Numerical solution of Poisson partial differential equation in high dimension using deep neural networks

Dus Mathias, Virginie Ehrlacher

Numerical solution of Poisson partial differential equations in high dimension using deep neural networks

Dus Mathias, Ehrlacher Virginie

May 5, 2023

Abstract

The aim of this article is to analyze numerical schemes using two-layer neural networks with infinite width for the resolution of the high-dimensional Poisson-Neumann partial differential equations (PDEs) with Neumann boundary conditions. Using Barron's representation of the solution [1] with a measure of probability, the energy is minimized thanks to a gradient curve dynamic on the 2 Wasserstein space of parameters defining the neural network. Inspired by the work from Bach and Chizat [2, 3], we prove that if the gradient curve converges, then the represented function is the solution of the elliptic equation considered. In contrast to the works [2, 3], the activation function we use here is not assumed to be homogeneous to obtain global convergence of the flow. Numerical experiments are given to show the potential of the method.

1 Introduction

1.1 Literature review

At the origin the building unit of a neural network *ie* the perceptron, was developed by a psychiatrist F. Rosenblatt [4] who wanted to design a simple model of neural network. Computer scientists got interested in his work and developed artificial neural networks [5] but the lack of computational power in the 60s prevented serious applications. Interesting advances were made with backprop:agation [6] and the use of convolutional neural networks for image processing [7] in the late 80's but it did not focused the attention credited today. The increasing availability of computational power and the efficiency of parallel computing shed light into the true potential of neural networks with famous algorithms : deepblue surpassing Kasparov or even AlexNet identifying cats and dogs on images. Now a lot of research resources are devoted on applications of these algorithms to more and more scientific (or not) fields as statistical physics or fluid dynamics. Nevertheless, the rigorous mathematical understanding of neural networks is still lacking and the vast majority of papers are empirical proofs of concept. In this context, it is of tremendous importance to understand why neural networks work so well in certain contexts in order to improve its efficiency and get an insight of why a particular neural network should be relevant to a specific task.

The first step towards a numerical theory of neural network is the identification of functional spaces suited for neural network approximation. The first important result in this direction is the celebrated theorem of approximation due to Cybenko [8] proving that two-layer neural networks can approximate an arbitrary smooth function on a compact of $\mathbb{R}^d$. However, this work does not give an estimation of the number of neurons needed even if it is of utmost importance to hope for tractable numerical methods. To answer this question, Yarotsky [9] gave bounds on the number of neurons necessary to represent smooth functions. This theory mainly relies on classical techniques of Taylor expansions and does not give computable architectures in the high dimensional regime. Another original point of view was given by Barron [1] who used Monte Carlo techniques from Maurey-Jones-Barron to prove that functions belonging to a certain metric space *ie* the Barron space, can be approximated by a two-layer NN with precision $O\left(\frac{1}{\sqrt{m}}\right)$, m being the width of the first layer. Initially, Barron's norm was characterized using Fourier analysis reducing the theory to domain where Fourier decomposition is available. Now other Barron type norms which does not suppose the existence of an harmonic decomposition [10], are also available.

In order to give a global idea of how this works, one can say that a Barron function $f_\mu : \mathbb{R}^d \to \mathbb{R}$ can be represented by a measure μ with second order moments :

$$f_\mu(x) := \int a\sigma(wx + b)d\mu(a, b, c)$$

where σ is an activation function and the Barron norm $\|f_\mu\|_{\mathcal{B}}$ is roughly speaking the second order moments of μ. Intuitively, the law of large number says that the function f_μ can be represented by a sum of Dirac corresponding to a two-layer neural network whose width equals the number of Dirac masses. The architecture of a two-layer neural network is recalled in Figure 1. Having said that, some important questions arises :

- What is the size of the Barron space and the influence of the activation function on such size ? Some works have been done in this direction for the ReLU activation function. In [11], it is proven that H^s functions are Barron if $s \geq \frac{d}{2} + 2$ and that f_μ can be decomposed by an infinite sum of f_{μ_i} whose singularities are located on a k ($k < d$) affine subspace of $\mathbb{R}^d$. For the moment, no similar result seems to hold with more regular activation functions.
- One can add more and more layers and observe the influence on the corresponding space. In [12], tree-like spaces $\mathcal{W}_L$ (where L is the number of hidden layers) are introduced using an iterative scheme starting from the Barron space. Of course, multi-layers neural networks naturally belong to these spaces. Nevertheless for a function belonging to $\mathcal{W}_L$, it is not clear that a multilayer neural network is more efficient than its two-layer counterpart for its approximation.
- Does solutions of classical PDEs belong to a Barron space ? In this case, there is a potential to solve PDEs without suffering from the curse of dimension. Some important advances have been made in this direction in [13] where authors considered the Poisson problem with Neumann boundary conditions on the d dimensional cube. If the source term is Barron, then it is proved that the solution is also Barron and there is hope for an approximation with a two-layer NN.

Using conclusions from [13], the object of this paper is to solve Poisson equation in the high dimensional regime with Barron source. Inspired from [2], we immerse the problem on the space of probability $\mathcal{P}_2$ with finite second order moments on the domain $\mathbb{R}^2 \times S_{\mathbb{R}^d}(1) \times \mathbb{R}$. This corresponds to finding a solution to the PDE thanks to infinitely wide two-layer neural networks. Then we interpret the learning phase of the network as a gradient curve in the space of probability measure. Finally under some hypothesis on the initial support, we prove that if the curve converges then it is necessarily towards a measure corresponding to the solution of the PDE considered. Note that our argumentation is different from [2] since the convergence proof is not based on topological degree. We rather use an homotopy argument taken from algebraic topology and a clever choice of activation function to prove that the dynamic of the support of the measure curve behaves nicely on the sphere. Numerical experiments are conducted to confirm the potential of the method proposed.

In Section 1, the problem is presented in a more precise way and the link between probability and Barron function is made clearly. In Section 3, the gradient curve is introduced; its well posedness and convergence is proved. Finally, numerical experiments are exposed in Section 4.

Notation : The $|\cdot|_i$ designates the i norm of a vector of arbitrary finite dimension with particular attention to $i = 2$ (euclidean norm) for which the notation $|\cdot|$ is preferred.

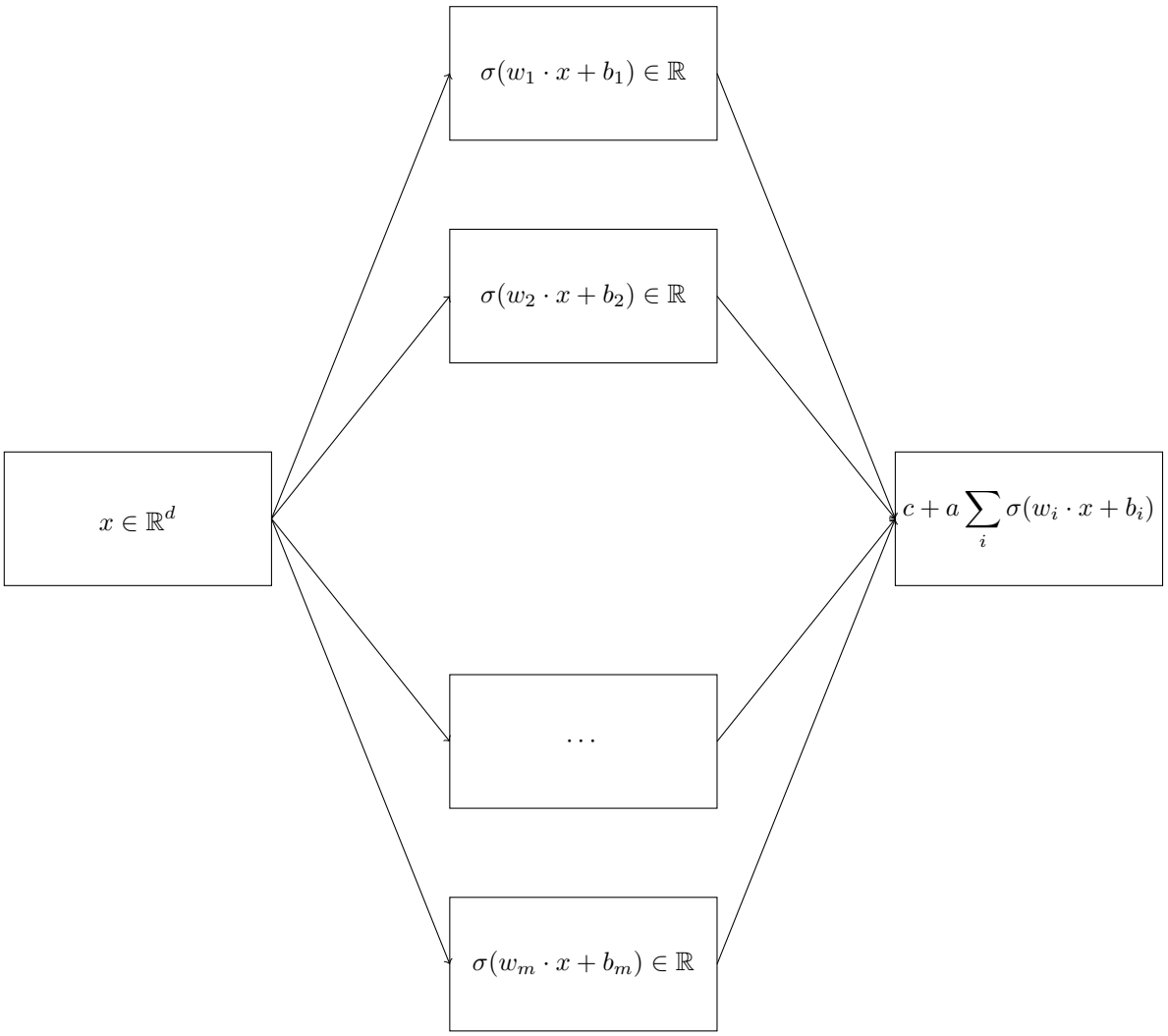


Figure 1: A two-layer neural network of width m

2 Preliminaries

This section introduces the mathematical framework to understand how neural networks are able to solve elliptic PDEs.

2.1 Problem setting

Poisson equation is considered on $\Omega := [0,1]^d$ ($d \in \mathbb{N}$) with Neumann boundary conditions: find $u^* \in H^1(\Omega)$ with $\int_\Omega u^* = 0$ solution to :

$$\begin{cases} -\Delta u^\star = f \text{ on } \Omega, \\ \partial_n u^\star = 0 \text{ on } \partial\Omega, \end{cases} \tag{1}$$

where $f \in L^2(\Omega)$ with $\int_\Omega f = 0$. Here (1) has to be understood in the variational sense, in the sense that u^* is equivalently the unique minimizer to:

$$u^\star = \underset{u \in H^1(\Omega)}{\operatorname{argmin}}\, \mathcal{E}(u), \tag{2}$$

where

$$\forall u \in H^1(\Omega), \quad \mathcal{E}(u) := \int_\Omega \left(\frac{|\nabla u|^2}{2} - fu \right) dx + \frac{1}{2} \Big(\int_\Omega u dx \Big)^2.$$

This can indeed be easily checked by classic Lax-Milgram arguments. The functional $\mathcal{E}$ is strongly convex and differentiable with derivative given by Lemma 1.

Lemma 1. *The functional $\mathcal{E} : H^1(\Omega) \to \mathbb{R}$ is continuous, differentiable and for all $u \in H^1(\Omega)$, it holds that*

$$\forall v \in H^1(\Omega), \ d\mathcal{E}\,|_u(v) = \int_\Omega (\nabla u \cdot \nabla v - fv)\, dx + \int_\Omega u dx \int_\Omega v dx.$$

It can be easily seen that points u where the differential is identically zero are solution to equation (1).

Remark 1. *The coercive symmetric bilinear form a involved in the definition of the energy writes :*

$$a(u, v) := \int_\Omega \nabla u \cdot \nabla v dx + \int_\Omega u dx \int_\Omega v dx.$$

The energy $\mathcal{E}$ can then be equivalently rewritten thanks to the bilinear form a :

$$\mathcal{E}(u) = \frac{1}{2} a(u - u^\star, u - u^\star) - \frac{1}{2} \int_\Omega |\nabla u^\star|^2 dx.$$

The aim of the present work is to analyze a numerical method based on the use of infinite-width two-layer neural networks for the resolution of (1) with a specific focus on the case when d is large.

2.2 Activation function

We introduce here the particular choice of activation function we consider in this work.

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the classical Rectified Linear Unit (ReLU) function where :

$$\forall y \in \mathbb{R}, \ \sigma(y) := \max(y, 0). \tag{3}$$

Let $\rho : \mathbb{R} \to \mathbb{R}$ be defined by

$$\begin{cases} Z \exp\left(-\dfrac{\tan(\frac{\pi}{2} y)^2}{2} \right) & \text{if } |y| \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where the constant $Z \in \mathbb{R}$ is defined such that the integral of ρ is equal to one. For all $\tau > 0$, we then define $\rho_\tau := \tau \rho(\tau \cdot)$ and $\sigma_\tau : \mathbb{R} \to \mathbb{R}$ the function defined by

$$\forall y \in \mathbb{R}, \ \sigma_\tau(y) := (\rho_\tau \star \sigma)(y). \tag{5}$$

We then have the following lemma.

Lemma 2. *For any $\tau > 0$, it holds that*

(i) $\sigma_\tau \in \mathcal{C}^\infty(\mathbb{R})$ is uniformly bounded as σ_τ',

(ii) for all $y < -1/\tau$, $\sigma_\tau(y) = 0$,

(iii) for all $y > 1/\tau$, $\sigma_\tau(y) = y$,

(iv) there exists $C > 0$ such that for all $\tau > 0$,

$$\|\sigma - \sigma_\tau\|_{H^1(\mathbb{R})} \leq \frac{C}{\sqrt{\tau}}.$$

Proof. The first item (i) is classic and left to the reader. For (ii), we have :

$$\sigma_\tau(y) = \int_{-1/\tau}^{1/\tau} \rho_\tau(y)\sigma(x-y)dy \tag{6}$$

and if $x < -1/\tau$ then $x - y < 0$ for $-1/\tau < y < 1/\tau$ and $\sigma(x-y) = 0$. This naturally gives :

$$\sigma_\tau(y) = 0.$$

For (iii), using again (6) and if $x > 1/\tau$, then $x - y > 0$ for $-1/\tau < y < 1/\tau$ and $\sigma(x-y) = x - y$. As a consequence,

$$\begin{aligned}\sigma_\tau(y) &= \int_{-1/\tau}^{1/\tau} \rho_\tau(y)(x-y)dy \\ &= x.\end{aligned}$$

where we have used the fact that $\int_{\mathbb{R}} \rho_\tau(y)dy = 1$ and $\int_{\mathbb{R}} y\rho_\tau(y)dy = 0$ by symmetry of ρ.

For (iv), we have by $(ii) - (iii)$:

$$\|\sigma - \sigma_\tau\|^2_{L^2(\mathbb{R})} = \int_{-1/\tau}^{1/\tau} (\sigma(x) - \sigma_\tau(x))^2 dx \leq \frac{8}{\tau^2}$$

where we used the fact that $|\sigma(x)|, |\sigma_\tau(x)| \leq 1/\tau$ on $[-1/\tau, 1/\tau]$. In a similar way,

$$\|\sigma' - \sigma'_\tau\|^2_{L^2(\mathbb{R})} = \int_{-1/\tau}^{1/\tau} (\sigma'(x) - \sigma'_\tau(x))^2 dx \leq \frac{1}{\tau}.$$

The two last inequalities gives (iv).

□

In this work, we will rather use a hat version of the regularized ReLU activation function. More precisely, we define:

$$\forall y \in \mathbb{R},\ \sigma_{H,\tau}(y) := \sigma_\tau(y+1) - \sigma_\tau(2y) + \sigma_\tau(y-1). \tag{7}$$

We call hereafer this activation function the regularized HReLU (Hat ReLU) activation. When $\tau = +\infty$, the following notation is proposed :

$$\forall y \in \mathbb{R},\ \sigma_H(y) := \sigma(y+1) - \sigma(2y) + \sigma(y-1). \tag{8}$$

The reasons why we use this activation is that it has a compact support and can be used to generate an arbitrary piecewise constant function on $[0, 1]$.

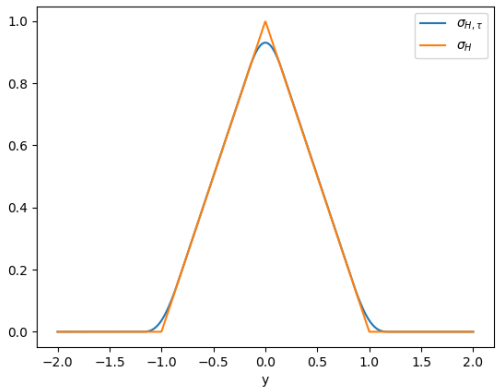


Figure 2: The hat activation function and its regularization ($\tau = 4$)

2.3 Spectral Barron space

The orthonormal basis in $L^2(\Omega)$ composed of the eigenfunctions $\{\phi_k\}_{k\in\mathbb{N}^d}$ of the Laplacian operator with Neumann boundary conditions, where

$$\forall k = (k_1, \dots, k_d) \in \mathbb{N}^d,\ \forall x := (x_1, \cdots, x_d) \in \Omega, \quad \phi_k(x_1, \dots, x_d) := \prod_{i=1}^{d} \cos(\pi k_i x_i). \tag{9}$$

Notice that $\{\phi_k\}_{k\in\mathbb{N}^d}$ is also an orthogonal basis of $H^1(\Omega)$. Using this basis, we have the Fourier representation formula for any function $u \in L^2(\Omega)$:

$$u = \sum_{k\in\mathbb{N}^d} \hat{u}(k)\phi_k,$$

where for all $k \in \mathbb{N}^d$, $\hat{u}(k) := \langle \phi_k, u\rangle_{L^2(\Omega)}$. This allows to define the (spectral) Barron space [13] as follows :

Definition 1. *For all $s > 0$, the Barron space $\mathcal{B}^s(\Omega)$ is defined as : [] []*

$$\mathcal{B}^s(\Omega) := \Big\{ u \in L^1(\Omega) : \sum_{k\in\mathbb{N}^d} (1 + \pi^s |k|_1^s)|\hat{u}(k)| < +\infty \Big\} \tag{10}$$

and the space $\mathcal{B}^2(\Omega)$ is denoted $\mathcal{B}(\Omega)$. Moreover, the space $\mathcal{B}^s(\Omega)$ is embedded with the norm :

$$\|u\|_{\mathcal{B}^s(\Omega)} := \sum_{k\in\mathbb{N}^d} (1 + \pi^s |k|_1^s)|\hat{u}(k)|. \tag{11}$$

[]

By [13, Lemma 4.3], it is possible to relate the Barron space to traditional Sobolev spaces :

Lemma 3. *The following continuous injections hold :*

- $\mathcal{B}(\Omega) \hookrightarrow H^1(\Omega)$,
- $\mathcal{B}^0(\Omega) \hookrightarrow L^\infty(\Omega)$.

