is constant (0-homogeneous) so that $\nabla f(x_0) \approx \nabla f(x)$ and we conclude that in a neighborhood of x_0 where the $\mathcal{O}$ correction is neglected:

$$\begin{aligned} f(\alpha x_0 + (1-\alpha)x) &= \alpha f(x) + (1-\alpha)x\nabla f(x) \\ &= \alpha f(x) + (1-\alpha) f(x) && \text{Prop. A.20\#2} \end{aligned}$$

Which concludes the initial computation on h, proving it is semiconvex locally for the same λ. □

Definition A.22 (Lipschitz function)**.** *A function f such that:*

$$\left\| f(x) - f(x') \right\| \leq C \left\| x - x' \right\| \quad \forall x, y \in \mathcal{X} \times \mathcal{X} \quad \textit{for some } C \in \mathbb{R}$$

Namely the growth is bounded by the norm.

Proposition A.23 (Lipschitz derivative bounds Taylor expansions in norm)**.** *For a C^2 function $f : \mathcal{X} \to \mathcal{Y}$ with L-Lipschitz derivative it holds that the norm of the remainder of a first order expansion is bounded:*

$$\|f(x+h) - f(x) - df_x(h)\|_{\mathcal{Y}} \leq \frac{L}{2} \|h\|_{\mathcal{X}}^2$$

Proof. Doing a first order Taylor expansion by continuity:

$$f(x+h) = f(x) + df_x \cdot h + \int_0^1 (df_{x+th} - df_x)h \; dt$$

And clearly:

$$\begin{aligned} \|f(x+h) - f(x) - df_x \cdot h\|_{\mathcal{Y}} &= \left\| \int_0^1 (df_{x+th} - df_x)h \; dt \right\|_{\mathcal{Y}} \\ &\leq \int_0^1 \|(df_{x+th} - df_x)h\|_{\mathcal{X}} \; dt \\ &\leq \int_0^1 L \|th\|_{\mathcal{X}} \|h\|_{\mathcal{X}} \; dt && \text{Lipschitz derivative} \\ &= \frac{L}{2} \|h\|_{\mathcal{X}}^2 \end{aligned}$$

□

Proposition A.24 (Projection onto convex set is Lipschitz)**.** *Let $K \subset \mathcal{X}$ be closed and convex and $\mathcal{X}$ is normed, then:*

$$\left\|\texttt{proj}_K(x) - \texttt{proj}_K(x')\right\| \leq \left\|x - x'\right\|$$

Proof. By closedness of K it is possible to attain the minimization. By convexity:

$$\begin{aligned}\|x - \texttt{proj}_K(x)\|^2 &\leq \|x - [\lambda k + (1-\lambda)\texttt{proj}_K(x)]\|^2 \\ &= \|x - \texttt{proj}_K(x) - \lambda(k - \texttt{proj}_K(x))\|^2 \qquad \forall k \in K, \forall \lambda \in (0,1)\end{aligned}$$

using the inner product and letting $\lambda \to 0$ we derive:

$$\langle x - \texttt{proj}_K(x), k - \texttt{proj}_K(x)\rangle \leq 0 \quad \forall k \in K$$

The same holds for x':

$$\left\langle x' - \texttt{proj}_K(x'), k - \texttt{proj}_K(x')\right\rangle \leq 0 \quad \forall k \in K$$

Using the particular solution for $k = \texttt{proj}_K(x')$ in the former and $k = \texttt{proj}_K(x)$ in the latter we get:

$$\left\langle x - y + [\texttt{proj}_K(x') - \texttt{proj}_K(x)], \texttt{proj}_K(x') - \texttt{proj}_K(x)\right\rangle \leq 0$$

from which we conclude that:

$$\left\|\texttt{proj}_K(x') - \texttt{proj}_K(x)\right\|^2 \leq \left\langle x - x', \texttt{proj}_K(x') - \texttt{proj}_K(x)\right\rangle \leq \left\|x' - x\right\| \left\|\texttt{proj}_K(x') - \texttt{proj}_K(x)\right\|$$

and simplyfying the equation we get the claim. □

Definition A.25 (Subgradient)**.** *For function $f : \mathbb{R}^d \to \overline{R}$ and a point $u_0 \in \mathbb{R}^d$ a subgradient is defined as:*

$$p \in \mathbb{R}^d \;:\; f(u) \geq f(u_0) + p \cdot (u - u_0) + o(u - u_0) \quad \forall u \in \mathbb{R}^d$$

Definition A.26 (Subdifferential of a function at a point $\partial f(u)$)**.** *For a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ and a point $u \in \mathbb{R}^d$ the set of subgradients is denoted as $\partial f(u)$*

Proposition A.27 (Subdifferential structure)**.** *$\partial f(u)$ is closed and convex $\forall u$.*

Proof. [Roc70]. □

Definition A.28 (Coercive map)**.** *A linear and continuous map $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is Banach and $\mathcal{Y}$ is Hilbert is coercive if:*

$$\exists c > 0 \quad : \quad \|f(x)\|_{\mathcal{Y}} \geq c \|x\|_{\mathcal{X}} \quad \forall x \in \mathcal{X}$$

Definition A.29 (Closed map)**.** *A map $f : \mathcal{X} \to \mathcal{Y}$ between Banach spaces such that for $x_n \to x$ in $\mathcal{X}$ and $f(x_n) \to y$ in $\mathcal{Y}$ it holds $f(x) = y$. Here convergence is in the norm of the Banach space.*

Proposition A.30 (Characterization of Coercive maps)**.** *A continuous map f between two Banach spaces $\mathcal{X}, \mathcal{Y}$ is coercive then it is closed.*

$$\exists c > 0 : \forall x \in \mathcal{X} \quad \|f(x)\|_{\mathcal{Y}} \geq c \|x\|_{\mathcal{X}} \implies f \text{ closed}$$

Proof. let $(y_n) \subset f(\mathcal{X}) \subset \mathcal{Y}$. Since $\mathcal{Y}$ is Banach we know $(y_n) \to y \in \mathcal{Y}$, and by assumption we also have:

$$(x_n) \subset \mathcal{X} \quad f(x_n) = y_n, \quad (x_n) \to x \in \mathcal{X}$$

Clearly then:

$$0 \leq \|y_n - f(x)\| = \|f(x_n) - f(x)\|$$

And by continuity if $\|x_n - x\| \to 0$ so does the function application. We then force the value to be between the continuity bound above and the coercivity bound below. By squeezing, it holds:

$$\|y - f(x_n)\|_{\mathcal{Y}} \to 0 \implies f(x_n) = y_n \to y \in \mathcal{Y}$$

Which is in line with the definition of closedness (Def. A.29). □

Lemma A.31 (Grownwall's Lemma)**.** *For f be a real valued continuous function on an interval I, differentiable on the interior I^0 of I. We say the interior is (a, b). Let β be continuous and real valued on I as well. Then:*

$$f'(t) \leq \beta(t)f(t) \implies f(t) \leq f(a) \exp\left\{ \int_a^t \beta(s)ds \right\} \quad \forall t \in I$$

Definition A.32 (Vector field)**.** *A vector field is simply a vector valued function on a space.*

In out case, the velocity is exactly a vector field.

Definition A.33 (Divergence operator `div`)**.** *For a smooth vector field $E = (E_i)_{i=1}^d : \mathbb{R}^d \to \mathbb{R}^d$ we define the divergence as:*

$$\texttt{div}(E) = \sum_{i=1}^{d} \frac{\partial E_i}{\partial x_i}$$

for any given point in $\mathbb{R}^d$, it can be seen as the density of the outward flux at the limit.

A.2.1 Arzelà–Ascoli Theorem

Definition A.34 (Uniformly bounded family)**.** *A collection of continuous functions $(f_n)_{n\in\mathbb{N}}$ is uniformly bounded on $I = [a, b]$ when:*

$$|f_n(x)| \leq M \quad \forall x \in [a, b], \forall n$$

Definition A.35 (Uniform equicontinuity)**.** *A sequence of functions $(f_n)_{n\in\mathbb{N}}$ is uniformly equicontinuous if each function shares the same limiting constants. Namely:*

$$\forall \epsilon > 0 \exists \delta(\epsilon) := \delta > 0 \quad : \quad |x - x'| < \delta \implies |f_n(x) - f_n(x')| < \epsilon \quad \forall n$$

Definition A.36 (Uniform convergence $\rightrightarrows$)**.** *A sequence $(f_n)_{n\in\mathbb{N}}$ such that $f_n : E \to \mathbb{R}$ is uniformly convergent to $f : E \to \mathbb{R}$ if:*

$$\forall \epsilon > 0 \exists N \in \mathbb{N} \text{ such that } \forall n \geq N, \forall x \in E \quad |f_n(x) - f(x)| < \epsilon$$

Namely, there is a common bound on the distance after sufficiently large n. It is often written as $f_n \rightrightarrows f$.

In our setting we work with Topologies, but it is important to get the idea of the Theorem we will need.

Theorem A.37 (Arzelà–Ascoli Theorem)**.** *We provide a weaker result. If a sequence of functions $(f_n)_{n\in\mathbb{N}}$ continuous and real valued defined on $I = [a, b] \subset \mathbb{R}$ is such that equicontinuity and boundedness hold uniformly over an interval I (Defs. A.34, A.35) then $\exists k \to n(k)$ subsequence $(f_{n(k)})_{k\in\mathbb{N}}$ which is uniformly convergent to some f over I.*

Definition A.38 (Topology of compact convergence)**.** *For a topological space $(X, \mathfrak{T})$ and a metric space (Y, d_Y) a sequence of functions $(f_n)_{n\in\mathbb{N}}$ where $f_n : X \to Y \forall n$ converges compactly to $f : X \to Y$ when:*

$$\forall K \subseteq X \ : \ compact \quad f_n(x) \rightrightarrows f(x) \forall x \in K$$

Or in other words:

$$\lim_{n\to\infty} \sup_{x\in K} d_Y(f_n(x), f(x)) = 0$$

Which is another classical definition of uniform convergence (Def. A.36) for metric spaces. .

Definition A.39 (Relative compactness or precompactness)**.** *For a topological space X a subset $K \subset X$ is precompact when its closure $\overline{K}$ is compact.*

Theorem A.40 (Topological Arzelà–Ascoli Theorem)**.** *A sequence of functions f_n such that $f : X \to Y$ where:*

- *X is topological*
- *Y is Hausdorff uniform*

collected in the bigger family $\mathfrak{F}(X, Y)$ of functions, belonging to the space of continuous functions $C(X, Y)$ together with the topology of compact convergence (Def. A.38) is such that if $H \subset C(X, Y)$ is a set of equicontinuous functions then

$$H(x) \ precompact \ in \ Y \ (Def.A.39) \ \forall x \in X \implies H \ precompact \ in \ C(X, Y)$$

Proof. [Bou98](Chap. X, Num. 2, nr. 5). □

A.2.2 Comments about Sard-type regularity

This formulation is sufficient for the context of our application. Note that when referring to points we mean elements of the domain, while values are elements of the codomain of a function. A good introductory reference with proofs of the two results is [Du16].

Definition A.41 (Regular values of a function)**.** *For $g : \Theta \to \mathbb{R}$ a regular value $\alpha \in \mathbb{R}$ satisfies:*

- *(part of range) $\exists \theta : g(\theta) = \alpha$*
- *(regularity) $g^{-1}(\alpha) \subset A_{good} \subseteq \Theta$ where A_{good} is such that:*
  - *it is open*
  - *for $\theta \in A_{good}$ g is differentiable*
  - *for $\theta \in A_{good}$ $dg \neq 0$*

If the inverse image is null, by convention α is a regular value.

Definition A.42 (Critical points of a function)**.** *For a function g, a point θ for which there are no associated regular values in the codomain is said to be a critical point. Intuitively, it is a point in which the derivative is 0 or singular (null or inexistent in a neighborhood). For a euclidean function $f : \mathbb{R}^n \to \mathbb{R}^m$ as in our case, this means that the Jacobian has rank $< m$.*

Definition A.43 (Measure zero set)**.** *A set $A \subset \mathbb{R}^m$ has measure zero if:*

$$\forall \epsilon > 0 \exists (U_n)_{n\in\mathbb{N}},\ U_n \text{ open } \forall n \quad \bigcup_n U_n \supset A, \quad \sum_n |U_n| < \epsilon$$

Namely, there is an arbitrary small in volume (norm) countable collection of sets that covers the whole set A. This idea originates from the fact that sets of this kind have the peculiarity that their realization is negligible.

Theorem A.44 (Regular value Theorem)**.** *Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be smooth and $\alpha \in \mathbb{R}^m$ be a regular value of g. Then $g^{-1}(\alpha)$ is a submanifold of dimension $n - m$.*

Lemma A.45 (Morse-Sard Lemma)**.** *Consider a function $f : \mathbb{R}^n \to \mathbb{R}^m$. Let the set $C \subset \mathbb{R}^n$ be the collection of all critical points (Def. A.42) of the function f. Then, $f(C) \subset \mathbb{R}^m$ has measure zero in the sense of Definition A.43.*

A.3 Measure Theory

A good reference from the statistics side is the first two chapters of [Çin11].