The space $\mathcal{B}(\Omega)$ has interesting approximation properties related to neural networks schemes. We introduce the following approximation space:

Definition 2. *Let $\chi : \mathbb{R} \to \mathbb{R}$ be measurable, $m \in \mathbb{N}^*$ and $B > 0$. The space $\mathcal{F}_{\chi,m}(B)$ is defined as :*

$$\mathcal{F}_{\chi,m}(B) := \Big\{ c + \sum_{i=1}^{m} a_i \chi(w_i \cdot x + b_i) : c, a_i, b_i \in \mathbb{R},\ w_i \in \mathbb{R}^d,\ |c| \le 2B, |w_i| = 1, |b_i| \le 1, \sum_{i=1}^{m} |a_i| \le 4B \Big\} \tag{12}$$

Now, we are able to state the main approximation theorem.

Theorem 1. *For any $u \in \mathcal{B}(\Omega)$, $m \in \mathbb{N}^*$:*

(i) there exists $u_m \in \mathcal{F}_{\sigma_H,m}(\|u\|_{\mathcal{B}(\Omega)})$ such that :

$$\|u - u_m\|_{H^1(\Omega)} \le \frac{C\|u\|_{\mathcal{B}(\Omega)}}{\sqrt{m}},$$

(ii) there exists $\tilde{u}_m \in \mathcal{F}_{\sigma_{H,m},m}(\|u\|_{\mathcal{B}(\Omega)})$ such that :

$$\|u - \tilde{u}_m\|_{H^1(\Omega)} \le \frac{C\|u\|_{\mathcal{B}(\Omega)}}{\sqrt{m}}. \tag{13}$$

where C is a universal constant which does not depend on d neither on u. []

Proof. Let $B := \|u\|_{\mathcal{B}(\Omega)}$. We just give a sketch of the proof of (ii), (i) being derived from similar arguments as in [13, Theorem 2.1] []

By (i), there exists $u_m \in \mathcal{F}_{\sigma_H,m}(B)$ such that

$$\|u - u_m\|_{H^1(\Omega)} \leq \frac{CB}{\sqrt{m}}.$$

The function u_m can be written as :

$$u_m(x) = c + \sum_{i=1}^{m} a_i \sigma_H(w_i \cdot x + b_i)$$

for some $c, a_i, b_i \in \mathbb{R}$, $w_i \in \mathbb{R}^d$ for $i = 1, \ldots, m$ with $|c| \leq 2B, |w_i| = 1, |b_i| \leq 1, \sum_{i=1}^{m} |a_i| \leq 4B$.

By Lemma 2 (iv), there exists $C > 0$ such that for all $\tau > 0$ $\|\sigma_H - \sigma_{H,\tau}\|_{H^1(\mathbb{R})} \leq \frac{C}{\sqrt{\tau}}$, it is easy to see that

$$\|\tilde{u}_m - u_m\|_{H^1(\Omega)} \leq \frac{CB}{\sqrt{m}}$$

where

$$\tilde{u}_m(x) = c + \sum_{i=1}^{m} a_i \sigma_{H,m}(w_i \cdot x + b_i).$$

Consequently,

$$\|u - \tilde{u}_m\|_{H^1(\Omega)} \leq \frac{CB}{\sqrt{m}}$$

which yields the desired result. □

Remark 2. *With other words, a barron function can be approximated in $H^1(\Omega)$ by a two-layer neural network of width m with precision $O\left(\frac{1}{\sqrt{m}}\right)$ when the activation function is the HReLU one.*

In the sequel, for all $r > 0$, we denote by $K_r := [-2r, 2r] \times [-4r, 4r] \times S_{\mathbb{R}^d}(1) \times [-1, 1]$.

Remark 3. *Let $m \in \mathbb{N}^*$, $u_m \in \mathcal{F}_{\chi,m}(B)$ with $B > 0$ and $\chi : \mathbb{R} \to \mathbb{R}$. Then, there exists $c, a_i, b_i \in \mathbb{R}$, $w_i \in \mathbb{R}^d$ for $i = 1, \ldots, m$ with $|c| \leq 2B, |w_i| = 1, |b_i| \leq 1, \sum_{i=1}^{m} |a_i| \leq 4B$ such that for all $x \in \Omega$,*

$$\begin{aligned} u_m(x) &= c + \sum_{i=1}^{m} a_i \chi(w_i \cdot x + b_i) \\ &= \sum_{i=1}^{m} \left(c + \sum_{j=1}^{m} |a_j| sign(a_i) \chi(w_i \cdot x + b_i) \right) \frac{|a_i|}{\sum_{j=1}^{m} |a_j|} \\ &= \int_{\Theta} [c + a\chi(w \cdot x + b)] d\mu_m(c, a, w, b), \end{aligned}$$

where

- *the space of parameters Θ is defined by $\Theta := \mathbb{R} \times \mathbb{R} \times S_{\mathbb{R}^d}(1) \times \mathbb{R}$ with $S_{\mathbb{R}^d}(1)$ the unit sphere of $\mathbb{R}^d$;*
- *the measure μ_m is a probability measure on Θ given by :*

$$\mu_m := \sum_{i=1}^{m} \frac{|a_i|}{\sum_{j=1}^{m} |a_j|} \delta_{(c, \sum_{j=1}^{m} |a_j| sign(a_i), w_i, b_i)}.$$

Remark that μ_m has support in K_B. Moreover, let $\mathcal{P}_2(\Theta)$ denote the set of probability measures on Θ with finite second-order moments. Then, it holds that the sequence $(\mu_m)_{m\in\mathbb{N}^*}$ is uniformly (wrt m) bounded in $\mathcal{P}_2(\Theta)$.

For a general domain Ω which is not of the form $\Omega = [0,1]^d$, the solution to equation (1) does not necessarily belong to the Barron space even if the source term has finite Barron norm. Nevertheless for our case ($\Omega = [0,1]^d$), there is an explicit bound of the Barron norm of the solution compared with the source one. This gives hope for a neural network approximation of the solution.

Theorem 2. *[13] Let u^* be the solution of the equation (1) with $f \in \mathcal{B}^0(\Omega)$, then $u^* \in \mathcal{B}(\Omega)$. Moreover, the following estimate holds :*

$$\|u^*\|_{\mathcal{B}(\Omega)} \leq d\|f\|_{\mathcal{B}^0(\Omega)}.$$

2.4 Infinite width two-layer neural networks

In order to ease the notation for future computations, for all $\tau > 0$, we introduce the function $\Phi_\tau : \Theta \times \Omega \to \mathbb{R}$ defined by

$$\forall \theta := (c, a, w, b) \in \Theta,\ \forall x \in \Omega, \quad \Phi_\tau(\theta; x) := c + a\sigma_{H,\tau}(w \cdot x + b) \tag{14}$$

and $\Phi_\infty : \Theta \times \Omega \to \mathbb{R}$ defined by such that:

$$\forall \theta := (c, a, w, b) \in \Theta,\ \forall x \in \Omega, \quad \Phi_\infty(\theta; x) := c + a\sigma_H(w \cdot x + b). \tag{15}$$

The space $\mathcal{P}_2(\Theta)$ is embedded with the 2-Wasserstein distance :

$$\forall \mu, \nu \in \mathcal{P}_2(\Theta), \quad W_2^2(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Theta^2} d(\theta, \tilde{\theta})^2 d\gamma(\theta, \tilde{\theta}),$$

where $\Gamma(\mu, \nu)$ is the set of probability measures on Θ^2 with marginals given respectively by μ and ν and where d is the geodesic distance in Θ. For the interested reader, the geodesic distance between $\theta, \tilde{\theta} \in \Theta$ can be computed as :

$$d(\theta, \tilde{\theta}) = \sqrt{(c - \tilde{c})^2 + (a - \tilde{a})^2 + d_{S_{\mathbb{R}^d}(1)}(w, \tilde{w}) + (b - \tilde{b})^2}.$$

For all $\tau, r > 0$, we introduce the operator P_τ and the functional $\mathcal{E}_{\tau,r}$ defined as follows:

Definition 3. *The operator $P_\tau : \mathcal{P}_2(\Theta) \to H^1(\Omega)$ is defined for all $\mu \in \mathcal{P}_2(\Theta)$ as :*

$$P_\tau(\mu) := \int_\Theta \Phi_\tau(\theta; x) d\mu(\theta).$$

Additionally, we define the functional $\mathcal{E}_{\tau,r}(\mu) : \mathcal{P}_2(\Theta) \to \mathbb{R}$ as :

$$\mathcal{E}_{\tau,r}(\mu) := \begin{cases} \mathcal{E}(P_\tau(\mu)) \text{ if } \mu(K_r) = 1 \\ +\infty \text{ otherwise.} \end{cases}$$

.

Proposition 1. *For all $0 < \tau, r < \infty$, the functional $\mathcal{E}_{\tau,r}$ is weakly lower semicontinuous.*

Proof. Let $(\mu_n)_{n\in\mathbb{N}^*}$ be a sequence of elements of $\mathcal{P}_2(\Theta)$ which narrowly converges towards some $\mu \in \mathcal{P}_2(\Theta)$. Without loss of generality, we can assume that μ_n is supported in K_r for all $n \in \mathbb{N}^*$. Then, it holds that :

- the limit μ has support in K_r (by Portmanteau theorem);
- moreover, let $u_n : \Omega \to \mathbb{R}$ be defined such that for all $x \in \Omega$,

$$u_n(x) := \int_\Theta \Phi_\tau(\theta; x) d\mu_n(\theta) = \int_{K_r} \Phi_\tau(\theta; x)\, d\mu_n(\theta).$$

Since for all $x \in \Omega$, the function $K_r \ni \theta \mapsto \Phi_\tau(\theta; x)$ is continuous and bounded, it then holds that, for all $x \in \Omega$,

$$u_n(x) \underset{n\to\infty}{\longrightarrow} u(x) := \int_{K_r} \Phi_\tau(\theta; x) d\mu(\theta) = \int_\Theta \Phi_\tau(\theta; x) d\mu(\theta),$$

where the last equality comes from the fact that μ is supported in K_r.

- It actually holds that the sequence $(u_n)_{n\in\mathbb{N}^*}$ is uniformly bounded in $\mathcal{C}(\Omega)$. Indeed, there exists $C_\tau > 0$ such that for all $x \in \Omega$ and $n \in \mathbb{N}^*$, we have

$$\begin{aligned} u_n(x)^2 &= \left(\int_{K_r} \Phi_\tau(\theta; x) d\mu_n(\theta) \right)^2 \\ &\leq \int_{K_r} \Phi_\tau^2(\theta; x) d\mu_n(\theta) \\ &\leq C r^4 \end{aligned}$$

where last inequality comes from the fact that Φ_τ is at most quadratic in the variable θ. As a consequence of the Lebesgue dominated convergence theorem, the sequence $(u_n)_{n\in\mathbb{N}^*}$ strongly converges towards u in $L^2(\Omega)$. Reproducing the same argument as above for the sequence $(\nabla u_n)_{n\in\mathbb{N}^*}$, one easily proves that this strong convergence holds in fact in $H^1(\Omega)$. The fact that the functional $\mathcal{E} : H^1(\Omega) \to \mathbb{R}$ is continuous allows us to conclude. □

Remark 4. *In $\mathcal{P}_2(\Theta)$, the weak convergence is metricized by the Wasserstein distance. Hence, $\mathcal{E}_\tau$ is lower semicontinuous as a functional from $(\mathcal{P}_2(\Theta), W_2)$ to $(\mathbb{R}, |\cdot|)$.*

Finally, the lower semicontinuity of $\mathcal{E}_{\tau,r}$ and the weak compactness of K_r allows to prove the existence of at least one solution to the following minimization problem :

Problem 1. *For $0 < \tau < \infty$ and $0 < r < +\infty$, let $\mu^\star_{\tau,r} \in \mathcal{P}_2(\Theta)$ be solution to*

$$\mu^\star_{\tau,r} \in \underset{\mu\in\mathcal{P}_2(\Theta)}{\operatorname{argmin}} \mathcal{E}_{\tau,r}(\mu). \tag{16}$$

For large values of τ and $r = d\|f\|_{\mathcal{B}^0(\Omega)}$, solutions of (16) enable to obtain accurate approximations of the solution of (1). This result is stated in Theorem 3.

Theorem 3. *There exists $C > 0$ such that for all $m \in \mathbb{N}^*$ and any solution $\mu^\star_{m,d\|f\|_{\mathcal{B}^0(\Omega)}}$ to (16) with $\tau = m$ and $r = d\|f\|_{\mathcal{B}^0(\Omega)}$, it holds that:*

$$\left\| u^\star - \int_\Theta \Phi_m(\theta; x) d\mu^\star_{m,d\|f\|_{\mathcal{B}^0(\Omega)}}(\theta) \right\|_{H^1(\Omega)} \leq C d \frac{\|f\|_{\mathcal{B}^0(\Omega)}}{\sqrt{m}}$$

where $u^\star$ is the solution of the equation (1).

Proof. For all $m \in \mathbb{N}^*$, let $\tilde{u}_m \in \mathcal{F}_{\sigma_{H,m},m}(\|u^*\|_{\mathcal{B}})$ satisfying (13) for $u = u^*$ (using Theorem 1). Since $\|u^*\|_{\mathcal{B}(\Omega)} \leq d\|f\|_{\mathcal{B}^0(\Omega)}$ thanks to Theorem 2 and by Remark 3, $\tilde{u}_m$ can be rewritten using a probability measure μ_m with support in $K_{d\|f\|_{\mathcal{B}^0(\Omega)}}$ as :

$$\forall x \in \Omega, \quad \tilde{u}_m(x) = \int_\Theta \Phi_m(\theta; x)\, d\mu_m(\theta).$$

Let $\mu^\star_{m,d\|f\|_{\mathcal{B}^0(\Omega)}}$ be a minimizer of (16) with $\tau = m$ and $r = d\|f\|_{\mathcal{B}^0(\Omega)}$. Then, it holds that:

$$\mathcal{E}_{m,d\|f\|_{\mathcal{B}^0(\Omega)}} \left(\mu^\star_{m,d\|f\|_{\mathcal{B}^0(\Omega)}} \right) \leq \mathcal{E}_{m,d\|f\|_{\mathcal{B}^0(\Omega)}}(\mu_m),$$

which by Remark 1, is equivalent to :

$$a(u^\star_m - u^\star, u^\star_m - u^\star) \leq a(\tilde{u}_m - u^\star, \tilde{u}_m - u^\star).$$

where for all $x \in \Omega$,

$$u_m^\star(x) := \int_\Theta \Phi_m(\theta; x)\, d\mu^\star_{m, d\|f\|_{\mathcal{B}^0(\Omega)}}(\theta).$$

Denoting by α and L respectively the coercivity and continuity constants of a, we obtain that

$$\|u_m^\star - u^\star\|_{H^1(\Omega)} \leq \frac{L}{\alpha}\|\tilde{u}_m - u^\star\|_{H^1(\Omega)} \leq Cd\frac{\|f\|_{\mathcal{B}^0(\Omega)}}{\sqrt{m}}.$$

□

3 Gradient curve

In this section, we solve Problem 1 using gradient curve techniques. More particularly, we will define and prove the existence of a gradient descent curve such that if the convergence is asserted, then it is necessarily towards a global minimizer.

3.1 Well-posedness

3.1.1 Well-posedness with a condition on the support

Let us fix some values of $r, \tau > 0$ in this section.

Let Γ be the set of constant speed geodesics of Θ *ie* the set of curves $\pi : [0,1] \to \Theta$ of geodesics. For all $s \in [0,1]$, we define the application map $e_s : \Gamma \to \Theta$ such that $e_s(\pi) := \pi(s)$. Owing this, McCann interpolation gives the fundamental characterization of constant speed geodesics in $\mathcal{P}_2(\Theta)$:

Proposition 2. *[14, Proposition 2.10] For all $\mu, \nu \in \mathcal{P}_2(\Theta)$ and any minimal geodesic $(\mu_t)_{t\in[0,1]}$ between them, there exists $\Pi \in \mathcal{P}_2(\Gamma)$ such that :*

$$\forall t \in [0,1],\ \mu_t = e_t \# \Pi.$$

Remark 5. *As $e_0\#\Pi = \mu$ and $e_1\#\Pi = \nu$, the support of Π is included in the set of geodesics $\pi : [0,1] \to \Theta$ such that $\pi(0)$ belongs to the support of μ and $\pi(1)$ belongs to the support of ν.*

The next result states smoothness properties of geodesics on Θ which are direct consequences of the smoothness properties of geodesics on the unit sphere of $\mathbb{R}^d$. It is a classical result and its proof is left to the reader.