Definition A.46 (Set of measures $\mathcal{M}(\Theta)$ setting)**.** *With the symbol $\mathcal{M}(\Theta)$ the authors refer to **finite** (i.e. $\mu(\Theta) < \infty$) signed measures on $\mathbb{R}^d$, endowed with the borel sigma algebra $\mathscr{B}(\Theta)$, concentrated on $X \subset \mathbb{R}^d$.*

Definition A.47 (Support $\mathtt{spt}\cdot$)**.** *For support of a measure we mean the complement of the largest open set of measure zero. Equivalently, it could be the closed set of points that have non zero measure neighborhoods.*

Definition A.48 (Concentrated measure)**.** *We say a measure is concentrated on $X \subset \mathbb{R}^d$ when $S^c \subset N : \mu(N) = 0$. Clearly then μ is concentrated on $\mathtt{spt}\,\mu$*

Theorem A.49 (Jordan Decomposition Theorem)**.** *For any $\mu \in \mathcal{M}(\mathbb{R}^d)$ there is a decomposition:*

$$\mu = \mu_+ - \mu_-, \quad \mu_+, \mu_- \in \mathcal{M}_+(\mathbb{R}^d)$$

Proof. [Coh13](Cor. 4.1.6), [Fis12]. □

Corollary A.50 (Minimality property of Jordan decomposition)**.** *If a measure μ has a Jordan decomposition, then:*

$$\mu = \mu_+ - \mu_- \quad \mu_+, \mu_- \in \mathcal{M}_+(\mathbb{R}^d)$$

It holds:

$$\mu_+(B) = \sup_{A\in\mathcal{B}(\mathbb{R}^d), A\subset B} \mu(A) \quad \mu_-(B) = -\inf_{A\in\mathcal{B}(\mathbb{R}^d), A\subset B} \mu(A)$$

And for any other decomposition of μ into finite non-negative measures on $\mathbb{R}^d$ it holds:

$$\mu = \nu_+ - \nu_- \implies \nu_+ \geq \mu_+ \quad \nu_- \geq \mu_-$$

Namely, the Jordan decomposition is the minimal decomposition of μ into non negative measures.

Proof. [Fis12]. □

Definition A.51 (Variation and total variation norm)**.** *Upon choosing μ_+, μ_- with minimal total mass ($\mu_+(\mathbb{R}^d)$ minimal and similar), which we can choose since the measures are assumed to be finite (Def. A.46), define:*

- *Variation $|\mu| = \mu_+ + \mu_-$, a measure*
- *total variation norm $|\mu|(\mathbb{R}^d)$*

Definition A.52 (Pushforward $T_{\#}\mu$)**.** *Let X, Y be measurable sets, and $T : X \to Y$ a measurable map. Then for any measure $\mu \in \mathcal{M}(X)$ there corresponds a measure $T_{\#}\mu \in \mathcal{M}(Y)$ which is the pushforward of μ by T. In particular*

$$T_{\#}\mu : Y \to \mathbb{R}_+ \quad T_{\#}\mu(B) = \mu(T^{-1}(B)) \; \forall B \subset Y, B \text{ measurable}$$

Definition A.53 (Integrability with respect to a measure)**.** *We say a function φ is integrable with respect to a measure μ whenever:*

$$\int |\varphi(x)| d\mu < \infty$$

Or say that such function is μ-integrable.

Proposition A.54 (Change of variable formula)**.** *A pushforward map and a measurable function $\varphi : Y \to \mathbb{R}$ such that $\varphi \circ T$ is μ-integrable satisfy:*

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

Proof. [Coh13](Prop. 2.6.8). □

Definition A.55 (Marginal $\pi^i_{\#}\cdot$)**.** *We denote as marginal of a measure the projection into a single dimension of the underlying space through the map:*

$$\pi^i : (x_1, \ldots) \to x_i$$

which is the pushforward $\pi^i_{\#}\mu$

Definition A.56 (Weak convergence or narrow convergence)**.** *A sequence of measures $(\mu_n) \in \mathcal{M}(\mathbb{R}^d)$ is weakly convergence to μ when:*

$$\int \varphi d\mu_n \to \int \varphi d\mu \quad \forall \varphi : \mathbb{R}^d \to \mathbb{R} \text{ bounded countinuous}$$

Definition A.57 (Bounded Lipschitz Norm)**.** *For a measure $\mu \in \mathcal{M}(\mathbb{R}^d)$ the bounded Lipschitz norm is:*

$$\|\mu\|_{BL} \coloneqq \sup \left\{ \int \varphi d\mu \; ; \; \varphi : \mathbb{R}^d \to \mathbb{R}, \; Lip(\varphi) \leq 1, \; \|\varphi\|_\infty \leq 1 \right\}$$

Where $Lip(\varphi)$ is the smallest Lipschitz constant for φ and $\|\cdot\|_\infty$ is the supremum norm.

Proposition A.58 (Equivalence of weak and bounded Lipschitz norm convergence)**.** *Consider a sequence (μ_n) such that it is bounded in total variation. Then:*

$$\text{weak convergence} \iff \lim_{n\to\infty} \|\mu_n\|_{BL} = \|\mu\|_{BL}$$

Proof. [Bog07](Sec. 8.3). □

Definition A.59 (σ-finite measure)**.** *A measure is σ-finite if a partition of its space gives finite measure to all subsets.*

$$\mu \in \mathcal{M}(\Omega) \quad \exists (A_n) \subset \Omega, \mu(A_n) < \infty \forall n, \bigcup_n A_n = \Omega$$

Definition A.60 (Singular measures, absolutely continuous measures)**.** *We briefly recall useful relations between measures. let $\mu, \nu \in \mathcal{M}(\Omega)$:*

- *singularity $\mu \perp \nu \iff \mathtt{spt}\mu \cap \mathtt{spt}\mu = \emptyset \;\; \mathtt{spt}\mu \cup \mathtt{spt} = \Omega$*
- *absolute continuity $\mu \ll \nu$ when $\nu(B) = 0 \implies \mu(B) = 0 \forall B \in \mathtt{spt}\mu$*

The latter is also used, with the additional assumption of σ-finiteness, as an hypothesis for the Radon Nykodym theorem, which states the a.e. unique existance of the function $f : \int_A d\mu = \int_A f d\nu$ for all plausible A.