Lemma 4. *There exists $C > 0$ such that for all $(\theta, \tilde{\theta})$ in Θ^2, all geodesic $\pi : [0,1] \to \Theta$ such that $\pi(0) = \theta$ and $\pi(1) = \widetilde{\theta}$ and all $0 \leq s \leq t \leq 1$,*

$$|\pi(t) - \pi(s)| \leq d(\pi(t), \pi(s)) = (t-s)d(\theta, \tilde{\theta}) \leq C(t-s)|\tilde{\theta} - \theta|$$

and

$$\left|\frac{d}{dt}\pi(t)\right| \leq d(\theta, \tilde{\theta}) \leq C|\tilde{\theta} - \theta|.$$

In order to prove the well-posedness, it is necessary to get information about the smoothness of $\mathcal{E}_{\tau,r}$.

Proposition 3. *The functional $\mathcal{E}_{\tau,r}$ is proper, coercive, differentiable on its domain $D(\mathcal{E}_{\tau,r})$. Moreover, for all $\mu, \nu \in D(\mathcal{E}_{\tau,r})$, $\gamma \in \Gamma(\mu, \nu)$:*

$$\mathcal{E}_{\tau,r}(\nu) = \mathcal{E}_{\tau,r}(\mu) + \int_{K_r^2} v_\mu^\tau(\theta) \cdot (\tilde{\theta} - \theta) d\gamma(\theta, \tilde{\theta}) + O(C_2^2(\gamma)) \tag{17}$$

with

$$v_\mu^\tau(\theta) := \nabla_\theta \phi_\mu^\tau(\theta)\ \mu\text{-almost everywhere} \tag{18}$$

where for all $\theta \in K_r$,

$$\begin{aligned}\phi_\mu^\tau(\theta) &:= \langle \nabla P_\tau(\mu), \nabla \Phi_\tau(\theta; \cdot)\rangle_{L^2(\Omega)} - \langle f, \Phi_\tau(\theta; \cdot)\rangle_{L^2(\Omega)} + \int_\Omega P_\tau(\mu)(x)dx \times \int_\Omega \Phi_\tau(\theta; x)dx \\ &= d\,\mathcal{E}\,|_{P_\tau(\mu)}(\Phi_\tau(\theta; \cdot)).\end{aligned} \tag{19}$$

The properness and coercivity are easy to prove and left to the reader.

Proof. First, we focus on the proof of (17)-(19). As Φ_τ and $\mathcal{E}$ are smooth, it holds that for all $x \in \Omega$, $\theta, \widetilde{\theta} \in \Theta$, $u, \tilde{u} \in H^1(\Omega)$,

$$\begin{cases} \Phi_\tau(\tilde{\theta}; x) = \Phi_\tau(\theta; x) + \nabla_\theta \Phi_\tau(\theta; x) \cdot (\tilde{\theta} - \theta) + M_\tau(\theta, \tilde{\theta}; x) \\ \quad \mathcal{E}(\tilde{u}) = \mathcal{E}(u) + d\,\mathcal{E}\,|_u(\tilde{u} - u) + N(\tilde{u} - u), \end{cases}$$

where $N(u) := \mathcal{E}(u)$ for all $u \in H^2(\Omega)$ and $M_\tau(\theta, \tilde{\theta}; x) := \displaystyle\int_0^1 (\tilde{\theta}-\theta)^T \nabla^2_\theta \Phi_\tau(\theta + t(\tilde{\theta}-\theta); x)(\tilde{\theta}-\theta)(1-t)dt$. More precisely,

$$\begin{cases} \dfrac{\partial \Phi_\tau(\theta; x)}{\partial c} = 1 \\ \dfrac{\partial \Phi_\tau(\theta; x)}{\partial a} = \sigma_\tau(w \cdot x + b) \\ \dfrac{\partial \Phi_\tau(\theta; x)}{\partial w} = a x \sigma'_\tau(w \cdot x + b) \\ \dfrac{\partial \Phi_\tau(\theta; x)}{\partial b} = a \sigma'_\tau(w \cdot x + b) \end{cases} \tag{20}$$

and the non zero terms of the hessian matrix write :

$$\begin{cases} \dfrac{\partial^2 \Phi_\tau(\theta; x)}{\partial a \partial w} = \sigma'_\tau(w \cdot x + b) x \\ \dfrac{\partial^2 \Phi_\tau(\theta; x)}{\partial a \partial b} = \sigma'_\tau(w \cdot x + b) \\ \dfrac{\partial^2 \Phi_\tau(\theta; x)}{\partial^2 w} = a \sigma''_\tau(w \cdot x + b) x x^T \\ \dfrac{\partial^2 \Phi_\tau(\theta; x)}{\partial w \partial b} = a \sigma''_\tau(w \cdot x + b) x \\ \dfrac{\partial^2 \Phi_\tau(\theta; x)}{\partial^2 b} = a \sigma''_\tau(w \cdot x + b). \end{cases} \tag{21}$$

As $|\sigma'_\tau| \leq 1, |\sigma''_\tau| \leq \tau, |\sigma'''_\tau| \leq \tau^2$, it then holds that :

- $\forall x \in \Omega, \theta \in K_r,\ |M_\tau(\theta, \tilde{\theta}; x)| \leq C r \tau |\theta - \tilde{\theta}|^2$ (Ω is bounded),
- $\forall x \in \Omega, \theta \in K_r,\ |\nabla M_\tau(\theta, \tilde{\theta}; x)| \leq C r^2 \tau^2 |\theta - \tilde{\theta}|^2$ (Ω is bounded).

Moreover, there exists a constant $C > 0$ such that for all $u \in H^1(\Omega)$,

$$|N(u)| = |\,\mathcal{E}(u)| \leq C \|u\|^2_{H^1(\Omega)}. \tag{22}$$

Thus for $\mu, \nu \in D(\mathcal{E}_{\tau,r})$ then γ is supported in K_r^2 and :

$$\begin{aligned} \mathcal{E}_{\tau,r}(\nu) &= \mathcal{E}\Big(\int_{K_r} \Phi_\tau(\tilde{\theta}; \cdot) d\nu(\tilde{\theta})\Big) \\ &= \mathcal{E}\Big(\int_{K_r^2} \Phi_\tau(\tilde{\theta}; \cdot) d\gamma(\theta, \tilde{\theta})\Big) \\ &= \mathcal{E}\Big(\int_{K_r^2} \Phi_\tau(\theta; \cdot) + \nabla_\theta \Phi_\tau(\theta; \cdot) \cdot (\tilde{\theta} - \theta) + M_\tau(\theta, \tilde{\theta}; \cdot) d\gamma(\theta, \tilde{\theta})\Big) \\ &= \mathcal{E}_{\tau,r}(\mu) + d\,\mathcal{E}\,|_{P_\tau(\mu)}\Big(\int_{K_r^2} \nabla_\theta \Phi_\tau(\theta; \cdot) \cdot (\tilde{\theta} - \theta) d\gamma\Big) \\ &+ N\Big(\int_{K_r^2} M_\tau(\theta, \tilde{\theta}; \cdot) d\gamma\Big). \end{aligned}$$

The remainder term is of order two since :

$$\begin{aligned}\Big\| \int_{K_r^2} M_\tau(\theta,\tilde\theta;\cdot)d\gamma\Big\|^2_{H^1(\Omega)} &= \Big\| \int_{K_r^2} M_\tau(\theta,\tilde\theta;\cdot)d\gamma\Big\|^2_{L^2(\Omega)} + \Big\| \int_{K_r^2} \nabla M_\tau(\theta,\tilde\theta;\cdot)d\gamma\Big\|^2_{L^2(\Omega)} \\ &\le \int_{K_r^2} \|M_\tau(\theta,\tilde\theta;\cdot)\|^2_{L^2(\Omega)}d\gamma + \int_{K_r^2} \|\nabla M_\tau(\theta,\tilde\theta;\cdot)\|^2_{L^2(\Omega)}d\gamma \\ &\le C(r\tau + r^2\tau^2)\int_{\Theta^2} |\tilde\theta-\theta|^2 d\gamma \\ &\le C(r\tau + r^2\tau^2)C_2^2(\gamma),\end{aligned}$$

where $C_2^2(\gamma) := \int_{\Theta^2} |\tilde\theta-\theta|^2 d\gamma$ and where we used Jensen inequality to get the first inequality and Lemma 4 to get the last inequality. Hence, with previous bound and (22)

$$\mathcal{E}_{\tau,r}(\nu) = \mathcal{E}_{\tau,r}(\mu) + d\,\mathcal{E}_{P_\tau(\mu)}\Big(\int_{K_r^2} \nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta)d\gamma\Big) + O(C_2^2(\gamma)). \tag{23}$$

Now we focus on the first order term and by Fubini :

$$\begin{aligned} d\,\mathcal{E}_{P_\tau(\mu)}\Big(\int_{K_r^2} \nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta)d\gamma\Big) &= \langle \nabla P_\tau(\mu), \nabla \int_{K_r^2} \nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta)d\gamma\rangle_{L^2(\Omega)} \\ &\quad - \langle f, \int_{K_r^2} \nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta)d\gamma\rangle_{L^2(\Omega)} \\ &\quad + \int_\Omega P_\tau(\mu)dx \times \int_\Omega \int_{K_r^2} \nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta)d\gamma dx \\ &= \int_{K_r^2} \langle \nabla P_\tau(\mu), \nabla\nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta)\rangle_{L^2(\Omega)}d\gamma \\ &\quad - \int_{K_r^2} \nabla_\theta \langle f, \Phi_\tau(\theta;\cdot)\rangle_{L^2(\Omega)}\cdot(\tilde\theta-\theta)d\gamma \\ &\quad + \int_{K_r^2}\int_\Omega P_\tau(\mu)dx \times \int_\Omega \nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta)dx d\gamma \\ &= \int_{K_r^2} v_\mu(\theta)\cdot(\tilde\theta-\theta)d\gamma. \end{aligned}$$

where :

$$v_\mu(\theta) := \nabla_\theta \phi_\mu(\theta) \ \gamma \text{ ae} \tag{24}$$

with

$$\phi_\mu(\theta) := \langle \nabla P_\tau(\mu), \nabla \Phi_\tau(\theta;\cdot)\rangle_{L^2(\Omega)} - \langle f, \Phi_\tau(\theta;\cdot)\rangle_{L^2(\Omega)} + \int_\Omega P_\tau(\mu)dx \times \int_\Omega \Phi_\tau(\theta;\cdot)dx.$$

Note that (24) is equivalent to

$$v_\mu(\theta) := \nabla_\theta \phi_\mu(\theta) \ \mu \text{ ae}$$

as v_μ depends only on θ.

□

To prove a well-posedness result, some convexity is needed. More precisely, one should be certain that $\mathcal{E}_{\tau,r}$ is convex along geodesics.

Proposition 4. *There exists $\lambda_{\tau,r} > 0$ such that for all $\mu,\nu \in D(\mathcal{E}_{\tau,r})$ with associated geodesic $\mu(t) := e_{t\#}\Pi$ given by Proposition 2, the functional $t \to \dfrac{d\,\mathcal{E}_{\tau,r}}{dt}(\mu_t)$ is $\lambda_{\tau,r}$ Lipschitz.*

Proof. First of all, one has to prove that for all $t \in [0,1]$, $\mu_t \in D(\mathcal{E}_{\tau,r})$. This is a direct consequence of the fact that μ, ν are supported in K_r, Remark 5 and that K_r is convex (in the geodesic sense). This proves that for all $0 \leq t \leq 1$, $\mu_t \in D(\mathcal{E}_{\tau,r})$.

Let $t, s \in [0,1]$ and take $\alpha_{t,s} := (e_t, e_s)_{\#}\Pi \in \Gamma(\mu_t, \mu_s)$. By (23), it holds :

$$\mathcal{E}_{\tau,r}(\mu_s) = \mathcal{E}_{\tau,r}(\mu_t) + \int_{\Theta^2} d\,\mathcal{E}_{P_\tau(\mu_t)} \Big(\nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde{\theta}-\theta)\Big) d\alpha_{t,s} + O(C_2^2(\alpha_{t,s}))$$

equivalent to :

$$\frac{\mathcal{E}_{\tau,r}(\mu_s) - \mathcal{E}_{\tau,r}(\mu_t)}{s-t} = \int_\Gamma d\,\mathcal{E}_{P_\tau(\mu_t)} \Big(\nabla_\theta \Phi_\tau(\pi(t);\cdot)\cdot\Big(\frac{\pi(s)-\pi(t)}{s-t}\Big)\Big) d\Pi(\pi) + O((t-s)C_2^2(\gamma))$$

Taking the limit as s goes to t, one concludes that $t \to \mathcal{E}_P(\mu_t)$ is differentiable with derivative equals to :

$$h(t) := \frac{d\mathcal{E}_{\tau,r}(\mu_t)}{dt} = \int_\Gamma d\,\mathcal{E}_{P_\tau(\mu_t)} \Big(\nabla_\theta \Phi_\tau(\pi(t);\cdot)\cdot\Big(\frac{d}{dt}\pi(t)\Big)\Big) d\Pi(\pi).$$

To conclude, one has the decomposition :

$$\begin{aligned} |h(t)-h(s)| \leq & \Big| \int_\Gamma d\,\mathcal{E}_{P_\tau(\mu_t)} \Big((\nabla_\theta \Phi_\tau(\pi(t);\cdot) - \nabla_\theta \Phi_\tau(\pi(s);\cdot))\cdot\Big(\frac{d}{dt}\pi(t)\Big)\Big) d\Pi(\pi)\Big| \\ & + \Big| \int_\Gamma (d\,\mathcal{E}_{P_\tau(\mu_t)} - d\,\mathcal{E}_{P_\tau(\mu_s)}) \Big(\nabla_\theta \Phi_\tau(\pi(s);\cdot)\cdot\Big(\frac{d}{dt}\pi(t)\Big)\Big) d\Pi(\pi)\Big| \\ & + \Big| \int_\Gamma d\,\mathcal{E}_{P_\tau(\mu_s)} \Big(\nabla_\theta \Phi_\tau(\pi(s);\cdot)\cdot\Big(\frac{d}{dt}\pi(t) - \frac{d}{dt}\pi(s)\Big)\Big) d\Pi(\pi)\Big|. \end{aligned} \tag{25}$$

Next we use the lemma of regularity stated below.

Lemma 5. *The following regularity estimates hold :*

- *For all $u, v \in H^1(\Omega)$,*

$$\|d\,\mathcal{E}_u - d\,\mathcal{E}_v\|_{\mathcal{L}(H^1(\Omega))} \leq C\|u-v\|_{H^1(\Omega)}$$

- *For all $\tau, r > 0$, there exists $C(\tau,r) > 0$ such that for all θ_1, θ_2 in K_r*

$$\begin{cases} \|\nabla_\theta \Phi_\tau(\theta_1;\cdot)\|_{H^1(\Omega)} \leq & C(\tau,r)|\theta_1|, \\ \|\nabla_\theta \Phi_\tau(\theta_1;\cdot) - \nabla_\theta \Phi_\tau(\theta_2;\cdot)\|_{H^1(\Omega)} \leq & C(\tau,r)|\theta_1-\theta_2|. \end{cases}$$

- *For all $r > 0, \mu, \nu \in D(\mathcal{E}_{\tau,r})$:*

$$\|P_\tau(\mu)\|^2_{H^1(\Omega)} \leq C(\tau)\int_\Theta |\theta|^2 d\mu(\theta). \tag{26}$$

and

$$\|P_\tau(\mu) - P_\tau(\nu)\|^2_{H^1(\Omega)} \leq C(\tau,r) W_2^2(\mu,\nu).$$

Proof. The proof of the first item is elementary and left to the reader. For the second one take $\tilde{\theta} \in \Theta$ and remark that :

$$\nabla \Phi_\tau(\theta; x) = a w \sigma_\tau'(w\cdot x + b) \tag{27}$$

and finally,

$$\begin{cases} \nabla \dfrac{\partial \Phi_\tau(\theta;x)}{\partial c} = 0 \\ \nabla \dfrac{\partial \Phi_\tau(\theta;x)}{\partial a} = w\sigma_\tau'(w\cdot x+b) \\ \nabla \dfrac{\partial \Phi_\tau(\theta;x)}{\partial w} = a\sigma_\tau'(w\cdot x+b)I_d + axw^T\sigma_\tau''(w\cdot x+b) \\ \nabla \dfrac{\partial \Phi_\tau(\theta;x)}{\partial b} = aw\sigma_\tau''(w\cdot x+b). \end{cases} \tag{28}$$

By (20),

$$|\nabla_\theta \Phi_\tau(\theta_1;x) - \nabla_\theta \Phi_\tau(\theta_2;x)| \le C(\tau)\max(|\theta_1|,|\theta_2|)|\theta_2-\theta_1|.$$

This is mainly because σ_τ' is Lipschitz and bounded and Ω is bounded. Whereas, by (28)

$$|\nabla(\nabla_\theta \Phi_\tau(\theta_1;x) - \nabla_\theta \Phi_\tau(\theta_2;x))| \le C(\tau)\max(|\theta_1|,|\theta_2|)\|\theta_2-\theta_1|.$$

This is mainly because $\sigma_\tau', \sigma_\tau''$ is Lipschitz and bounded and Ω is bounded. Hence, if $\theta_1,\theta_2 \in K_r$

$$\|\nabla_\theta \Phi_\tau(\theta_1;\cdot) - \nabla_\theta \Phi_\tau(\theta_2;\cdot)\|_{H^1(\Omega)} \le C(\tau,r)|\theta_1-\theta_2|.$$

The bound on $\|\nabla_\theta \Phi_\tau(\theta_1;\cdot)\|_{H^1(\Omega)}$ is obtained in a similar way. For the third item, by Jensen's inequality

$$\begin{aligned} \|P_\tau(\mu)\|^2_{H^1(\Omega)} &= \int_\Omega \left(\int_\theta \Phi_\tau(\theta;x)d\mu(\theta)\right)^2 dx + \int_\Omega \left|\int_\theta \nabla\Phi_\tau(\theta;x)d\mu(\theta)\right|^2 dx \\ &\le \int_\Omega\int_\theta \Phi_\tau^2(\theta;x)d\mu(\theta)dx + \int_\Omega\int_\theta |\nabla\Phi_\tau(\theta;x)|^2 d\mu(\theta)dx \\ &\le C(\tau)\int_\Theta |\theta|^2 d\mu(\theta) \end{aligned}$$

where we used the fact that $\Phi_\tau, \nabla\Phi_\tau$ are sublinear and Ω is bounded in the second inequality. The bound for the difference term $P_\tau(\mu) - P_\tau(\nu)$ can be proven in a similar way using Lemma 4.

□

Recalling (25), denoting $\gamma := (e_0,e_1)_\#\Pi$ and by previous lemma :

$$\begin{aligned} |h(t)-h(s)| &\le C(\tau,r)\Big(\|P_\tau(\mu_t)\|_{H^1(\Omega)}\int_\Gamma |\pi(t)-\pi(s)|\left|\frac{d}{dt}\pi(t)\right| d\Pi(\pi) \\ &+ \|P_\tau(\mu_t)-P_\tau(\mu_s)\|_{H^1(\Omega)}\int_\Gamma |\pi(s)|\left|\frac{d}{dt}\pi(t)\right| d\Pi(\pi) \\ &+ \|P_\tau(\mu_s)\|_{H^1(\Omega)}\int_\Gamma |\pi(s)|\left|\frac{d}{dt}\pi(t)-\frac{d}{dt}\pi(s)\right| d\Pi(\pi)\Big) \\ &\le C(\tau,r)\Big(|t-s|\|P_\tau(\mu_t)\|_{H^1(\Omega)}\int_\Gamma |\pi(1)-\pi(0)|^2 d\Pi(\pi) \\ &+ \|P_\tau(\mu_t)-P_\tau(\mu_s)\|_{H^1(\Omega)}\int_\Gamma \sup_u|\pi(u)||\pi(1)-\pi(0)|d\Pi(\pi) \\ &+ |t-s|\|P_\tau(\mu_s)\|_{H^1(\Omega)}\int_\Gamma \sup_u|\pi(u)|\sup_u\left|\frac{d^2\pi(u)}{dt^2}\right| d\Pi(\pi)\Big) \\ &\le C(\tau,r)\left(|t-s|\left(\sqrt{\int_{\Theta^2}|\theta|^2 d\mu_t(\theta)}+\sqrt{\int_{\Theta^2}|\theta|^2 d\mu_s(\theta)}\right)(1+C_2^2(\gamma)) + W_2(\mu_t,\mu_s)C_2^2(\gamma)\right) \end{aligned}$$

where we have used Lemma 4 to get the second inequality and the fact that $\sup_u\left|\frac{d^2\pi(u)}{dt^2}\right|$ is uniformly bounded (the curvature of Θ is bounded) to get the last one. Owing that :

- By Remark 5 and the convexity of K_r (in the geodesic sense), for all $0 \leq t \leq 1$:

$$\int_\Theta |\theta|^2 d\mu_t(\theta) = \int_{\Theta^2} |\pi(t)|^2 d\Pi(\pi) \leq C(1+r^2).$$

- The natural estimate holds :

$$\begin{aligned} W_2^2(\mu_t, \mu_s) &\leq \int_{\Theta^2} d(\theta, \tilde{\theta})^2 d\alpha_{t,s} \\ &\leq \int_\Gamma d(\pi(t), \pi(s))^2 d\Pi(\pi) \\ &= |t-s| \int_\Gamma d(\pi(1), \pi(0))^2 d\Pi(\pi) \\ &= |t-s| \int_{\Theta^2} d(\theta, \tilde{\theta})^2 d\gamma(\theta, \tilde{\theta}). \end{aligned}$$

This allows us to conclude that :

$$|h(t) - h(s)| \leq C(\tau, r)(1 + C_2^2(\gamma))|t-s|.$$

As the plan γ is supported in K_r^2, we get :

$$|h(t) - h(s)| \leq \lambda_{\tau,r}|t-s|.$$

for some $\lambda_{\tau,r} > 0$ and this finishes the proof. □

The characterization of the velocity field allows to get a bound on its amplitude. This is given by the next corollary which will be useful later in the paper.

Corollary 1. *For all $\mu \in D(\mathcal{E}_{\tau,r})$ and $\theta \in \Theta$:*

$$|v_\mu(\theta)| \leq C(\tau) r |\theta|.$$

Proof. This can be proved combining (26), (20) and (28). The rest is just basic computations and left to the reader. □

An important consequence of last proposition is that $\mathcal{E}_P$ is convex along geodesics. Before going into the main result of this section, we recall the basic definition of local slope [15].

Definition 4. *At every $\mu \in D(\mathcal{E}_{\tau,r})$, the local slope writes :*

$$|\nabla^- \mathcal{E}_{\tau,r}|(\mu) := \limsup_{\nu \to \mu} \frac{(f(\mu) - f(\nu))_+}{W_2(\mu, \nu)}$$

which may be infinite.

To understand the upcoming result, the careful reader should read appendix A which introduces the concept of gradient curve.

Theorem 4. *For all $\tilde{\mu}_0 \in D(\mathcal{E}_{\tau,r})$, there exists a unique locally Lipschitz gradient curve $(\mu_t)_t$ which is also a curve of maximal slope with respect to the upper gradient $|\nabla^- \mathcal{E}_{\tau,r}|$. Moreover, there exists a vector field v_t such that for almost all $t \geq 0$:*

$$\int_\Theta v_t^2 d\mu_t = \|v_t\|^2_{L^2(\Theta; d\mu_t)} < +\infty \tag{29}$$

and :

$$\begin{cases} \partial_t \mu_t + div(v_t \mu_t) = 0 \\ \mu_0 = \tilde{\mu}_0 \\ \mu_t \in D(\mathcal{E}_{\tau,r}). \end{cases} \tag{30}$$

Proof. The functional $\mathcal{E}_{\tau,r}$ is lower semicontinuous by Remark 4 and it is $-\lambda_{\tau,r}$ convex along generalized geodesics. Moreover, the space Θ has a curvature bounded from below which ensures that it is an Alexandrov space of curvature bounded from below. We apply [16, Theorem 5.9, 5.11] to get the existence and the uniqueness of a gradient curve in the sense of [16, Definition 5.8]. Being a gradient curve, it is also a curve maximal slope in the sense of [15, Definition 1.3.2].

The existence of the vector field v_t is given by the absolute continuity of the curve $(\mu_t)_t$ (because it is a gradient curve) and by [17, Proposition 2.5].

□

The work is not done here since we do not have any knowledge about the velocity field v_t and the well-posedness result is proved only for $\mathcal{E}_{\tau,r}$ with $r < \infty$.

3.1.2 Identification of the vector field v_t

To pursue our reasoning, the operators G, S_h for $0 \leq h \leq 1$ are introduced as follows :

$$G := \left\{ \begin{array}{ccc} \Gamma & \to & T\Theta \\ \pi & \mapsto & (\pi(0), \dot{\pi}(0)) \end{array} \right.$$

and :

$$S_h := \left\{ \begin{array}{ccc} T\Theta & \to & T\Theta \\ (\theta, v) & \mapsto & \left(\theta, \dfrac{v}{h}\right). \end{array} \right.$$

Furthermore, next lemma gives the local behavior of couplings along $(\mu_t)_t$:

Lemma 6. *If μ_t is solution to (30) and $h \to \gamma_h = (e_0, e_1)_{\#}\Pi_h \in \Gamma_o(\mu_t, \mu_{t+h})$ (Π_h defined in Proposition 2) then for almost all t :*

$$\lim_{h\to 0} (S_h \circ G)_{\#}\Pi_h = (i \times v_t)_{\#}\mu_t \text{ in } \mathcal{P}_2(T\Theta).$$

Moreover,

$$\lim_{h\to 0} \frac{W_2^2(\mu_{t+h}, \exp(hv_t)_{\#}\mu_t)}{h^2} = 0.$$

Proof. Let ϕ be in $C_c^\infty(\Theta)$. The continuity equation gives :

$$\int_{\mathbb{R}^+} \eta'(t) \int_\Theta \phi d\mu_t dt = -\int_{\mathbb{R}^+} \eta(t) \int_\Theta \nabla_\theta \phi \cdot v_t d\mu_t dt$$

for η smooth compactly supported in $\mathbb{R}^+$. Taking η as an approximation of the characteristic function of $[t, t+h]$ and passing to the limit, one gets :

$$\mu_t(\phi) - \mu_{t+h}(\phi) = -\int_t^{t+h} \int_\Theta \nabla_\theta \phi \cdot v_t d\mu_t dt.$$

Passing to the limit in h, one gets the differentiability almost everywhere of $t \to \mu_t(\phi)$ with (29) and :

$$\lim_{h\to 0} \frac{\mu_{t+h}(\phi) - \mu_t(\phi)}{h} = \int_\Theta \nabla_\theta \phi \cdot v_t d\mu_t.$$

Let $\nu_h := (S_h \circ G)_{\#}\Pi_h$ and take a limit point ν_0 of $(\nu_h)_h$ wrt the narrow convergence on $\mathcal{P}_2(T\Theta)$. Then,

$$\begin{aligned}
\frac{\mu_{t+h}(\phi)-\mu_t(\phi)}{h} &= \frac{1}{h}\int_{\Theta^2}(\phi(\tilde{\theta})-\phi(\theta))d\gamma_h \\
&= \frac{1}{h}\int_{\Gamma}(\phi(\pi(1))-\phi(\pi(0)))d\Pi_h(\pi) \\
&= \frac{1}{h}\int_{T\Theta}(\phi(\exp_\theta(v))-\phi(\theta))dG_{\#}\Pi_h(\theta,v) \\
&= \frac{1}{h}\int_{T\Theta}(\phi(\exp_\theta(hv))-\phi(\theta))d(S_h\circ G)_{\#}\Pi_h(\theta,v) \\
&= \int_{T\Theta}\nabla\phi(\theta)\cdot v d(S_h\circ G)_{\#}\Pi_h(\theta,v) \\
&+ \int_{T\Theta}R(h,\theta,v)d\nu_h(\theta,v) \\
&\to \int_{T\Theta}\nabla\phi(\theta)\cdot v d\nu_0(\theta,v)
\end{aligned}$$

where $R(h,\theta,v) = \dfrac{\phi(\exp_\theta(hv))-\phi(\theta)}{h} - \nabla\phi(\theta)\cdot v$ is bounded by $C(\phi)|v|^2h$ ($\phi \in C_c^\infty(\Theta)$ and the euclidean curvature in Θ is uniformly bounded; see [18, Chapter 8] for the definition of euclidean curvature). In fact to get the last limit, we need a bit of work detailed below :

- For the first term :

$$\begin{aligned}
\int_{T\Theta}|\nabla\phi(\theta)\cdot v|d(S_h\circ G)_{\#}\Pi_h(\theta,v) &= \int_{T\Theta}\frac{1}{h^2}|\nabla\phi(\pi(0))\cdot\dot{\pi}(0)|d\Pi_h(\pi) \\
&\leq \int_{T\Theta}\frac{C(\phi)}{h^2}d(\pi(1),\phi(0))d\Pi_h(\pi) \\
&\leq \frac{C(\phi)}{h^2}\int_{T\Theta}d(\pi(1),\pi(0))^2d\Pi_h(\pi) \\
&\leq C(\phi)\frac{W_2^2(\mu_t,\mu_{t+h})}{h^2}
\end{aligned}$$

  where we used the fact that $\phi \in C_c^\infty(\Theta)$ in the first inequality. As $(\mu_t)_t$ is locally Lipschitz by Theorem 4, $|\nabla\phi(\theta)\cdot v|$ is uniformly integrable wrt $((S_h\circ G)_{\#}\Pi_h)_h$ and the passage to the limit is allowed.

- For the second one,

$$\begin{aligned}
\int_{T\Theta}|R(h,\theta,v)|d\nu_h(\theta,v) &\leq C(\phi)h\int_{T\Theta}|v|^2d\nu_h(\theta,v) \\
&= C(\phi)h\frac{W_2^2(\mu_t,\mu_{t+h})}{h^2}
\end{aligned}$$

  and using again the local Lipschitz property, we can pass to the limit which is zero.

As a consequence,

$$\int_{T\Theta}\nabla_\theta\phi(\theta)\cdot v d\nu_0(\theta,v) = \int_\Theta \nabla_\theta\phi(\theta)\cdot v_t(\theta)d\mu_t(\theta)$$

which is no more than (by disintegration) :

$$\int_\Theta \nabla_\theta\phi(\theta)\cdot\int_{T_\theta\Theta} v d\nu_{0,\theta}(v)d\mu_t(\theta) = \int_\Theta\nabla_\theta\phi(\theta)\cdot v_t(\theta)d\mu_t(\theta).$$

Noting $\tilde{v_t}(\theta) := \displaystyle\int_{T_\theta\Theta} v d\nu_{0,\theta}(v)$, last equation is equivalent to :

$$div((\tilde{v_t}-v_t)\mu_t) = 0.$$

Aside that, as $(\theta, v) \to |v|^2$ is positive and lower semicontinuous and for $t \geq 0$ such that $\lim_{h\to 0} \frac{W_2(\mu_t, \mu_{t+h})}{h}$ $= |\mu_t'|(t)$ (this is a dense set of $\mathbb{R}^+$ as $(\mu_t)_t$ is locally Lipschitz) :

$$\begin{aligned}
\int_\Theta \int_{T_\theta \Theta} |v|^2 d\nu_{0,\theta}(v) d\mu_t(\theta) &\leq \liminf_{h\to 0} \int_{T\Theta} |v|^2 d\nu_h(\theta, v) \\
&= \liminf_{h\to 0} \frac{1}{h^2} \int_{T\Theta} |v|^2 dG_\# \Pi_h(\theta, v) \\
&= \liminf_{h\to 0} \frac{1}{h^2} \int_{T\Theta} |\dot{\pi}(0)|^2 d\Pi_h(\pi) \\
&= \liminf_{h\to 0} \frac{1}{h^2} \int_{\Theta^2} d(\theta, \tilde{\theta})^2 d\gamma_h(\theta, \tilde{\theta}) \\
&= \liminf_{h\to 0} \frac{W_2^2(\mu_t, \mu_{t+h})}{h^2} \\
&= |\mu_t'|^2(t).
\end{aligned} \tag{31}$$

As a consequence and by Jensen inequality,

$$\|\tilde{v}_t\|^2_{L^2(\Theta; d\mu_t)} \leq \int_\Theta \int_{T_\theta \theta} |v|^2 d\nu_{0,\theta}(v) d\mu_t(\theta) \leq |\mu_t'|^2(t) = \|v_t\|^2_{L^2(\Theta; d\mu_t)}. \tag{32}$$

By [17, Lemma 2.4], one gets $\tilde{v}_t = v_t$. Reconsidering (32), one gets the equality case in Jensen inequality *ie* :

$$\int_\Theta |\tilde{v}_t(\theta)|^2 \, d\mu_t(\theta) = \int_\Theta \int_{T_\theta \Theta} |v|^2 d\nu_{h,\theta}(v) d\mu_t(\theta)$$

and as a consequence $\nu_{0,\theta} = \delta_{v_t(\theta)}$ μ_t almost everywhere. And

$$\lim_{h\to 0} (S_h \circ G)_\# \Pi_h = (i \times v_t)_\# \mu_t$$

in the sense of the narrow convergence. The convergence of the v moment is given by (31)-(32) where inequalities can be replaced by equalities (as $\tilde{v}_t = v_t$) and the lim inf can be replaced by a lim as $\lim_{h\to 0} \frac{W_2(\mu_t, \mu_{t+h})}{h} = |\mu_t'|(t)$ exists :

$$\int_\Theta \int_{T_\theta \Theta} |v|^2 d\nu_{0,\theta}(v) d\mu_t(\theta) = \lim_{h\to 0} \int_{T\Theta} |v|^2 d\nu_h(\theta, v). \tag{33}$$

For the θ moment, it is more obvious as for all $h > 0$:

$$\int_{T\Theta} |\theta|^2 d\nu_h(\theta, v) = \int_\Theta |\theta|^2 d\mu_t(\theta)$$

and

$$\int_{T\Theta} |\theta|^2 d\nu_0(\theta, v) = \int_{T\Theta} |\theta|^2 d(i \times v_t)_\# \mu_t(\theta) = \int_\Theta |\theta|^2 d\mu_t(\theta).$$

Consequently,

$$\int_{T\Theta} |\theta|^2 d\nu_0(\theta, v) = \lim_{h\to 0} \int_{T\Theta} |\theta|^2 d\nu_h(\theta, v). \tag{34}$$

With (33)-(34), the convergence of moments is asserted. The narrow convergence combined with the convergence of moments gives the convergence in $P_2(\Theta)$ and the proof of the first part of the lemma is finished.