Proposition A.61 (Finite measure expression)**.** *Any finite measure $\mu \in \mathcal{M}(\Omega)$ can be decomposed into a probability measure and a function integrable by the latter.*

$$\mu \quad \textit{finite} \implies \exists \sigma \in \mathcal{P}(\Omega), f \in L^1(\sigma) \quad \mu = f\sigma$$

Proof. For $|\mu|(\Omega) = 0$ any probability measure and $f \equiv 0$ satisfy the claim.
For $|\mu|(\Omega) \neq 0$ we could say:

$$\begin{aligned}
\mu(A) &= \int_A d\mu \\
&= \int_A \underbrace{|\mu|(\Omega)}_{:=f} d \left(\underbrace{\frac{d\mu}{|\mu|(\Omega)}}_{:=\sigma} \right) \\
&= \int_A f d\sigma \qquad\qquad f \in L^1(\sigma),\ \sigma \in \mathcal{P}(\Omega) \qquad\qquad \forall A
\end{aligned}$$

□

Theorem A.62 (Lebesgue decomposition)**.** *Let μ, ν be σ-finite on a measurable space. Then:*

$$\exists \mu^*, \mu^\perp \quad : \quad \nu = \mu^* + \mu^\perp$$

Where $\mu^* = f\mu \ll \mu$, $f \in L^1(\mu)$ (absolutely continuous part) and $\mu^\perp$ is singular wrt μ (independent part).

Proof. [Coh13](Thm. 4.3.2) □

A.4 Optimal Transport

Definition A.63 (p-Wasserstein metric for measures)**.** *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $\Pi(\mu, \nu) \subset \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ be the space of measures of which marginals coincide with μ on the first factor and ν on the second. Then:*

$$W_p(\mu, \nu) \coloneqq \left(\min_{\gamma \in \Pi(\mu,\nu)} \int |y - x|^p d\gamma(x, y) \right)^{\frac{1}{p}}$$

Definition A.64 (Notation for $\mathcal{P}_2(\mathbb{R}^d)$)**.** *We denote:*

$$\mathcal{P}_2(\mathbb{R}^d) \coloneqq \left(\left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \mid \int |x|^2 d\mu < \infty \right\}, \, W_2(\mu, \nu) \right)$$

Meaning the tuple of probability measures with finite second moments and the W_2 distance.

Proposition A.65 (Completeness of $\mathcal{P}_2(\mathbb{R}^d)$)**.** *The tuple $\mathcal{P}_2(\mathbb{R}^d)$ is complete in the sense of Definition A.13.*

Proof. [Vil09](Thm. 6.18), but also Section 6 in general. □

Proposition A.66 (Convergence equivalence)**.** *A sequence $(\mu_m) \in \mathcal{P}_2(\mathbb{R}^d)$ is such that:*

$$\lim_{m\to\infty} \mu_m = \mu \in \mathcal{P}_2(\mathbb{R}^d) \iff \lim_{m\to\infty} \left(\int \varphi d\mu_m \right) = \int \varphi d\mu \quad \forall \varphi : \mathbb{R}^d \to \mathbb{R} \text{ continuous, subquadratic}$$

Where by subquadratic we mean that the functions φ have at most quadratic growth.
*Note that requiring **subquadratic** growth it could still be the case that $\lim_{x\to\infty} \varphi(x) = \infty$, making it **possibly unbounded**. Then, this **result is stronger than weak convergence**.*

Proof. [AGS05](Prop. 7.1.5). □

Definition A.67 (Duality Formula for W_1)**.** *This is the definition of [San15], (Eqn. 3.1).*
For a distance cost function such as our case it holds:

$$W_1(\mu, \nu) = \min_{\gamma} \int_{\Omega\times\Omega} |y - x| d\gamma(x, y) = \max_{Lip(u)\leq 1} \int_{\Omega} u d(\mu - \nu) = \max_{Lip(f-g)\leq 1} \int_{\Omega} f(x) d\mu - \int_{\Omega} g(y) d\nu$$

Proposition A.68 (Order of norms)**.** *Using the duality formula of W_1 and Jensen's inequality:*

$$\|\mu - \nu\|_{BL} \leq W_1(\mu, \nu) \leq W_2(\mu, \nu)$$

Here by Jensen's inequality we mean:

$$\rho\left(\int x d\mu) \right) \leq \int \rho(x) f d\mu$$

For ρ a convex function. Then we basically prove in general $W_p \leq W_q$ for $p \leq q$.
In terms of duality, intuitively, it is a switch of sup and min to make W_1 comparable with the BL norm.

Lemma A.69 (Wasserstein continuity of F)**.** *Under Assumptions 1.10 the function F of Equation 1.9 is continuous for the metric $W_2(\cdot, \cdot)$. Namely:*

$$\forall \epsilon > 0 \exists \delta > 0 \quad s.t. \quad W_2(\mu, \nu) < \delta \implies |F(\mu) - F(\nu)| < \epsilon$$

This is not common for Wasserstein Gradient Flows.

Proof. The statement is equivalent to:

$$(\mu_m) \subset \mathcal{P}_2(\Omega), \mu \in \mathcal{P}_2(\Omega), \quad (\mu_m) \overset{W_2}{\to} \mu \implies F(\mu_m) \overset{|\cdot|}{\to} F(\mu) \tag{$\diamondsuit$}$$

We will prove $\diamondsuit$.
Consider the hypothesis. Then, by Assumption 1.10-3.(c) the norms $\|\Phi\|$ and $|V|$ have at most quadratic growth. Indeed we bound the sup of their differentials to be sublinear. Then:

- $\int V d\mu_m \to \int V d\mu$ (by Prop. A.66)
- By the properties of Bochner Integrals [Coh13](Prop. E5) also:

  $$\|\int \Phi d\mu_m - \int \Phi d\mu\| \leq \int \|\Phi\| \, d(\mu_m - \mu)$$

  Which implies that strongly in $\mathcal{F}$:

  $$\int \Phi d\mu_m \to \int \Phi d\mu \implies R\left(\int \Phi d\mu_m\right) \overset{|\cdot|}{\to} R\left(\int \Phi d\mu\right)$$

  since R is just a loss between functions and the arguments converge in their functional norm $\|\cdot\|$.

So that for $\mu_m \to \mu$ it holds:

$$\begin{aligned} |F(\mu_m) - F(\mu)| &= \left| R\left(\int \Phi d\mu_m\right) + \int V d\mu_m - R\left(\int \Phi d\mu\right) - \int V d\mu \right| \\ &\leq \left| R\left(\int \Phi d\mu_m\right) - R\left(\int \Phi d\mu\right) \right| + \left| \int V d\mu_m - \int V d\mu \right| \\ &\overset{|\cdot|}{\to} 0 \end{aligned}$$

which proves the claim. □

Definition A.70 (Absolutely continuous function)**.** ***(easy case)****A function $f : \mathbb{R} \to \mathbb{R}^d$ such that:*

1. *f is a.e. differentiable, meaning $\exists \frac{d}{dt} f$ $\forall t$ excluding those with measure zero*
2. *$f(t) - f(s) = \int_s^t f'(r) dr \quad \forall s < t$*

(almost general case) *for a metric space $(\mathcal{X}, d)$ a function $f : I \to \mathcal{X}$ where $I \subset \mathbb{R}$ is absolutely continuous if $\forall \epsilon > 0 \exists \delta > 0$ such that for a disjoint finite collection of subintervals $\{[t_k, t'_k]\}_{k=1}^n \subset I$ it holds:*

$$\sum_{k=1}^n |t'_k - t_k| < \delta \implies \sum_{k=1}^n d(f(t'_k), f(t_k)) < \epsilon$$

The former is rather the Lebesgue type definition of absolute continuity, a result of additional reasonings. We briefly list them for the sake of understanding the objects in place:

1. *almost general case $\implies$ bounded variation*
2. *apply Jordan decomposition (Thm. A.49) by bounded variation*
3. *Jordan decomposition holds a.s. and we split the function into two*
4. *use Lebesgue differentiation theorem to both, both are differentiable a.e.*

5. *the sum of differentiable a.e. functions is differentiable a.e.*
6. *absolutely continuous functions arise as Lebesgue integrals of those derivatives*

Also the notion requires a finite collection of intervals since it is easily recovered for an countably infinite case as well. The trick is just setting a $\frac{\epsilon}{2}$ bound on the finite sum on the LHS and sum over n.

We show next that the Wasserstein Gradient flow in our setting is just the pushforward of the intial measure *dragged* by the velocity fields. Namely, if it exists, then it has a specific relationship with the velocity.

Lemma A.71 (A classical Wgf representation)**.** *Consider the setting or Proposition 2.26. Namely:*

- *Assumptions 1.10 hold*
- *the starting measure $\mu_0 \in \mathcal{P}_2(\Omega)$ is concentrated on $Q_{r_0} \subset \Omega$ where $r_0 > 0$*

Then a Wgf $(\mu_t)_{t\geq 0}$ with velocity fields $(v_t)_{t\geq 0}$ is such that a flow of the form[29]

$$X : \mathbb{R}_+ \times \Omega \to \Omega \quad X(0,u) = u \quad \partial_t X(t,u) = v_t(X(t,u)) \quad a.e.\ t \geq 0$$

is:

- *uniquely well-defined*
- *continuous*
- *$X(t,\cdot)$ is Lipschitz on Q_r, uniformly on compact time intervals $\forall r > 0$*
- *such that the homeomorphism $\mu_t = (X_t)_{\#}\mu_0$ holds*

Proof. In [AGS05] it is shown how the velocity field v_t on Q_r satisfies a Lipschitz uniform on compact time intervals bound (Lem. 8.1.4).
The pushforward identity is a property of the continuity equation [AGS05](Prop. 8.1.8). □

A.5 Distribution Theory

The following subsection aims to explain the intution behind distributional solutions for Differential Equations. It is mostly a rewrite of [CG21](Sec. 2.7), and other sources, mentioned by the authors [WZ77; Bre11; Rud13]. It serves as an introduction to the tools needed for Subsection B.2. We denote the space of linear and continous functionals (i.e. the dual space) of $L^p(\Omega)$ as $(L^p(\Omega))'$.

Proposition A.72 (Subsequence a.s. convergence by convergence in norm)**.** *Consider $\Omega \subseteq \mathbb{R}^d$ measurable and $p \in [1,\infty]$. For $(f_m) \subset L^p(\Omega), f \in L^p(\Omega)$ it holds:*

$$f_m \overset{L^p}{\to} f \implies \exists k \to m(k) \quad f_{m(k)} \overset{a.s.}{\to} f$$

Namely, there is a subsequence which tends to f for almost every $x \in \Omega$. For $p = \infty$, the whole sequence converges.

Definition A.73 (Hölder conjugates)**.** *Two numbers $p, q \in [1,\infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$.*

[29] this is a description of how the parameters evolve according to the velocity

Proposition A.74 (Hölder's Inequality)**.** *In a measure space (S, Σ, μ) let $p, q \in [1, \infty]$ be Hölder's conjugates (Def. A.73). Then for all Borel functions f, g defined on S it holds:*

$$\|fg\|_{L^1(\mu)} \leq \|f\|_{L^p(\mu)} \|g\|_{L^q(\mu)}$$

Theorem A.75 (Riesz Representation for Dual spaces)**.** *For $\Omega \subset \mathbb{R}^d$ measurable and $p \in [1, \infty)$, $q \in (1, \infty]$ its Hölder conjugate (Def. A.73) it holds:*

$$\forall \varphi \in (L^p(\Omega))' \quad \begin{cases} \exists! u \in L^q(\Omega) \ : \ \langle \varphi, f \rangle = \int_\Omega uf \quad \forall f \in L^p(\Omega) \\ \|u\|_{L^q} = \|\varphi\|_{(L^p)'} \end{cases}$$

That is, we can identify the space of linear and continuous functionals of L^p with L^q where q is Hölder conjugate with p. Notice however that this does not hold for $p = \infty$, where we should provide further care.

We are now in the position to recover a different perspective on weak convergence in L^p.

Definition A.76 (Weak convergence in integrable functions space)**.** *We refine the notions of convergence for $\Omega \subset \mathbb{R}^d$ measurable, $p \in [1, \infty]$, q its Hölder conjugate:*

- *$p \in [1, \infty), (f_m) \subset L^p(\Omega), f \in L^p(\Omega)$:*

$$\int_\Omega f_m \varphi \to \int_\Omega f\varphi \quad \forall \varphi \in L^q(\Omega) \implies f_m \overset{w}{\to} f$$

- *$p = \infty, (f_m) \subset L^p(\Omega), f \in L^p(\Omega)$:*

$$\int_\Omega f_m \varphi \to \int_\Omega f\varphi \quad \forall \varphi \in L^1(\Omega) \implies f_m \overset{w*}{\to} f$$

Thanks to this updated notion of weak and weak* convergence, we can enlarge our perspective. Define, for $\Omega \subset \mathbb{R}^d$ open the space:

$$\mathcal{D}(\Omega) \coloneqq \{f \in C^\infty(\Omega) : \texttt{spt}\, f \textit{ compact in } \Omega\}$$

Which means infinitely smooth functions vanishing at the boundary of the domain. Such functions have a very nice property.