For the second part, one has $(\exp(hv_t) \times i)_\# \gamma_h$ belongs to $\Gamma(\exp(hv_t)_\# \mu_t, \mu_{t+h})$

$$\begin{aligned}\frac{W_2^2(\mu_{t+h}, \exp(hv_t)_\#\mu_t)}{h^2} &\leq \frac{1}{h^2}\int_{\Theta^2} d(\theta,\tilde{\theta})^2 d(\exp(hv_t)\times i)_\#\gamma_h \\ &\leq \frac{1}{h^2}\int_{\Theta^2} d(\exp_\theta(hv_t(\theta)),\tilde{\theta})^2 d\gamma_h \\ &\leq \frac{1}{h^2}\int_{\Theta^2} d(\exp_\theta(hv_t(\theta)),\exp_\theta(hv))^2 d\nu_h \\ &\leq C\int_{\Theta^2} |v_t(\theta)-v|^2 d\nu_h \\ &\to_{h\to 0} 0\end{aligned}$$

where we have used the boundedness of the euclidean curvature of the manifold Θ in last inequality and the fact that $d\nu_h \to (i\times v_t)_\#\mu_t$ proved earlier to get the limit. □

We introduce the operator of projection on the manifold Θ :

Definition 5. *For all θ in Θ, the projection on the tangent space of Θ is given by the operator $\Pi_\theta : \mathbb{R}^{d+3} \to T_\theta\Theta$. The operator Π denotes the corresponding projection on vector fields ie $(\Pi X)(\theta) := \Pi_\theta X(\theta)$ for any vector field X and any θ in Θ.*

Now we are able to identify the velocity field given in Theorem 4 under a support hypothesis.

Proposition 5. *Let $t \geq 0$. If there exists $\delta > 0$ such that $supp(\mu_t) \subset \dot{K_{r-\delta}}$ then for μ_t almost everywhere, the velocity v_t in (30) is equal to $-\Pi v_{\mu_t}$.*

Proof. On the one hand, for $\gamma_h := (e_0, e_1)_\#\Pi_h \in \Gamma_o(\mu_t, \mu_{t+h})$, by Proposition 3 and the fact that for all $t \geq 0$, $\mu_t \in D(\mathcal{E}_{\tau,r})$:

$$\mathcal{E}_{\tau,r}(\mu_{t+h}) - \mathcal{E}_{\tau,r}(\mu_t) = \int_{\Theta^2} v_{\mu_t}(\theta)\cdot(\tilde{\theta}-\theta)d\gamma_h(\theta,\tilde{\theta}) + o(W_2(\mu_t,\mu_{t+h}))$$

which is equivalent to

$$\frac{\mathcal{E}_{\tau,r}(\mu_{t+h}) - \mathcal{E}_{\tau,r}(\mu_t)}{h} = \int_{T\Theta} v_{\mu_t}(\theta)\cdot\frac{\exp_\theta(hv)-\theta}{h} d(S_h\circ G)_\#\Pi_h(\theta,v) + \frac{1}{h}o(W_2(\mu_t,\mu_{t+h})).$$

Then, one can use the decomposition :

$$\begin{aligned}\int_{T\Theta} v_{\mu_t}(\theta)\cdot\frac{\exp_\theta(hv)-\theta}{h} d(S_h\circ G)_\#\Pi_h(\theta,v) &= \int_{T\Theta} v_{\mu_t}(\theta)\cdot v\, d(S_h\circ G)_\#\Pi_h(\theta,v) \\ &\quad + \int_{T\Theta} v_{\mu_t}(\theta)\cdot R(h,\theta,v) d(S_h\circ G)_\#\Pi_h(\theta,v)\end{aligned}$$

where $R(h,\theta,v) := \dfrac{\exp_\theta(hv)-\theta}{h} - v$ is bounded by $Ch|v|^2$ due to the uniform boundedness of euclidean curvature in Θ. Passing to the limit as h goes to zero and using Lemma 6, one gets the differentiability of $t \to \mathcal{E}_{\tau,r}(\mu_t)$ almost everywhere and for almost all $t \geq 0$:

$$\frac{d\,\mathcal{E}_{\tau,r}(\mu_t)}{dt} = \int_{T\Theta} v_{\mu_t}(\theta)\cdot v_t(\theta) d\mu_t(\theta,v).$$

Note that to pass to the limit to obtain last equation, we need the two following points :

- First that $v\cdot v_{\mu_t}(\theta)$ is at most quadratic in (θ, v) which is given by Corollary 1.
- $|v_{\mu_t}(\theta)\cdot R(h,\theta,v)| \leq Cr|\theta|h|v|^2$ by Corollary 1 and thus uniformly integrable wrt $((S_h\circ G)_\#\Pi_h)_h$ as it is supported in K_r in the θ variable and :

$$\int_{T\Theta} |v|^2 d(S_h\circ G)_\#\Pi_h(\theta,v) = \frac{W_2^2(\mu_t,\mu_{t+h})}{h^2}$$

which is bounded by the local Lipschitz property of $(\mu_t)_t$.

Next as $\Pi v_t = v_t$, it holds :

$$\frac{d\,\mathcal{E}_{\tau,r}(\mu_t)}{dt} = \int_{\Theta^2} \Pi v_{\mu_t}(\theta)\cdot v_t(\theta) d\mu_t(\theta). \tag{35}$$

On the other hand, consider the curve $t \to \tilde{\mu}_h$ satisfying :

$$\tilde{\mu}_h := \exp(-h\Pi v_{\mu_t})_{\#}\mu_t.$$

As $\mathrm{supp}(\mu_t) \subset K_{r-\delta}$, there exists a small time interval around zero such that $(\tilde{\mu}_h)_h$ is in $D(\mathcal{E}_{\tau,r})$. So with $\gamma_h = (i \times \exp(-h\Pi v_{\mu_t}))_{\#}\mu_t \in \Gamma(\mu_t, \tilde{\mu}_h)$:

$$\begin{aligned}\mathcal{E}_{\tau,r}(\tilde{\mu}_h) - \mathcal{E}_{\tau,r}(\mu_t) &= \int_{\Theta^2} \Pi v_{\mu_t}(\theta)\cdot(\tilde{\theta}-\theta) d\gamma_h(\theta,\tilde{\theta}) + o(W_2(\mu_t,\tilde{\mu}_h)) \\ &= h\int_{\Theta^2} \Pi v_{\mu_t}(\theta)\cdot \frac{\exp_\theta(-h\Pi v_{\mu_t}(\theta)) - \theta}{h} d\mu_t(\theta) + o(W_2(\mu_t,\tilde{\mu}_h))\end{aligned}$$

Hence,

$$\frac{\mathcal{E}_{\tau,r}(\tilde{\mu}_h) - \mathcal{E}_{\tau,r}(\mu_t)}{W_2(\tilde{\mu}_h,\mu_t)} = \frac{h}{W_2(\tilde{\mu}_h,\mu_t)}\int_{\Theta^2} \Pi v_{\mu_t}(\theta)\cdot \frac{\exp_\theta(-h\Pi v_{\mu_t}(\theta)) - \theta}{h} d\mu_t(\theta) + o(1)$$

and getting the limsup as h goes to zero (proceeding in the similar way as above to get the limit of the first term on the right hand side) and owing the fact that $\limsup_{h\to 0} \frac{W_2(\tilde{\mu}_h,\mu_t)}{h} \le \|\Pi v_{\mu_t}\|_{L^2(\Theta;d\mu_t)}$ (left to the reader) :

$$|\nabla^-\,\mathcal{E}_{\tau,r}\,|(\mu_t) \ge \|\Pi v_{\mu_t}\|_{L^2(\Theta;d\mu_t)}. \tag{36}$$

As $t \to \mu_t$ is a curve of maximal slope with respect to the upper gradient $|\nabla^-\,\mathcal{E}_{\tau,r}\,|$ of $\mathcal{E}_{\tau,r}$, one has :

$$\begin{aligned}\frac{d\,\mathcal{E}_{\tau,r}(\mu_t)}{dt} &= \int_{\Theta} \Pi v_{\mu_t}(\theta)\cdot v_t(\theta) d\mu_t(\theta) \le -\frac{1}{2}\|v_t\|_{L^2(\Theta;d\mu_t)} - \frac{1}{2}|\nabla^-\,\mathcal{E}_{\tau,r}\,|^2(\mu_t) \\ &\le -\frac{1}{2}\|v_t\|^2_{L^2(\Theta;d\mu_t)} - \frac{1}{2}\|\Pi v_{\mu_t}\|^2_{L^2(\Theta;d\mu_t)}\end{aligned}$$

where we have used (36). As a consequence,

$$\int_\Theta \left(\frac{1}{2}(\Pi v_{\mu_t})^2(\theta) + \frac{1}{2}v_t^2(\theta) - \Pi v_{\mu_t}(\theta)\cdot v_t(\theta)\right) d\mu_t(\theta) \le 0$$

and

$$v_t = -\Pi v_{\mu_t}\ \mu_t \text{ a.e.}$$

□

The identification of the velocity field when the support condition is satisfied allows to give an explicit formula for the gradient curve. It is given by the characteristics :

Proposition 6. *Let χ_t be the flow associated to the velocity field $-\Pi v_{\mu_t}$:*

$$\begin{cases}\dfrac{\partial \chi_t}{\partial t}(\theta) = -\Pi v_{\mu_t}(\theta) \\ \chi_t(\theta) = \theta.\end{cases}$$

Then χ is uniquely defined, continuous, Lipschitz on K_r. Moreover, as long as $supp(\mu_t) \subset K_{r-\delta}$ for some $\delta > 0$:

$$\mu_t = \chi_{t\#}\mu_0.$$

Proof. This is a direct consequence of the fact that $v_t = -\Pi v_{\mu_t} = -\Pi\nabla_\theta\phi_{\mu_t}$ is C^∞. □

Next lemma ensures that the curve $h \to \exp(hv_t)_\# \mu_t$ is a geodesic. This will be useful later to prove that the velocity field characterizes the gradient curve.

Lemma 7. *For $\mu \in \mathcal{P}_2(\Theta)$ with $supp(\mu) \subset K_{r-\delta}$ for some $\delta > 0$, the map $\nu_h : h \to \exp(-h\Pi v_\mu / \|\Pi v_\mu\|_{L^2(\Theta;d\mu)})_\# \mu$ is differentiable at $h = 0$ and moreover :*

$$\nu_0' = \nabla_- \mathcal{E}_{\tau,r}(\mu)/|\nabla_- \mathcal{E}_{\tau,r}|(\mu).$$

Proof. First, we claim that $|\nabla_- \mathcal{E}_{\tau,r}(\mu)| = \|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}$. In order to prove it, take an arbitrary unit speed geodesic $s \to (e_s)_\# \Pi$ starting at μ for which there exists a time interval around zero such that $(e_s)_\# \Pi$ belongs to $D(\mathcal{E}_{\tau,r})$. Note that normally, we should write a unit speed geodesic as $(e_{\delta s})_\# \Pi$ with $\delta > 0$ a scaling factor but for simplicity here, we take $\delta = 1$. As a consequence, one can write for all $s > 0$ sufficiently small :

$$\begin{aligned}\mathcal{E}_{\tau,r}((e_s)_\#\Pi) &= \mathcal{E}_{\tau,r}(\mu) + \int_{\Theta^2} v_\mu(\theta) \cdot (\tilde{\theta} - \theta) d(e_0, e_s)_\# \Pi(\theta, \tilde{\theta}) + o(W_2(\mu, (e_s)_\#\Pi)) \\ &= \mathcal{E}_{\tau,r}(\mu) + \int_{T\Theta} v_\mu(\theta) \cdot (\exp_\theta(sv) - \theta) dG_\# \Pi(\theta, v) + o(W_2(\mu, (e_s)_\#\Pi)).\end{aligned}$$

Dividing by s and passing to the limit as s goes to zero, one obtains :

$$\frac{d}{ds} \mathcal{E}_{\tau,r}((e_s)_\#\Pi) = \int_{T\Theta} v_\mu(\theta) \cdot v dG_\# \Pi(\theta, v).$$

Note to get the last equation, we need to prove that $\eta : s \to v_\mu(\theta) \cdot \dfrac{\exp_\theta(sv) - \theta}{s}$ is uniformly integrable wrt $G_\#\Pi$. In fact, this is given by Corollary 1 and the uniform curvature bound on Θ giving $|\eta(s)| \leq Csr|\theta||v|^2$. As the term $Cr|\theta||v|^2$ is integrable wrt measure $G_\#\Pi$ (recall that it is $\mathcal{P}_2$ and supported in K_r in the θ variable), we have the desired uniform integrability.

Moreover, by Cauchy-Schwartz :

$$\begin{aligned}\frac{d}{ds} \mathcal{E}_{\tau,r}((e_s)_\#\Pi) &\geq -\|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)} \sqrt{\int_{T\Theta} v^2 dG_\# \Pi(\theta, v)} \\ &= -\|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}\end{aligned}$$

where last equality comes from :

$$\begin{aligned}\int_{T\Theta} v^2 dG_\# \Pi(\theta, v) &= \int_\Gamma \dot{\pi}(0)^2 d\Pi(\pi) \\ &= \int_\Gamma d(\pi(0), \pi(1))^2 d\Pi(\pi) \\ &= W_2^2((e_0)_\#\Pi, (e_1)_\#\Pi) \\ &= 1.\end{aligned}$$

The last equality is derived from the fact that $s \to (e_s)_\#\Pi$ is a unit speed geodesic. To conclude, we have proved that for all unit speed geodesic $(\alpha, 1) \in C_\mu(D(\mathcal{E}_{\tau,r}))$

$$D_\mu \mathcal{E}_{\tau,r}((\alpha, 1)) \geq -\|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}$$

which by [16, Lemma 4.3], asserts that :

$$|\nabla_- \mathcal{E}_{\tau,r}|(\mu) \leq \|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}. \tag{37}$$

Aside that, let $h > 0$:

$$\begin{aligned}W_2^2(\nu_h, \nu_0) &\leq \int_\Theta d^2(\exp_\theta(-h\Pi v_\mu(\theta)/\|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}), \theta) d\mu(\theta) \\ &\leq h^2 \int_\Theta d^2(\exp_\theta(-\Pi v_\mu(\theta)/\|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}), \theta) d\mu(\theta) \\ &= h^2\end{aligned}$$

and :

$$\limsup_{h\to 0} \frac{W_2(\nu_h,\nu_0)}{h} \leq 1. \tag{38}$$

Moreover as $\text{supp}(\mu) \subset K_{r-\delta}$, v_μ is bounded in L^∞ and for a small time interval around zero $\nu_h \in D(\mathcal{E}_{\tau,r})$. Consequently :

$$\begin{aligned}\mathcal{E}_{\tau,r}(\nu_h) - \mathcal{E}_{\tau,r}(\mu) &= \int_{\Theta^2} v_\mu(\theta)\cdot(\tilde{\theta}-\theta) d(i\times \exp(-h\Pi v_\mu/\|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}))_\# \mu(\theta) \\ &+ o\left(\int_\Theta d(\theta, \exp(-h\Pi v_\mu(\theta)/\|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}) d\mu(\theta)\right) \\ &= \int_\Theta v_\mu(\theta)\cdot(\exp(-h\Pi v_\mu(\theta)/\|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}) - \theta) d\mu(\theta) + o(h).\end{aligned}$$

Dividing by h and passing to the limit as h goes to zero (justifying the passage to the limit as above), it holds :

$$\lim_{h\to 0} \frac{\mathcal{E}_{\tau,r}(\nu_h) - \mathcal{E}_{\tau,r}(\mu)}{h} = -\|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}. \tag{39}$$

Additionally, with (38) :

$$\limsup_{h\to 0} \frac{\mathcal{E}_{\tau,r}(\nu_h) - \mathcal{E}_{\tau,r}(\mu)}{W_2(\nu_h,\nu_0)} \leq -\|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}. \tag{40}$$

To conclude :

- With (40) and (37), the claim is proved :

$$|\nabla_-\, \mathcal{E}_{\tau,r}\,|(\mu) = \|\Pi v_\mu(\theta)\|_{L^2(\Theta;d\mu)}.$$

- Owing this, (38) and (40) the curve $h \to \nu_h$ is differentiable at $h=0$ by [16, Proof of (ii) Lemma 5.4] and :

$$\nu_0' = \nabla_-\, \mathcal{E}_{\tau,r}(\mu)/|\nabla_-\, \mathcal{E}_{\tau,r}\,|(\mu).$$

This finishes the proof of the lemma.

□

3.1.3 Existence with no support limitation

Note that for the moment the domain of $\mathcal{E}_{\tau,r}$ is reduced to measures supported in K_r. Using a bootstrapping argument, the existence theorem 5 holds for the energy $\mathcal{E}_{\tau,+\infty}$.

Theorem 5. *For all μ_0 compactly supported, there exists a curve $(\mu_t)_t$ such that :*

$$\begin{cases} \partial_t \mu_t + div((-\Pi v_{\mu_t})\mu_t) = 0 \\ \mu_{t=0} = \mu_0 \end{cases} \tag{41}$$

and for almost all t :

$$\int_\Theta |\Pi v_{\mu_t}|^2 d\mu_t = \|\Pi v_{\mu_t}\|^2_{L^2(\Theta;d\mu_t)} < +\infty.$$

Moreover, the solution satisfies :

$$\forall t \geq 0, \mu_t = \chi_{t\#}\mu_0$$

with :

$$\begin{cases} \dfrac{\partial \chi_t}{\partial t}(\theta) = -\Pi v_{\mu_t}(\theta) \\ \chi_0(\theta) = \theta. \end{cases}$$

Proof. Let :

- r_0 be such that $\text{supp}(\mu_0) \subset K_{r_0}$,
- denote $(\mu_t^r)_t$ the gradient curve associated to $\mathcal{E}_{\tau,r}$ for $r > r_0$.