Proposition A.77 (Compactly supported functions are dense in good integrable spaces)**.** *For $\Omega \subset \mathbb{R}^d$ measurable and $p \in [1, \infty)$ the set $\mathcal{D}(\Omega)$ is dense in $L^p(\Omega)$. Notice that we are leaving out the special case $p = \infty$.*

Definition A.78 (Locally integrable space $L^p_{loc}(\Omega)$)**.** *For $\Omega \subseteq \mathbb{R}^d$ open , $p \in [1, \infty]$ we give a symbol to the space:*

$$L^p_{loc}(\Omega) \coloneqq \{f \in L^p(K), \forall K \subset \Omega, \ K \textit{ compact}\}$$

Namely, well-behaved functions except at the boundary of Ω.

Definition A.79 (Local norm convergence)**.** *The natural modification to L^p convergence for elements in L^p_{loc} for $(u_m) \subset L^p_{loc}(\Omega), u \in L^p_{loc}(\Omega)$ is:*

$$u_m \overset{L^p(K)}{\to} u \quad \forall K \textit{ compact} \implies u_m \overset{L^p_{loc}}{\to} u$$

Lemma A.80 (Fundamental Lemma of Calculus of variations)**.** *We have two statements:*

1. $\Omega \subseteq \mathbb{R}^d$ *open,* $f \in L^1_{loc}(\Omega)$*:*

$$\int_\Omega f\varphi = 0 \quad \forall \varphi \in \mathcal{D}(\Omega) \implies f \overset{a.s.}{=} 0$$

2. *For* $(a,b) \subset \mathbb{R}$ *and* $g, h \in C^0[a,b]$

$$\int_a^b (g(t)v(t) + h(t)\dot{v}(t))\ dt = 0 \quad \forall v \in C^1[a,b] : v(a) = 0, v(b) = 0 \implies h \in C^1[a,b], \dot{h} = g$$

The second statement is also used to derive the famous Euler-Lagrange Equation for Calculus of Variations. If it is verified, it might be easier to read it as:

$$\int_a^b \dot{h}(t)v(t)\ dt = -\int_a^b h(t)\dot{v}(t)\ dt$$

or by highlighting the integration by parts we may perform:

$$\int_a^b \dot{h}(t)v(t)\ dt + \int_a^b h(t)\dot{v}(t)\ dt = \int_a^b h(t)v(t)\ dt = 0$$

Now, we refer to **multi-index** as vectors $\alpha \in \mathbb{N}^d$ with $|\alpha| = \sum_{i=1}^d \alpha_i$. This notation allows us to define convergence $\mathcal{D}(\Omega)$.

Definition A.81 (Convergence in smooth compactly supported functions)**.** *Denote the* $\alpha \in \mathbb{N}^d$ *derivative as:*

$$D^\alpha \varphi = \frac{\partial^{|\alpha|}\varphi}{\partial x_1^{\alpha_1} \cdot \partial x_n^{\alpha_n}}$$

Then for $(\varphi_k) \subset \mathcal{D}(\Omega), \varphi \in \mathcal{D}(\Omega)$ *we say the sequence converges to* φ *and write* $\varphi_k \overset{\mathcal{D}(\Omega)}{\to} \varphi$ *if the following hold:*

- *the sequence is concentrated on a compact subset of the domain, namely* $\exists K \subset \Omega$ *compact such that* $\mathtt{spt}\,\varphi_k \subset K \forall k \in \mathbb{N}$
- *all the derivatives converge uniformly (Def. A.36), namely:*

$$D^\alpha \varphi_k \rightrightarrows D^\alpha \varphi \quad \forall \alpha \in \mathbb{N}^d$$

Given a convergence notion, we can identify the space of distributions, as linear and continuous functionals on compactly supported smooth functions.

Definition A.82 (Space of distributions $\mathcal{D}'(\Omega)$)**.** *The set of functionals which are linear and continuous and emerge from convergence of smooth differentiable functions:*

$$\mathcal{D}'(\Omega) \coloneqq \left\{ \Lambda(\varphi) : \mathcal{D}(\Omega) \to \mathbb{R}, \quad \Lambda(\varphi_k) \to \Lambda(\varphi)\ \forall (\varphi_k) : \varphi_k \overset{\mathcal{D}(\Omega)}{\to} \varphi \right\}$$

We denote the application of the functional $\Lambda \in \mathcal{D}'(\Omega)$ *to the function* $\varphi \in \mathcal{D}(\Omega)$ *as* $\langle \Lambda, \varphi \rangle \coloneqq \Lambda(\varphi)$*.*

Definition A.83 (Convergence of distributions)**.** *For a sequence of distributions (functionals)* $(u_k) \subset \mathcal{D}'(\Omega)$ *and* $u \in \mathcal{D}'(\Omega)$ *we define convergence in the sense of distributions:*

$$\langle u_k, \varphi \rangle \to \langle u, \varphi \rangle \quad \forall \varphi \in \mathcal{D}(\Omega)$$

and denote it as $u_k \overset{\mathcal{D}'(\Omega)}{\to} u$.

Definition A.84 (Distributional derivative)**.** *Let* $u \in \mathcal{D}'(\Omega)$ *and* $\alpha \in \mathbb{N}^d$*. The* α*-derivative of the distribution* $D^\alpha u$ *is a distribution defined as:*

$$\langle D^\alpha u, \varphi \rangle \coloneqq (-1)^\alpha \langle u, D^\alpha \varphi \rangle \quad \forall \varphi \in \mathcal{D}(\Omega)$$

A.5.1 Intuition for distributional derivatives

We are basically defining the derivative of something which might not be differentiable in terms of a duality with the space of functions that are smooth, differentiable, and vanishing at the boundary (on a compact support). This is inspired from integration by parts. To give a motivating example without much detail consider the sufficiently regular case[30] in which we do a first derivation:

$$\int_\Omega u' \varphi = \int_\Omega u\varphi - \int_\Omega u\varphi'$$

Where we hope for the first term to be null.
For a real valued function and a vector field, namely $u, \mathbf{V}$ we could apply the algebra of divergence

$$\texttt{div}(u\mathbf{V}) = u\nabla\mathbf{V} + \mathbf{V}\texttt{div}(u)$$

and by the divergence theorem:

$$\int_{\partial\Omega} u\mathbf{V} \cdot \vec{n} = \int_\Omega \texttt{div}(u\mathbf{V}) = \int_\Omega u\nabla\mathbf{V} + \int_\Omega \mathbf{V}\texttt{div}(u) \tag{A.85}$$

Where we recognize on the RHS the decomposition of integration by parts of the two derivatives *alternated*, and on the LHS a boundary integral. The latter is null since $\varphi \in \mathcal{D}(\Omega)$, and the support does not reach the boundary $\partial\Omega$.

B Auxiliary Results

B.1 Gradient Flow

For more references on Gradient Flows we refer to [AGS05]. This is basically a rewrite of a useful blog post [Bac20b].
Gradient descent implements a discrete update for differentiable functions of the form $F_m(\mathbf{u})$:

$$\mathbf{u}_{n+1} = \mathbf{u}_n - \epsilon\nabla F_m(\mathbf{u}_n) \quad \epsilon > 0$$

[30] Smooth boundary and $u \in C^1(\overline{\Omega}), \varphi \in \mathcal{D}(\Omega)$

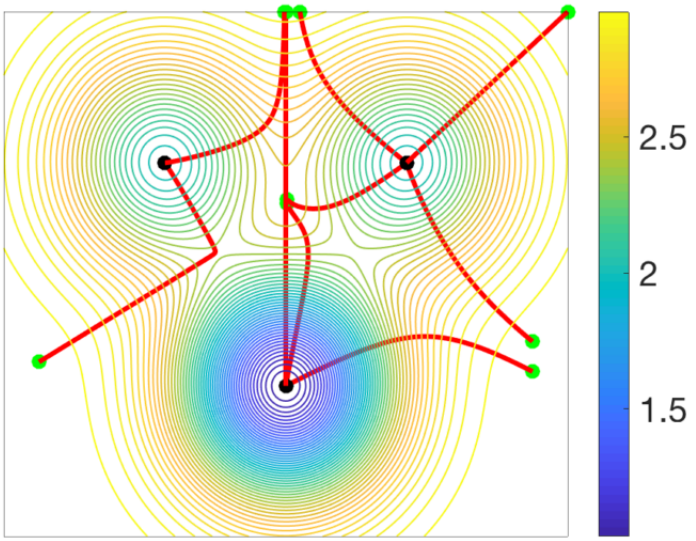


Figure 8: Gradient Flow static, Source [Bac20b]
Trajectories start at green points and end at black points. Below a GIF version of the dynamics.

We will see that this is a discretization of the dynamics on a space Ω described by:

$$\mathbf{u}'(t) = \mathcal{V}(\mathbf{u}(t)) = \begin{bmatrix} v_t(\mathbf{u}_1(t)) \\ \vdots \\ v_t(\mathbf{u}_m(t)) \end{bmatrix} = -\nabla F_m(\mathbf{u}(t))$$

Where $\mathcal{V}$ is interpreted as the velocity of evolution.
Indeed, we can interpret the update as a function:

$$X : \mathbb{R}_+ \to \Omega \quad \mathbf{u}_n = X(n\epsilon)$$

Where ϵ is the step size. By the differentiability of F_m, the expression is also seen by a Taylor expansion:

$$t = n\epsilon \quad X(t+\epsilon) = \mathbf{u}_{n+1} = \mathbf{u}_n - \epsilon \nabla F_m(\mathbf{u}_n) = X(t) - \epsilon \nabla F_m(X(t))$$

which is a piecewise affine interpolation obtained by discarding the remainder $\mathcal{O}(\epsilon)$. Which for stepsize $\epsilon \to 0$ is an ODE:

$$\lim_{\epsilon \to 0} \frac{X(t+\epsilon) - X(t)}{\epsilon} = X'(t) = -\nabla F_m(X(t))$$

Up to regularity assumptions which are verified in our case, see Proposition 2.3. Although we established an ODE for $X(t)$, it is evident that by $X(t) = \mathbf{u}_{n\epsilon}$ we are defining a gradient flow of $\mathbf{u}$ with steps in the whole $\mathbb{R}_+$. A very interesting property is that the underlying function decreases along the trajectory:

$$\frac{d}{dt} F_m(X(t)) = \nabla F_m(X(t))^T \frac{d}{dt} X(t) = \nabla F_m(X(t))^T (-\nabla F_m(X(t))) = -\left\| \nabla F_m(X(t)) \right\|_2^2 \leq 0$$

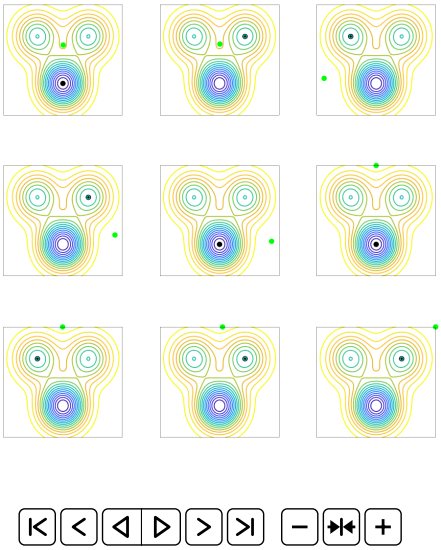


Figure 9: Animated version of Figure 8, Source [Bac20b]
Convergence to different local minima.

And if convergence holds, it is necessarily at a stationary point with $\nabla F_m(X(t)) = 0$, otherwise the dynamics would not stop. A different viewpoint is to question whether the dynamics converge or **oscillate indefinitely**. This is discussed at the proper time in the document.
Concerning Wasserstein Gradient Flows for probability measures, the construction of Definition 2.9 is much more elaborate. To grasp an intuition, a classical reference is [AGS05], while a review is [San17]. For examples of different formulations, it is also possible to consider this publication [Fan+22] and two interesting blogs discussing the topic [Rot20; Ans20].

B.2 Continuity Equation

This Section is more intuition based, for a rigorous treatment, see [Amb03](Prop. 16.3), [AGS05](Thm. 8.3.1), [AG13](Thm. 2.29), [San15](Sec. 4.2). There, the specific integrability condition is proved, as well as more comments are made on the space of test functions and the various formulations that can be done depending on the assumptions.
We inspect the dynamics of a measure $\mu \in \mathcal{M}(\Omega)$ expressed as (μ_t) being under the influence of a field (v_t).