By Corollary 1, $|v_\mu(\theta)| \leq Cr|\theta|$. Hence characteristics χ_t starting from K_{r_0} verifies $|\chi_t| \leq r_0 e^{Crt}$ and for time t in $\left[0, T_r := \frac{1}{Cr}\log\left(\frac{r+r_0}{2r_0}\right)\right]$, $\text{supp}(\mu_t^r) \subset K_{(r+r_0)/2} \subset K_r$. By the definition of the gradient curve :

$$\forall t \in [0, T_r],\ (\mu_t^r)' = \nabla_- \mathcal{E}_{\tau,r}(\mu_t^r) = (h \to \exp\left(-\Pi v_{\mu_t^r} h\right))'(0) \tag{42}$$

by Lemma 7. Note that the right hand side of last equation does not depend explicitly on r but on $\mu_\cdot^r$.

We construct the curve $(\mu_t)_t$ for all time imposing :

$$\forall t \in [0, T_r],\ r > r_0\ \mu_t := \mu_t^r.$$

This is well-defined since by uniqueness of the gradient curve wrt $\mathcal{E}_{\tau,r}$, $\mu_t^{r_1} = \mu_t^{r_2}$ on $[0, \min(T_{r_1}, T_{r_2})]$ for $r_0 < r_1 \leq r_2$. Defining the sequence :

$$r_n := (n+1)r_0,$$

we can build inductively a gradient curve on $\left[0, \frac{1}{Cr_0}\sum_{i=1}^{n}\frac{1}{(i+1)}\log\left(\frac{i+2}{2(i+1)}\right)\right]$. As the width of this interval is diverging, it is possible to construct a gradient curve on $\mathbb{R}^+$.

All the properties given by the theorem comes from the properties of $(\mu_t^r)_t$ derived in Theorem 4 and Proposition 6. □

Remark 6. *Two important remarks to make :*

- *We did not prove the existence of a gradient curve wrt $\mathcal{E}_{\tau,\infty}$ because this functional is not convex along geodesics and it is impossible to define gradients without such assumption.*
- *The uniqueness of a solution to (41) is out of the scope of this article. To prove it, one should link (41) and the support condition to prove that locally in time, a solution to (41) coincide with the gradient curve of $\mathcal{E}_{\tau,r}$ for some $r > 0$ large enough. Nevertheless, this link seems to be difficult to prove and we discard it for our purposes.*

3.2 Link with backpropagation in neural network

Let $m > 0$ be an integer. A two-layer neural network u with σ as activation function can always be written as :

$$u = \frac{1}{m}\sum_{i=1}^{m}\Phi_\tau(\theta_i, \dot{)} \tag{43}$$

with $\theta_i \in \Theta$. Then, we differentiate the functional $\mathcal{F} : (\theta_1, \cdots, \theta_m) \to \mathcal{E}\left(\frac{1}{m}\sum_{i=1}^{m}\Phi_\tau(\theta_i, \cdot)\right)$:

$$d\,\mathcal{F}_{\theta_1,\cdots,\theta_m}(d\theta_1, \cdots, d\theta_m) = d\,\mathcal{E}_u\left(\frac{1}{m}\sum_{i=1}^{m}\nabla_\theta\Phi_\tau(\theta_i, \cdot)\cdot d\theta_i\right).$$

Consequently, the gradient of $\mathcal{F}$ is given by

$$\begin{aligned}\nabla_{\theta_i}\mathcal{F}(\theta_1, \cdots, \theta_m) &= \frac{1}{m}\nabla_\theta(d\,\mathcal{E}_u(\Phi_\tau(\theta, \cdot)))_{\theta_i} \\ &= \frac{1}{m}\nabla_\theta\phi_\mu(\theta)_{\theta_i}\end{aligned}$$

where :

$$\mu := \frac{1}{m}\sum_{i=1}^{m}\delta_{\theta_i} \in \mathcal{P}_2(\Theta). \tag{44}$$

As a consequence, a gradient descent of $\mathcal{F}$ in the sense that :

$$\begin{cases} \dfrac{d}{dt}(\theta_1(t),\cdots,\theta_m(t)) = -m\nabla\mathcal{F}(\theta_1(t),\cdots,\theta_m(t)) \\ (\theta_1(0),\cdots,\theta_m(0)) = (\theta_{1,0},\cdots,\theta_{m,0}) \end{cases}$$

is equivalent to the gradient curve of $\mathcal{E}_{\tau,+\infty}$ where $\mu(0) := \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\delta_{\theta_{i,0}}$.

Theorem 6. *Let $\mu_0 \in \mathcal{P}_2(\Omega)$ compactly supported, $(\mu_{0,m})_m$ be such that $\lim\limits_{m\to+\infty} W_2(\mu_{0,m},\mu_0) = 0$ and composed by a finite number of Dirac masses located in the support of μ_0. Then for all $T > 0$, the associated gradient curves constructed in Theorem 5 converge in $C([0,T],\mathcal{P}_2(\Omega))$.*

Proof. Returning back to the proof of Theorem 5 and for all time $T > 0$, one can find $r > 0$ large enough such that $\mu_t, \mu_{t,m}$ coincide with gradient curves on $[0,T]$ wrt $\mathcal{E}_{\tau,r}$. As gradient curves wrt $\mathcal{E}_{\tau,r}$ verifies the following semigroup property [16, Theorem 5.11]

$$\forall t \in [0,T],\ W_2(\mu_t,\mu_{t,m}) \le e^{\lambda_{\tau,r}t}W_2(\mu_0,\mu_{0,m}),$$

the expected convergence on $C([0,T],\mathcal{P}_2(\Omega))$ holds by the convergence of initial measures.

□

3.3 Convergence to optimum

In this section, a LaSalle's principle argument is invoked in order to prove that the gradient curve converges to a global minimizer of $\mathcal{E}_{\tau,\infty}$. for simplicity, we note $\mathcal{E}_\tau := \mathcal{E}_{\tau,\infty}$ for $0 < \tau < +\infty$.

3.3.1 Characterization of optima

In this part, we focus on a characterization of global optima. For convenience, we extend the operator $\mathcal{E}_\tau$ to signed finite measures.

Lemma 8. *For all $\mu \in \mathcal{M}(\Theta)$, there exists a probability measure μ_p such that $\mathcal{E}_\tau(\mu) = \mathcal{E}_\tau(\mu_p)$.*

Proof. Let us first consider a positive signed measure $\mu \in \mathcal{M}^+(\Theta)$. If $\mu(\Theta) = 0$, $\Phi(\theta,.) = 0$ μ almost everywhere and $\mathcal{E}_\tau(\mu) = 0$. Take $\mu_p := \delta_{(0,0,w,b)}$ with w, b taken arbitrary to prove the result. Now if $\mu(\Theta) \neq 0$, consider $\mu_p := T_\#\left(\dfrac{\mu}{\mu(\Theta)}\right)$ where $T : (c,a,w,b) \to (c\mu(\Theta), a\mu(\Theta), w, b)$. In this case :

$$\begin{aligned}\int_\Theta \Phi(\theta;\cdot)d\mu &= \int_\Theta \mu(\Theta)\Phi(\theta;\cdot)\frac{d\mu(\theta)}{\mu(\Theta)} \\ &= \int_\Theta \Phi(T\theta;\cdot)\frac{d\mu(\theta)}{\mu(\Theta)} \\ &= \int_\Theta \Phi(\theta;\cdot)d\mu_p(\theta)\end{aligned}$$

where we have used the form of Φ (14)-(15) to get the last inequality.

Now take an arbitrary signed measure $\mu \in \mathcal{M}(\Theta)$. By Hahn/Jordan decomposition theorem, there exists P, N measurable sets (for μ) such that $P \cup N = \Theta$ and μ is respectively positive (negative) on P(N). The signed measure μ can be written as :

$$\mu = \mu_P - \mu_N$$

where $\mu_P, \mu_N \in \mathcal{M}^+(\Theta)$. Consider following map :

$$G(c,a,w,b) := \begin{cases} (-c,-a,w,b) & \text{if } (a,b,w,c) \in N \\ (c,a,w,b) & \text{if } (a,b,w,c) \in P \end{cases}$$

and the measure :

$$\mu_G := G_{\#}(\mu_P + \mu_N) \in \mathcal{M}^+(\Theta).$$

By construction, we have $P_\tau\left(T_{\#}\left(\frac{\mu_G}{\mu_G(\Theta)}\right)\right) = P_\tau(\mu)$ and consequently, $\mathcal{E}_\tau(\mu) = \mathcal{E}_\tau\left(T_{\#}\left(\frac{\mu_G}{\mu_G(\Theta)}\right)\right)$. □

Lemma 9. *The measure $\mu \in \mathcal{P}_2(\Theta)$ is optimal for Problem 1 iff $\phi_\mu = 0$ everywhere.*

Proof. Suppose $\mu \in \mathcal{P}_2(\Theta)$ optimal then for all $\nu := f\mu + \nu^\perp \in \mathcal{M}(\Theta)$ (Lebesgue decomposition of ν wrt μ with $f \in L^1(\Theta;\mu)$) and owing Lemma 8 :

$$\begin{aligned}\mathcal{E}_\tau(\mu + t\nu) &= \mathcal{E}(P_\tau(\mu) + tP_\tau(\nu)) \\ &= \mathcal{E}_\tau(\mu) + td\,\mathcal{E}_{P_\tau(\mu)}(P_\tau(\nu)) + o(t).\end{aligned}$$

Hence as μ is optimal

$$\begin{aligned}0 = \frac{d\,\mathcal{E}_\tau(\mu + t\nu)}{dt} &= d\,\mathcal{E}_{P_\tau(\mu)}(P_\tau(\nu)) \\ &= \int_\Theta d\,\mathcal{E}_{P_\tau(\mu)}(\Phi_\tau(\theta;\cdot))d\nu(\theta) \\ &= \int_\Theta \phi_\mu(\theta)d\nu(\theta) \\ &= \int_\Theta \phi_\mu(\theta)f(\theta)d\mu(\theta) + \int_\Theta \phi_\mu(\theta)d\nu^\perp(\theta).\end{aligned}$$

As this is true for all $f \in L^1(\Theta,\mu)$, one gets:

$$\phi_\mu = 0 \ \mu \text{ ae }, \phi_\mu = 0 \ \nu^\perp \text{ ae} \tag{45}$$

for all $\nu^\perp \perp \mu$. As ϕ_μ is continuous this is equivalent to $\phi_\mu = 0$ everywhere. Indeed, suppose (45) for all $\nu^\perp \perp \mu$. Take $\theta \in \Omega$, then if θ belongs to $supp(\mu)$ then by definition of the support $\mu(B(\theta,\varepsilon)) > 0$ for all $\varepsilon > 0$. Thus, one can take $\theta_\varepsilon \in B(\theta,\varepsilon)$ with $\phi_\mu(\theta_\varepsilon) = 0$. As $\theta_\varepsilon \to_{\varepsilon\to 0} \theta$ and by the continuity of ϕ_μ, $\phi_\mu(\theta) = 0$. If $\theta \notin supp(\mu)$, then $\delta_\theta \perp \mu$ and necessarily, $\phi_\mu(\theta) = 0$. The reverse implication is trivial.

Conversely suppose $\phi_\mu = 0$ everywhere and take $\nu \in P_2(\Theta)$, then by previous computations and the convexity of $\mathcal{E}$ (slopes are increasing)

$$0 = \frac{d\,\mathcal{E}(\mu + t(\mu-\nu))}{dt} = \frac{d\,\mathcal{E}(P_\tau(\mu) + tP_\tau(\mu-\nu))}{dt} \leq \mathcal{E}(P_\tau(\nu)) - \mathcal{E}(P_\tau(\mu))$$

which implies that

$$\mathcal{E}_\tau(\mu) \leq \mathcal{E}_\tau(\nu)$$

and μ is optimal.

□

3.3.2 Escape from critical points

In this section, we use the notation :

$$\theta = (a, c, w, b) =: (a, c, \omega)$$

to make the difference between "linear" variables and "nonlinear" ones.

Lemma 10. *For all μ, ν in $\mathcal{P}_2(\Theta)$:*

$$\forall \theta \in \Theta,\ |\phi_\mu(\theta) - \phi_\nu(\theta)| \leq C\left(\int_\Theta |\theta_1|^2 d\mu + \int_\Theta |\theta_2|^2 d\nu\right) W_2^2(\mu,\nu)(1+|\theta|^2)$$

$$\forall \theta \in \Theta,\ |v_\mu(\theta) - v_\nu(\theta)| \leq C\left(\int_\Theta |\theta_1|^2 d\mu + \int_\Theta |\theta_2|^2 d\nu\right) W_2^2(\mu,\nu)(1+|\theta|^2)$$

Proof. Here we focus on v_μ, the proof for ϕ_μ being very similar. Considering (18)-(19), one can decompose v_μ as

$$v_\mu =: v_{\mu,1} + v_2 + v_{\mu,3}.$$

By standard computations and for $\gamma \in \Gamma_o(\mu,\nu)$

$$v_{\mu,1} - v_{\nu,1} = \int_{\Theta^2}\int_\Omega \nabla_\theta \nabla\Phi_\tau(\theta;x)(\nabla\Phi_\tau(\theta_1;x) - \nabla\Phi_\tau(\theta_2;x))dxd\gamma(\theta_1,\theta_2).$$

Owing (27)-(28), one gets

$$\begin{aligned}|v_{\mu,1}(\theta) - v_{\nu,1}(\theta)| &\leq C(\tau)\int_{\Theta^2}\max(|\theta_1|,|\theta_2|)|\theta_1-\theta_1||\theta|^2 dxd\gamma(\theta_1,\theta_2)\\ &\leq C(\tau)\left(\int_\Theta |\theta_1|^2 d\mu + \int_\Theta |\theta_2|^2 d\nu\right) W_2^2(\mu,\nu)|\theta|^2\end{aligned}$$

where we used the Cauchy-Schwartz inequality. For $v_{\mu,3}$, one has :

$$v_{\mu,3} - v_{\nu,3} = \int_{\Theta^2}\int_\Omega \Phi_\tau(\theta_1;\cdot) - \Phi_\tau(\theta_2;\cdot)dxd\gamma \times \int_\Omega \nabla_\theta\Phi_\tau(\theta;\cdot)dx.$$

Owing (20), one gets :

$$\begin{aligned}|v_{\mu,3}(\theta) - v_{\nu,3}(\theta)| &\leq C(\tau)\int_{\Theta^2}\int_\Omega \max(|\theta_1,\theta_2|)|\theta_1-\theta_1|dxd\gamma(\theta_1,\theta_2)|\theta|\\ &\leq C(\tau)\left(\int_\Theta |\theta_1|^2 d\mu + \int_\Theta |\theta_2|^2 d\nu\right) W_2^2(\mu,\nu)|\theta|\end{aligned}$$

where we used again the Cauchy-Schwartz inequality.

□

Proposition 7. *Let $\mu \in \mathcal{P}_2(\Theta)$ such that $\phi_\mu \neq 0$ somewhere. Then there exist a set $A \subset \Theta$ and $\varepsilon > 0$ such that if there exists $t_0 > 0$ with $W_2(\mu_{t_0},\mu) \leq \varepsilon$ and $\mu_{t_0}(A) > 0$, then there exists a time $0 < t_0 < t_1 < +\infty$ such that $W_2(\mu_{t_1},\mu) > \varepsilon$.*

Proof. As ϕ_μ is linear in a and c, it can be written as :

$$\phi_\mu(\theta) =: a\tilde{\phi}_\mu(\omega) + cr_\mu.$$

By hypothesis, the set

$$A_0 := \{\theta \in \Theta \mid \phi_\mu(\theta) \neq 0\}$$

is a non empty (open set). This is equivalent to say that either $\tilde{\phi}_\mu(\omega)$ or r_μ is non zero somewhere. Suppose that $\tilde{\phi}_\mu$ is non zero somewhere, the case for r_μ being similar. Note

$$\begin{cases} A_0^+ = \tilde{\phi}_\mu^{-1}(]0,+\infty[),\\ A_0^- = \tilde{\phi}_\mu^{-1}(]-\infty,0[).\end{cases}$$

Additionally, $\tilde{A}_\alpha$ designates the α sub-level set of $\tilde{\phi}_\mu$. Now we focus on A_0^- and suppose that this set is non empty. The case where A_0^+ is non empty is similar to handle and left to the reader.

By Lemma 11 and the regular value theorem, there exists $\eta > 0$ such that $\partial\tilde{A}_{-\eta} = \tilde{\phi}_\mu^{-1}(\{-\eta\})$ is a $(d+1)-$orientable manifold on which $\nabla_{\tilde{\theta}}\tilde{\phi}_\mu$ is non zero. With our choice of activation function $\sigma_{H,\tau}$, it is easy to prove that $\tilde{A}_{-\eta}$ is a bounded set. Indeed, if b is large enough then $x \to \sigma_{H,\tau}(w\cdot x + b)$ is zero on Ω and $\tilde{\phi}_\mu(w,b)$ is zero.