Continuity equation the authors claim that arguments from fluid mechanics suggest that the relationship between vector field and mass satisfies the continuity equation:

$$\partial_t \mu_t = -\texttt{div}(v_t \mu_t) \quad in\ (0,\infty) \times \Omega \tag{B.1}$$

Yet we are not restricted to smooth densities in our setting, and need to resort to a distributional description of the identity.
We now recognize that the continuity equation B.1 can be interpreted in the distribution sense as per Definition A.84. In particular, a measure μ induces a distribution via Definition A.83:

$$\langle \mu, \varphi \rangle = \int \varphi d\mu \quad \varphi \in \mathcal{D}((0,\infty) \times \Omega)$$

So that we work on $(0,T)\times\Omega$ where the measure is indexed by time $\mu:(0,T)\to\mathcal{P}_2(\mathbb{R}^d)$. We are now in the position to give more context to the statement. Indeed for all compactly supported test functions φ the distributional derivatives are:

$$\langle\partial_t\mu_t,\varphi\rangle=-\langle\mu_t,\partial_t\varphi\rangle$$

Regarding the vector field, we could perform the following switch:

$$\langle\mu_t v_t,\varphi\rangle=\langle\mu_t,v_t\varphi\rangle=\left\langle\mu_t,(v_t^{(1)},\dots,v_t^{(d)})\varphi\right\rangle=\left(\left\langle\mu_t,v_t^{(1)}\varphi\right\rangle,\dots,\left\langle\mu_t,v_t^{(d)}\varphi\right\rangle\right)=\langle\mu_t,v_t\cdot\varphi\rangle$$

from which we infer for $E=\mu_t v_t$ a vector field:

$$\begin{aligned}\langle\texttt{div}(E),\varphi\rangle&=-\langle E,\nabla_u\varphi\rangle && \text{See Subsec. A.5.1}\\ &=-\langle\mu_t,v_t\nabla_u\varphi\rangle && \text{switch above}\end{aligned}$$

Therefore an adaptation of Lemma A.80 eventually gives us the definition stated by the authors [CB18]:

$$0=\langle\varphi,\partial_t\mu_t+\texttt{div}(\mu_t v_t)\rangle=-\left\langle\mu_t,\frac{d}{dt}\varphi+v\cdot\nabla_u\varphi\right\rangle\implies\int_0^\infty\int_\Omega\left(\partial_t\varphi_t(u)+\nabla_u\varphi_t(u)\cdot v_t(u)\right)d\mu_t(u)dt=0$$

Which holds for all $\varphi\in\mathcal{D}((0,T)\times\Omega)$.
Namely $\forall\varphi:(0,\infty)\times\mathbb{R}^d$ test functions over the optimization space and time such that:

- φ is smooth
- $\texttt{spt}\ \varphi$ is compact

It holds that

$$\int_0^\infty\int_{\mathbb{R}^d}\Big(\partial_t\varphi_t(u)+\nabla_u\varphi_t(u)\cdot v_t(u)\Big)d\mu_t(u)dt=0 \qquad \text{distributional version of B.1} \tag{B.2}$$

$$\int_0^{t_0}\int_{\mathbb{R}^d}|v_t(u)|d\mu(u)dt<\infty\quad\forall t_0<T \qquad \text{integrability condition holds} \tag{B.3}$$

Where the integrability condition is a requirement we do not comment on. A reference is [AGS05](Thm. 8.1.3).
The continuity equation can be seen as a **conservation of mass constraint**. The integrability condition is that guaranteeing absolute continuity in W_2. For more context, the references at the beginning of this subsection should be satisfactory.

B.3 Applying Hölder's inequality to the transport cost

We provide a result which is sufficient for the proofs of the document. In our context, the transportation cost defined in Equation 2.13 benefits from Hölder's inequality (Prop. A.74). We briefly recall the expression:

$$C_p(\gamma):=\left(\int|y-x|^p d\gamma(x,y)\right)^{\frac{1}{p}}\quad p\geq 1$$

And aim to show that in the probability space $(\Omega\times\Omega,\mathcal{B}(\Omega\times\Omega,\gamma)$ where Ω is a Euclidean space we have:

$$C_1^2(\gamma)\leq C_2^2(\gamma)$$

To do so, we simply use the fact that $p = 2$ is Hölder conjugate with itself, so that the inequality reads:

$$\int_{\Omega\times\Omega} |fg| d\gamma(x,y) = \|fg\|_{L^1(\gamma)} \leq \|f\|_{L^2(\gamma)} \|g\|_{L^2(\gamma)} = \left(\int_{\Omega\times\Omega} |f|^p d\gamma(x,y)\right)\left(\int_{\Omega\times\Omega} |g|^q d\gamma(x,y)\right)$$

which for an appropriate choice of Borel functions $f = y - x$ and $g \equiv 1$, after squaring, is equivalent to the result we look for. Indeed:

$$\begin{aligned}
C_1^2(\gamma) &= \left(\int |y-x| d\gamma(x,y)\right)^2 \\
&= \|fg\|^2_{L^1(\gamma)} && f = y - x, g \equiv 1 \\
&\leq \left(\|f\|_{L^2(\gamma)} \|g\|_{L^2(\gamma)}\right)^2 \\
&= \left(\int |f|^2 d\gamma(x,y)\right)^2 \left(\int |g|^2 d\gamma(x,y)\right)^2 \\
&= \left(\underbrace{\int |y-x|^2 d\gamma(x,y)}_{=C_2(\gamma)}\right)^2 \left(\underbrace{\int |1|^2 d\gamma(x,y)}_{=1}\right)^2 && \gamma \text{ is a probability measure} \\
&= C_2^2(\gamma)
\end{aligned}$$

B.4 Neural Networks regularity check

In the statement of Theorem 4.5 it is assumed that the boundary conditions (Ass. 3.4#3-(a)) hold. Below we briefly touch upon why this is required for a self-contained result. Consider the easiest possible functional loss with respect to the optimal Bayes regressor[31]:

$$R(f) = \|f - f^*\|^2_{\mathcal{F}}$$

Recall the notion of regular value (Def. A.41), for a candidate function:

$$\mathcal{F} \ni f = R'\left(\int \Phi d\mu\right) = \int \Phi d\mu - f^* \quad \mu \in \mathcal{P}_2(\Omega)$$

and the construction of the function of regular values (Lem. 4.4):

$$g_f(r\theta) \coloneqq \langle f, \phi(r\theta)\rangle = \int f(x)\sigma(r\theta \cdot (x,1)) d\rho_x(x) \overset{r\to\infty}{\rightrightarrows} \int_{\theta\cdot(x,1)\geq 0} f(x) d\rho_x(x) = \overline{g}_f(\theta)$$

Where the limiting function is continuously differentiable on $\mathbb{S}^{d-2}$ if:

- $\rho_x \in C_0(\mathbb{R}^{d-2})$
- f is bounded continuous (this holds by the construction of $f = \int \Phi d\mu - f^*$)

If these two hold then $g_f(r\cdot) \to \overline{g}_f$ is in C^1. The problem is that the requirement for f is not restricted. The function is not guaranteed to have $d-1$ derivatives, which require $d-1$ μ-bounded moments since they are inside the integral in $d\mu$. The measure μ only belongs to $\mathcal{P}_2(\Omega)$. For this reason, Morse Sard (Lem. A.45) is not applicable.

[31] A regressor that minimizes the probability of making a mistake. It is safe to assume it is smooth. In principle one could also assume that the true underlying function or generative model is smooth and reason equivalently. We just need f to be *nice*.