On $\partial\tilde{A}_{-\eta}$, the gradient $\nabla\tilde{\phi}_\mu$ is pointing outward $\tilde{A}_{-\eta}$ and on $\partial\tilde{A}_{-\eta}$ the outward component $|\nabla_{\tilde{\theta}}\tilde{\phi}_\mu \cdot n_{out}| > \beta$ for some $\beta > 0$ since it is non zero on a compact set. Hence, defining :

$$A := \{(a,c,\omega) \mid \omega \in \tilde{A}_{-\eta}\}$$

and owing that $v_{\mu,a} = \tilde{\phi}_\mu(\omega), v_{\mu,\omega} = a\nabla_\omega\tilde{\phi}_\mu(\omega)$, it holds :

$$\begin{cases} v_{\mu,a} < \eta \text{ on } A \\ v_{\mu,\omega} \cdot n_{out} > \beta \text{ on } \partial \tilde{A}_{-\eta}. \end{cases} \tag{46}$$

By contradiction, suppose that μ_{t_0} has non zero mass on A and that $W_2(\mu, \mu_t) \leq \varepsilon$ (with ε fixed later) for all time $t \geq t_0$. Then using Lemma 10, one has :

$$|v_{\mu_t}(\theta) - v_\mu(\theta)| \leq C(\tau, \mu)(1 + |\theta|^2)\varepsilon \tag{47}$$

and

$$|\phi_{\mu_t}(\theta) - \phi_\mu(\theta)| \leq C(\tau, \mu)(1 + |\theta|^2)\varepsilon.$$

One takes $\varepsilon := \dfrac{\eta}{2C(\tau, \mu)r}$ where r is a solution of :

$$(r-1)\mu_{t_0}(A) > \int |\theta|^2 d\mu + \frac{\eta}{2C(\tau, \mu)r}$$

which exists since $\mu_{t_0}(A) > 0$ by hypothesis. On the set $\{\theta \in A \mid 1 + |\theta|^2 \leq r\}$ and by (47), we have :

$$|v_{\mu_t}(\theta) - v_\mu(\theta)| \leq \frac{\eta}{2}$$

and so by (46) and the fact that $v_t = -v_{\mu_t}$:

$$\begin{cases} v_{t,a} > \eta/2 \text{ on } A \\ v_{t,\omega} \cdot n_{out} < -\beta/2 \text{ on } \partial \tilde{A}_{-\eta}. \end{cases}$$

The general picture is given by Figure 3. As a consequence, there exists a time t_1 such that the set $\{\theta \in A \mid 1 + |\theta|^2 \leq r\}$ has no mass and

$$\int |\theta|^2 d\mu_t(\theta) \geq (r-1)\mu_t(A) \geq (r-1)\mu_{t_0}(A).$$

At the same time, as $W_2(\mu, \mu_t) \leq \varepsilon$:

$$\int |\theta|^2 d\mu_t(\theta) \leq \int |\theta|^2 d\mu(\theta) + \varepsilon = \int |\theta|^2 d\mu(\theta) + \frac{\eta}{2C(\tau, \mu)}$$

and this a contradiction with the definition of r.

□

Remark 7. *The set A constructed in the proof of previous lemma is of the form :*

$$A := \{(a, c, \omega) \mid \omega \in \tilde{A}_{-\eta_1}\} \cup \{(a, c, \omega) \mid \omega \in \tilde{A}^c_{\eta_2}\} \tag{48}$$

where η_1, η_2 are strictly positive.

Lemma 11. *For all $\mu \in P_2(\Theta)$, if $\tilde{\phi}_\mu < 0$ somewhere, there exists a strictly negative regular value $-\eta$ ($\eta > 0$) of $\tilde{\phi}_\mu$.*

Proof. As $\tilde{\phi}_\mu < 0$ somewhere and by continuity, there exists a non empty open $O \subset]-\infty, 0[$ such that $O \subset range(\tilde{\phi}_\mu)$. Next, we use the Sard-Morse theorem recalled below :

Theorem 7 (Sard-Morse)**.** *Let $\mathcal{M}$ be a differentiable manifold and $f : \mathcal{M} \to \mathbb{R}$ of class C^n, then the image of the critical points of f (where the gradient is zero) is Lebesgue negligible in $\mathbb{R}$.*

This result applies to $\tilde{\phi}_\mu$ and the image of critical points of $\tilde{\phi}_\mu$ is Lebesgue negligible. As a consequence, there exists a point $o \in O$ which is a regular value of $\tilde{\phi}_\mu$. As $o \in O$, it is strictly negative and this finishes the proof of the lemma.

□

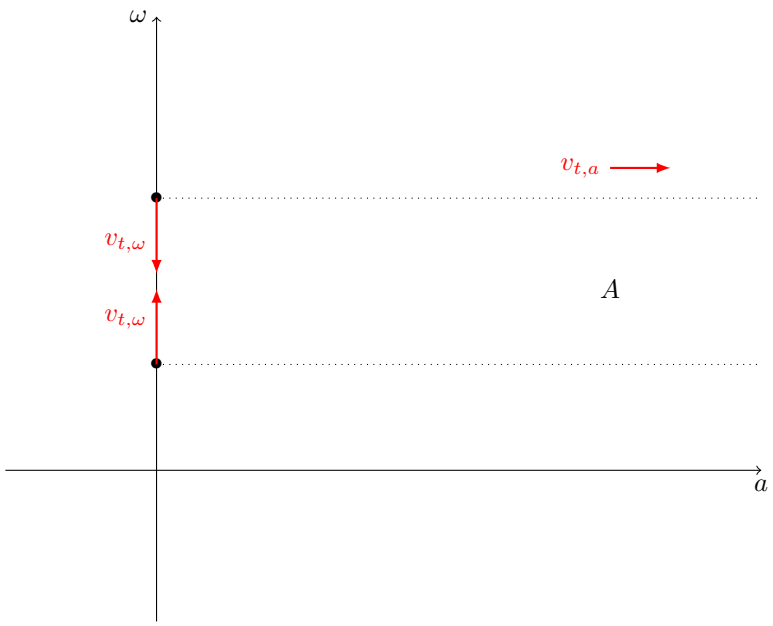


Figure 3: The escape of mass towards large values of a

3.3.3 Convergence

The proof of convergence is based on the following hypothesis on the initial measure μ_0 :

Hypothesis 1. *The support of the measure μ_0 verifies :*

$$\{0\} \times \{0\} \times S_{\mathbb{R}^d}(1) \times [-\sqrt{d}-2, \sqrt{d}+2] \subset supp(\mu_0)$$

This preliminary lemma gives an insight of why Hypothesis 1 is useful :

Lemma 12. *For all $\mu \in \mathcal{P}_2(\Theta)$, $\theta \notin \mathbb{R}^2 \times S_{\mathbb{R}^d}(1) \times]-\sqrt{d}-2, \sqrt{d}+2[, \tau > 1$, the potential writes :*

$$\phi_\mu(\theta) = cr_\mu$$

where r_μ is a constant that depends on μ. In particular, $\phi_\mu(\theta)$ does not depend on a, w, b.

Proof. For all $x \in \Omega, |b| > \sqrt{d}+2, \tau > 1$:

$$|w \cdot x + b| \geq |b| - |x|_\infty |w|_1 > 2$$

and

$$\sigma_{H,\tau}(w \cdot x + b) = 0.$$

This implies that for $|b| \geq \sqrt{d}+2, \mu \in \mathcal{P}_2(\Theta)$, the potential ϕ_μ writes $\phi_\mu = cr_\mu$ where r_μ is a constant. □

In fact Hypothesis 1 is verified by the gradient curve $(\mu_t)_t$ for all time. This is proved in the next lemma.

Lemma 13. *If μ_0 satisfies Hypothesis 1 then for all $t \geq 0$ and all open set $O \subset S_{\mathbb{R}^d}(1) \times [-\sqrt{d}-2, \sqrt{d}+2]$,*

$$\mu_t(\mathbb{R}^2 \times O) > 0$$

The arguments of the proof of last lemma are based on fine tools of algebraic topology. One can find a nice introduction to the topic in the reference book [19]. With simple words, we enjoy the homotopy properties on the sphere to prove that the measure μ_t keeps a large enough support.

Proof. As $\mu_t = (\chi_t)_{\#}\mu_0$, we have [2, Lemma C.8] :

$$\text{supp}(\mu_t) = \overline{\chi_t(\text{supp}(\mu_0))}. \tag{49}$$

Now let $\xi_t(w,b) := (P_{S_{\mathbb{R}^d}(1)\times\mathbb{R}} \circ \chi_t)((0,0,w,b))$ where $P_{S_{\mathbb{R}^d}(1)\times\mathbb{R}}$ is the projection on $S_{\mathbb{R}^d}(1)\times\mathbb{R}$ (w,b variables). We claim that the choice of the function of activation lets the extremal spheres invariant **ie** $\xi_t(w,\pm(\sqrt{d}+2)) = (w,\pm(\sqrt{d}+2))$. Indeed, by Lemma 12 for $\theta = (c,a,w,\pm(\sqrt{d}+2))$, $\phi_\mu(\theta) = cr_\mu$ giving :

$$\begin{cases} v_{\mu,w}(\theta) = & 0, \\ v_{\mu,b}(\theta) = & 0 \end{cases}$$

and the claim is proven. Consequently by Lemma 14, the continuous map ξ_t is surjective.

Now let $O \subset S_{\mathbb{R}^d}(1)\times[-\sqrt{d}-2,\sqrt{d}+2]$ be an open set. By what precedes, the exists a point $\omega \in S_{\mathbb{R}^d}(1)\times[-\sqrt{d}-2,\sqrt{d}+2]$ such that $\xi_t(\omega)\in O$ and $\chi_t((0,0,\omega)) \in \mathbb{R}^2\times O$. As $(0,0,\omega)$ belongs to the support of μ_0 by hypothesis then $\chi_t((0,0,\omega))$ belongs to the support of μ_t by (49) and :

$$\mu_t(\mathbb{R}^2\times O) > 0$$

which finishes the proof of the lemma.

□

Lemma 14 gives conditions for the surjectivity of a continuous map on a cylinder.

Lemma 14. *Let f be a continuous map $f : S_{\mathbb{R}^d}(1)\times[0,1] \to S_{\mathbb{R}^d}(1)\times[0,1] =: C$, homotopic to the identity such that :*

$$\forall w \in S_{\mathbb{R}^d}(1),\ \begin{cases} f(w,0) = (w,0), \\ f(w,1) = (w,1). \end{cases}$$

Then f is surjective.

Proof. Suppose that f misses a point p, then necessarily $p = (w,t)$ with $0 < t < 1$. We can write :

$$g : C \to C \setminus \{p\}$$

the restriction of f on its image. The induced homomorphism on homology groups writes :

$$g_\star : H_{d-1}(C) \to H_{d-1}(C\setminus\{p\}).$$

Aside that, we have the classic information on homology groups of C and $C\setminus\{p\}$:

$$\begin{cases} H_{d-1}(C) = H_{d-1}(S_{\mathbb{R}^d}(1)) & \simeq \mathbb{Z}, \\ H_{d-1}(C\setminus\{p\}) = H_{d-1}(S_{\mathbb{R}^d}(1)\vee S_{\mathbb{R}^d}(1)) & \simeq \mathbb{Z}^2 \end{cases}$$

where $\vee$ designates the wedge sum. Thus, the homomorphism $g_\star$ can be written as :

$$g_\star : \mathbb{Z} \to \mathbb{Z}^2.$$

As g lets the two spheres $w \to (w,0), w \to (w,1)$ invariant, we have :

$$g_\star(1) = (1,1).$$

Now we note $i : C\setminus\{p\} \to C$ the canonical inclusion map. For all $(a,b)\in\mathbb{Z}^2$,

$$i_\star(a,b) = a+b.$$

By hypothesis, f is homotopic to the identity so $f_\star = I_\star$ and $f_\star(1) = 1$ but at the same time :

$$f_\star(1) = i_\star g_\star(1) = i_\star((1,1)) = 2$$

which gives a contradiction.

□

It allows to conclude on the convergence,

Theorem 8. *If μ_0 satisfies Hypothesis 1 and $(\mu_t)_t$ converges towards $\mu^\star \in \mathcal{P}_2(\Theta)$ then $\mu^\star$ is optimal for Problem 1.*

Proof. By contradiction, suppose $\mu^\star$ is not optimal. Then by Lemma 9, $\phi_{\mu^\star} \neq 0$ somewhere. Reusing the separation of variables (see the proof of Proposition 7), $\phi_{\mu^\star}$ writes :

$$\phi_{\mu^\star}(\theta) = a\tilde{\phi}_\mu(w,b) + cr_\mu.$$

Hence either :

- r_μ is not zero and $v_{\mu,c} \neq 0$ and one can prove that some mass escapes at $c = \infty$ as in the proof of Proposition 7.
- $\tilde{\phi}_\mu$ is not identically zero and the set A defined in (48) is not empty and verifies :

$$A \subset \mathbb{R}^2 \times S_{\mathbb{R}^d(1)} \times [-\sqrt{d}-2, \sqrt{d}+2] \tag{50}$$

  by Lemma 12.

We focus on the last item. By Proposition 7, there exists $\varepsilon > 0$ such that if $W_2(\mu_{t_0}, \mu^\star) \leq \varepsilon$ for some t_0 and $\mu_{t_0}(A) > 0$ then there exists a further time t_1 with $W_2(\mu_{t_0}, \mu^\star) > \varepsilon$. As $(\mu_t)_t$ converges towards $\mu^\star$, there exists t_0 such that :

$$\forall t \geq t_0,\ W_2(\mu_{t_0}, \mu^\star) \leq \varepsilon.$$

But by Lemma 13 and (50), for all time $\mu_t(A) > 0$ and consequently there exists a time $t_1 > t_0$ with :

$$W_2(\mu_{t_0}, \mu^\star) > \varepsilon$$

which gives the contradiction.

□

4 Numerical experiments

In this section we will conduct numerical experiments to evaluate, in a concrete way, the potential of the method proposed.

4.1 The effect of frequency

First, the influence of the frequency on the approximation is investigated. To do so, we consider $d = 1$ and the following source term for which the solution is a cosinus mode :

$$f_k(x) := \pi^2 |k|^2 \cos(\pi k \cdot x).$$

In higher dimension, we use the corresponding source term which is a tensor product of its one dimensional counterpart :

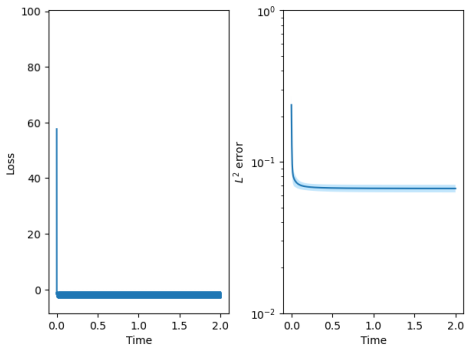
$$f_k(x_1, \cdots, x_d) := \pi^2 |k|_{l^2}^2 \cos(\pi k_1 \cdot x_1) \cdots \cos(\pi k_d \cdot x_d).$$

The code is written using python supplemented with Keras/Tensorflow framework. One should remember the following implementation facts :

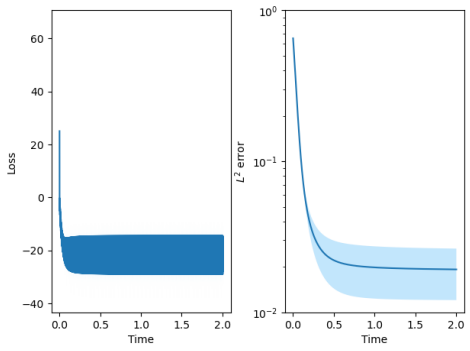
- The neural network represents the numerical approximation taking values of $x \in \Omega$ as input and giving a real as output.
- The loss function is approximated with a Monte Carlo sampling for the integrals where the measure is uniform on Ω. For each training phase, we use batches of size 10^2 obtained from a dataset of 10^5 samples, the number of epochs is calculated to have a time of optimization equals to 2 (learning rate $\times$ number steps $= 2$). Note that the dataset is shuffled at each epoch.

- The derivative involved in the loss is computed thanks to automatic differentiation.
- The training routine is given by the algorithm of backpropagation coupled with a gradient descent optimizer for which the learning rate $\zeta := \dfrac{1}{2nm}$ where n is the batch size and m is the width of the neural network involved. This choice will be explained later in the analysis.
- In all the plots, the reader will see the mean curve and a shaded zone representing the interval whose width is two times the standards deviation. Each simulation is run 4 times to calculate these statistical parameters.

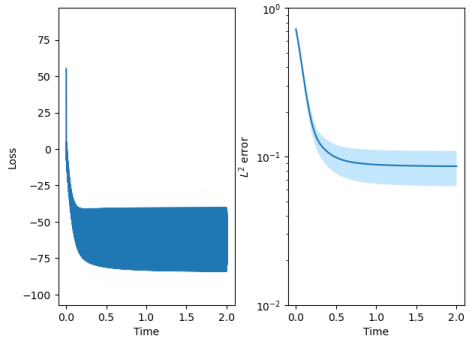
For $d = 1$ and a width $m = 1000$, the simulations are reported in Figure 4 for which very satisfactory results for $k = 1, 3$ are observed, the same conclusions hold for $d = 2$.



(a) The case $d = 1$ and $k = (1)$

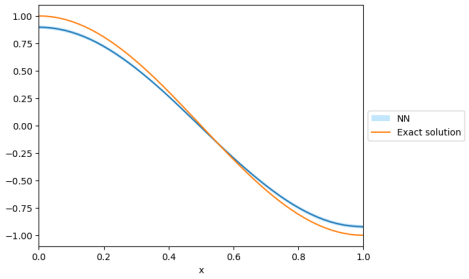


(b) The case $d = 1$ and $k = (3)$

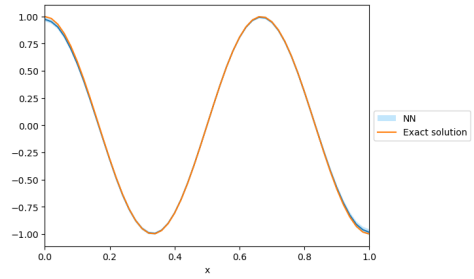


(c) The case $d = 1$ and $k = (5)$

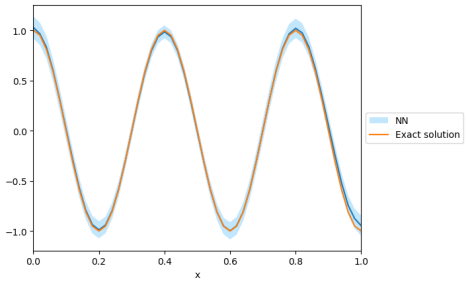
Figure 4: The effect of frequency on the approximation when $d = 1$ and $m = 1000$



(a) The case $d = 1$ and $k = (1)$

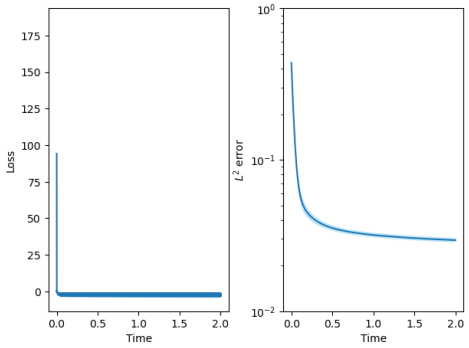


(b) The case $d = 1$ and $k = (3)$

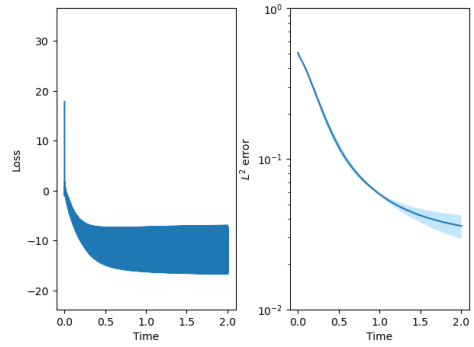


(c) The case $d = 1$ and $k = (5)$

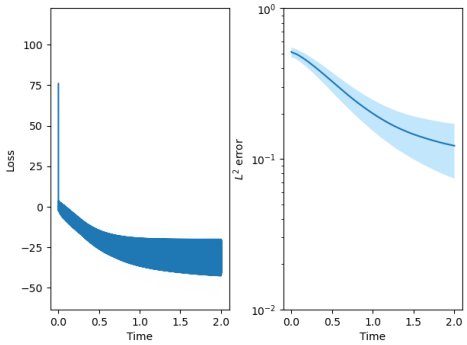
Figure 5: The numerical solutions when $d = 1$ and $m = 1000$



(a) The case $d = 2$ and $k = (1, 1)$



(b) The case $d = 2$ and $k = (3, 1)$



(c) The case $d = 2$ and $k = (5, 1)$

Figure 6: The effect of frequency on the approximation when $d = 2$

Remark 8. *In this remark, we expose not rigorous heuristic arguments for the scaling related to the*

learning rate :

$$\xi := \frac{1}{2nm}.$$

It is possible to write the learning scheme as follows :

$$\frac{\theta_{t+1} - \theta_t}{dt} = -\nabla_\theta \phi^n_{\mu_t^m}(\theta_t) \tag{51}$$

where :

$$\phi^n_{\mu_t^m}(\theta) := \frac{1}{nm} \sum_{i,j} \nabla\Phi(\theta_j, x_i) \cdot \nabla\Phi(\theta, x_i) - f(x_i)\Phi(\theta, x_i) + \left(\frac{1}{nm} \sum_{i,j} \Phi(\theta, x_i) \right)^2 \tag{52}$$

with $(x_i)_i$ are n samples taken uniformly on the d dimensional cube.

By analogy, equations (51)-(52) can be interpreted as an explicit finite element scheme for the heat equation where the space discretization parameter is $h := \frac{1}{\sqrt{nm}}$. This gives the CFL condition :

$$2dt \leq h^2$$

which is equivalent to :

$$dt \leq \frac{1}{2nm}.$$

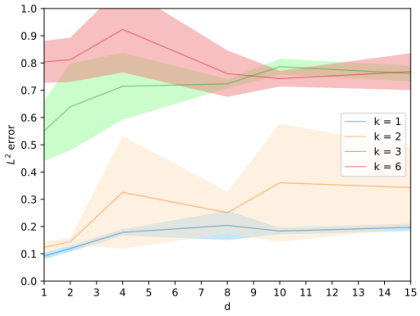
In practice, one can observe that if one takes $dt > O\left(\frac{1}{nm}\right)$ then the scheme diverges in the same way as a classic finite elements scheme.

The CFL condition is bad news since it prevents the use of large batch sizes necessary to get a good precision. In practice, the maximum on can do with a standard personal computer is $n, m = 10^2$.

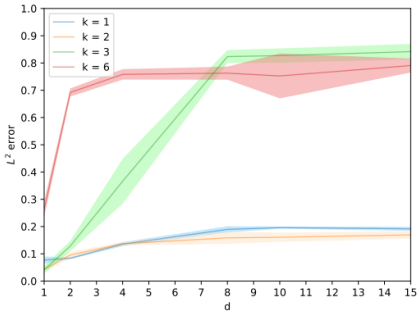
4.2 The effect of dimension

To evaluate the effect of dimension on performance, we consider frequencies of the form $k = (\bar{k}, 0, \cdots, 0)$ where $\bar{k}$ is an integer, and plot the L^2 error as a function of the dimension for different $\bar{k}$. This is done in Figure 7 where several observations can be made :

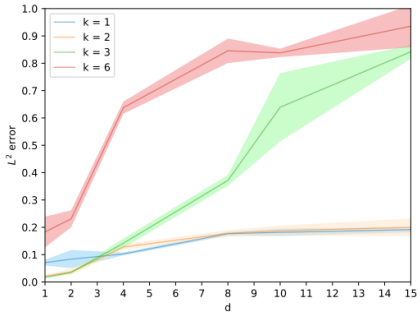
- For low frequency, the precision is not affected by dimension.
- At high frequency, performance are deteriorated as dimension increases.
- Having a larger neural network captures better high frequency modes up to a certain dimension.
- Variance increases with frequency but not with dimension.



(a) $m = 10$



(b) $m = 100$



(c) $m = 1000$

Figure 7: The effect of dimension for different frequencies and width

For completeness we plot in Figure 8 a high dimensional example where $d = 10$, $k = (1, 1, 0, \cdots, 0)$ to show that the proposed method works well in the high dimensional/low frequency regime. The contour plot shows the function's values on the slice $(x_1, x_2, 0.5, \cdots, 0.5)$.

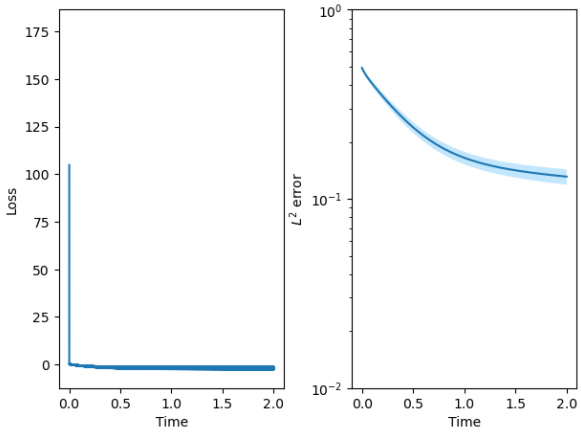


Figure 8: The case $d = 10$, $k = (1, 1, 0, \cdots, 0)$ and $m = 1000$

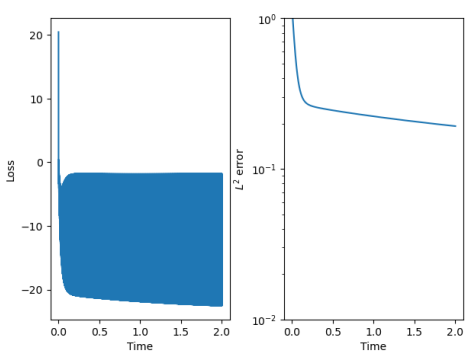
Finally we show an example where a lot of low frequencies are involved in the high dimensional regime :

$$f(x) = 2\pi^2 \sum_{k=1}^{d-1} \cos(\pi \cdot x_k) \cos(\pi \cdot x_{k+1})$$

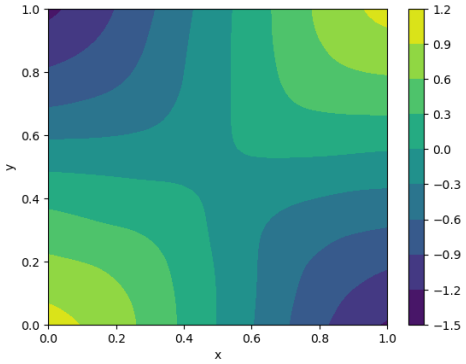
whose solution is :

$$u^\star(x) = \sum_{k=1}^{d-1} \cos(\pi \cdot x_k) \cos(\pi \cdot x_{k+1}).$$

For $d = 6$, $m = 1000$ and all other parameters being identical to previous cases, one gets convergence of the solution on Figure 9 where the contour plot still shows the function's values on the slice $(x_1, x_2, 0.5, \cdots, 0.5)$.



(a) The loss and the L^2 error



(b) The numerical solution

Figure 9: The mixed mode solution

5 Conclusion

In this article, the ability of two layers neural networks to solve Poisson equation is investigated. First the PDE problem commonly understood in the Sobolev sense, is reinterpreted in the perspective of probability measures by writing the energy functional as a function over probabilities. Then, we propose to solve the obtained minimization problem thanks to gradient curves for which an existence result is shown. To justify this choice of method, the convergence towards an optimal measure is proved assuming the convergence of the gradient curve. Finally, numerical illustrations with a detailed analysis of the effects of dimension and frequency are presented. With this work, it becomes clear that neural networks is a viable method to solve Poisson equation even in the high dimensional regime; something out of reach for classical methods. Nonetheless, some questions and extensions deserve more detailed developments. First, the main remark to observe is that the convergence is not proved theoretically even if it is observed in practice. Additionally, the domain considered is very peculiar $\Omega = [0,1]^d$ and it is not obvious that one could generalize such theory on domain where sin/cosine decomposition is not available. In numerical illustrations, integrals involved in the cost were not computed exactly but approximated by uniform sampling. It should be interesting to study the convergence of gradient curves with respect to the number of samples.

A The differential structure of Wasserstein spaces over compact Alexandrov spaces

The aim of this section is to get acquainted of the differential structure of $\mathcal{P}_2(\Theta)$. All the results presented here are not rigorously proved and we rather give a didactic introduction to the topic, the main reference being [16].

A.1 The differential structure of Alexandrov spaces

An Alexandrov space (A, d) is a geodesic space embedded with its distance d having a nice concave property on triangles. With big words, Alexandrov spaces are space where the curvature is bounded from below by a uniform constant. Before going further, we need to introduce some notation :

Definition 6. *Let α be a unit speed geodesic with $\alpha(0) = a \in A$ and $s \geq 0$, then we introduce the notation :*

$$(\alpha, s) := t \to \alpha(st)$$

the associated geodesic of velocity s. The space of directions is the space unit speed geodesics α and it is denoted $\Sigma_a(A)$. The tangent cone ie the set of geodesics departing from a at speed s, of the form (α, s) is denoted $C_a(A)$. We make the following correspondence :

$$\text{“}(\alpha, 1) = \alpha\text{”}$$

for α in $\Sigma_a(A)$.

It is not so important to focus on a rigorous definition of such spaces but one should remember the following fundamental property of existence of a tangential cone structure :

Theorem 9. *Let α, β be two unit speed geodesics and $s, t \geq 0$ with $\alpha(0) = \beta(0) =: a \in A$. Then the limit :*

$$\sigma_a((\alpha, s), (\beta, t)) := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} d(\alpha(s\varepsilon), \beta(t\varepsilon))$$

exists. Moreover,

$$\frac{1}{2st}\left(s^2 + t^2 - \sigma_a((\alpha, s), (\beta, t))\right) \tag{53}$$

does not depend on s, t.

Previous theorem is very important as it introduces a notion angle and scalar product :

Corollary 2. *One can define the local angle between $(\alpha, s), (\beta, t)$ by :*

$$\cos(\angle_a((\alpha, s), (\beta, t))) := \frac{1}{2st}\left(s^2 + t^2 - \sigma_a((\alpha, s), (\beta, t))\right)$$

and a local scalar product :

$$\langle (\alpha, s), (\beta, t)) \rangle_a := st \cos(\angle_a((\alpha, s), (\beta, t))).$$

Remark 9. *In fact, the space of directions $\Sigma_a(A)$ is the completion of :*

$$\{(\alpha, 1) \mid \alpha \text{ unit speed geodesic departing from } a\}$$

quotiented by the relation $\sigma_a = 0$ wrt to the distance σ_a. The same is true for the tangent cone $C_a(A)$.

A major result from [16] is that if the underlying space A is Alexandrov then the space over probabibilty $\mathcal{P}_2(A)$ is also an Alexandrov space and all the differential structure presented above is available. The proof of this result is based on MCann interpolation which allows to make the link between probability geodesics and geodesics of the underlying space.

Moreover, it is possible to define a notion of differentiation ;

Definition 7. *For a curve $(a_t)_t$ of A, it is said to be differentiable at $t = 0$ if there exists $(\alpha, \tau) \in C_a(A)$ such that for all $\alpha_i \in \Sigma_a(A), t_i \geq 0$ with $\lim_{i \to \infty} t_i = 0$, linking a_0 and a_{t_i} then :*

$$\lim_{i \to \infty} (\alpha_i, d(a_0, a_{t_i})/t_i) = (\alpha, t)$$

where the convergence has to be understood in the sense of the distance σ_a. Moreover, the derivative of the curve at $t = 0$ writes :

$$a_0' := (\alpha, t).$$

A.2 The notion of gradient

Now an energy $\mathcal{E} : A \to \mathbb{R}$ is introduced with the following property of convexity.

Definition 8. *We say that $\mathcal{E}$ is convex along geodesics if there exists $K \in \mathbb{R}$ such that for all rescaled geodesics $\alpha : [0,1] \to A$:*

$$\mathcal{E}(\alpha(\lambda)) \leq (1-\lambda)\,\mathcal{E}(\alpha(0)) + \lambda\,\mathcal{E}(\alpha(1)) - \frac{K}{2}\lambda(1-\lambda)d(\alpha(0),\alpha(1)).$$

Assuming such convexity, it is possible to define the gradient's direction of $\mathcal{E}$ using the differential structure of A (see [16, Lemma 4.3]). Before doing this, it is necessary to introduce the directional derivative :

Definition 9. *For $a \in A$ and $(\alpha, s) \in C_a(A)$, one defines :*

$$D_a\,\mathcal{E}((\alpha,s)) := \lim_{\varepsilon \to 0} \frac{\mathcal{E}(\alpha(s\varepsilon)) - \mathcal{E}(\alpha(0))}{\varepsilon}.$$

One can prove that the limit above exists using the convexity assumption of $\mathcal{E}$. Owing this, there exists a direction for which the local slope (see Definition 4) is attained in the sense defined below.

Theorem 10. *For all $a \in D(\mathcal{E})$ such that $|\nabla_-\,\mathcal{E}\,|(a) < \infty$, there exists a unique direction $\alpha \in \Sigma_a(A)$ such that :*

$$D_a\,\mathcal{E}((\alpha,1)) = -|\nabla_-\,\mathcal{E}\,|(a).$$

This direction α is denoted :

$$\frac{\nabla_-\,\mathcal{E}(a)}{|\nabla_-\,\mathcal{E}\,|(a)}$$

which means that :

$$D_a\,\mathcal{E}((\alpha, |\nabla_-\,\mathcal{E}\,|(a))) := -|\nabla_-\,\mathcal{E}\,|^2(a).$$

With this, it is straightforward to define the notion of gradient curve.

Definition 10. *A Lipschitz curve $(a_t)_t$ is said to be a gradient curve wrt $\mathcal{E}$ if it is differentiable for all $t \geq 0$ and :*

$$\forall t \geq 0,\ a_t' = \left(\frac{\nabla_-\,\mathcal{E}(a_t)}{|\nabla_-\,\mathcal{E}\,|(a_t)}, |\nabla_-\,\mathcal{E}\,|(a_t)\right) \in C_{a_t}(A).$$

In [16], results about existence and uniqueness of gradient curve on $\mathcal{P}_2(A)$ are given.

References

[1] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945, 1993.

[2] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport, 2018.

[3] F. Bach and L. Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization, 2021.

[4] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[5] M. Minsky and S. Papert. *Perceptrons; an Introduction to Computational Geometry.* MIT Press, 1969.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[7] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[8] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, 2(4):303–314, 1989.

[9] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

[10] W. E, C. Ma, and L. Wu. The Barron space and the flow-induced function spaces for neural network models. *Constr. Approx.*, 55(1):369–406, 2022.

[11] W. E and S. Wojtowytsch. Representation formulas and pointwise properties for Barron functions. *Calc. Var. Partial Differential Equations*, 61(2):37–46, 2022.

[12] W. E and S. Wojtowytsch. On the Banach spaces associated with multi-layer relu networks: Function representation, approximation theory and gradient descent dynamics. *CSIAM Transactions on Applied Mathematics*, 1(3):387–440, 2020.

[13] Y. Lu, J. Lu, and M. Wang. A priori generalization analysis of the deep ritz method for solving high dimensional elliptic partial differential equations. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3196–3241. PMLR, 15–19 Aug 2021.

[14] J. Lott and C. Villani. Ricci curvature for metric-measure spaces via optimal transport. *Ann. of Math. (2)*, 169(3):903–991, 2009.

[15] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures.* Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[16] S.-I. Ohta. Gradient flows on Wasserstein spaces over compact Alexandrov spaces. *Amer. J. Math.*, 131(2):475–516, 2009.

[17] M. Erbar. The heat equation on manifolds as a gradient flow in the Wasserstein space. *Ann. Inst. Henri Poincaré Probab. Stat.*, 46(1):1–23, 2010.

[18] J. M. Lee. *Riemannian manifolds*, volume 176 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1997. An introduction to curvature.

[19] A. Hatcher. *Algebraic topology.* Cambridge University Press, Cambridge, 2002.