Numerical solution of Poisson partial differential equation in high dimension using two-layer neural networks

Mathias Dus, Virginie Ehrlacher

Numerical solution of Poisson partial differential equation in high dimension using two-layer neural networks

Dus Mathias, Ehrlacher Virginie

July 13, 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 partial differential equation (PDE) with Neumann boundary condition. Using Barron's representation of the solution [1] with a probability measure defined on the set of parameter values, the energy is minimized thanks to a gradient curve dynamic on the 2-Wasserstein space of the set of parameter values 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

The motivation of our work is to bring some contributions on the mathematical understanding of neural-network based numerical schemes, typically Physically-Informed-Neural-Networks (PINNs) [4, 5, 6, 7, 8, 9] approaches, for the resolution of some high-dimensional Partial Differential Equations (PDEs). In this context, it is of tremendous importance to understand why neural networks work so well in some 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 mathematical analysis theory of neural network-based numerical methods 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 [10] 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 [11] 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 [12], 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, a mix of 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 arise :

- 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 [13], 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 [14], 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 [15] where authors considered the Poisson problem with Neumann boundary condition 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 [15], the object of this paper is to propose and analyze a neural-network based numerical approach for the resolution of the Poisson equation in the high dimensional regime with Barron source. Inspired from [2], we immerse the problem on the space of probability measures with finite second order moments defined on the parametric domain. This corresponds to finding a solution to the PDE thanks to an infinitely wide two-layer neural network. 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 necessarily does towards a measure corresponding to the solution of the PDE considered. Note that our argumentation is different from [2, 3] since the convergence proof is not based on topological degree nor the topological properties of the sphere. We rather use a homology argument taken from algebraic topology and a clever choice of activation function to prove that the dynamic of the support of the gradient curve of measure behaves nicely. Numerical experiments are conducted to confirm the potential of the method proposed.

In Section 2, the problem is presented in a more precise way and the link between probability and Barron functions is made clearly. In Section 3, the gradient curve is introduced and our main theorems on its well-posedness and convergence are presented and proved. Finally, numerical experiments are exposed in Section 4.

Notation : For $1 \leq p \leq \infty$, the notation $|\cdot|_p$ designates the ℓ^p norm of a vector of arbitrary finite dimension with particular attention to $p = 2$ (euclidean norm) for which the notation $|\cdot|$ is preferred.

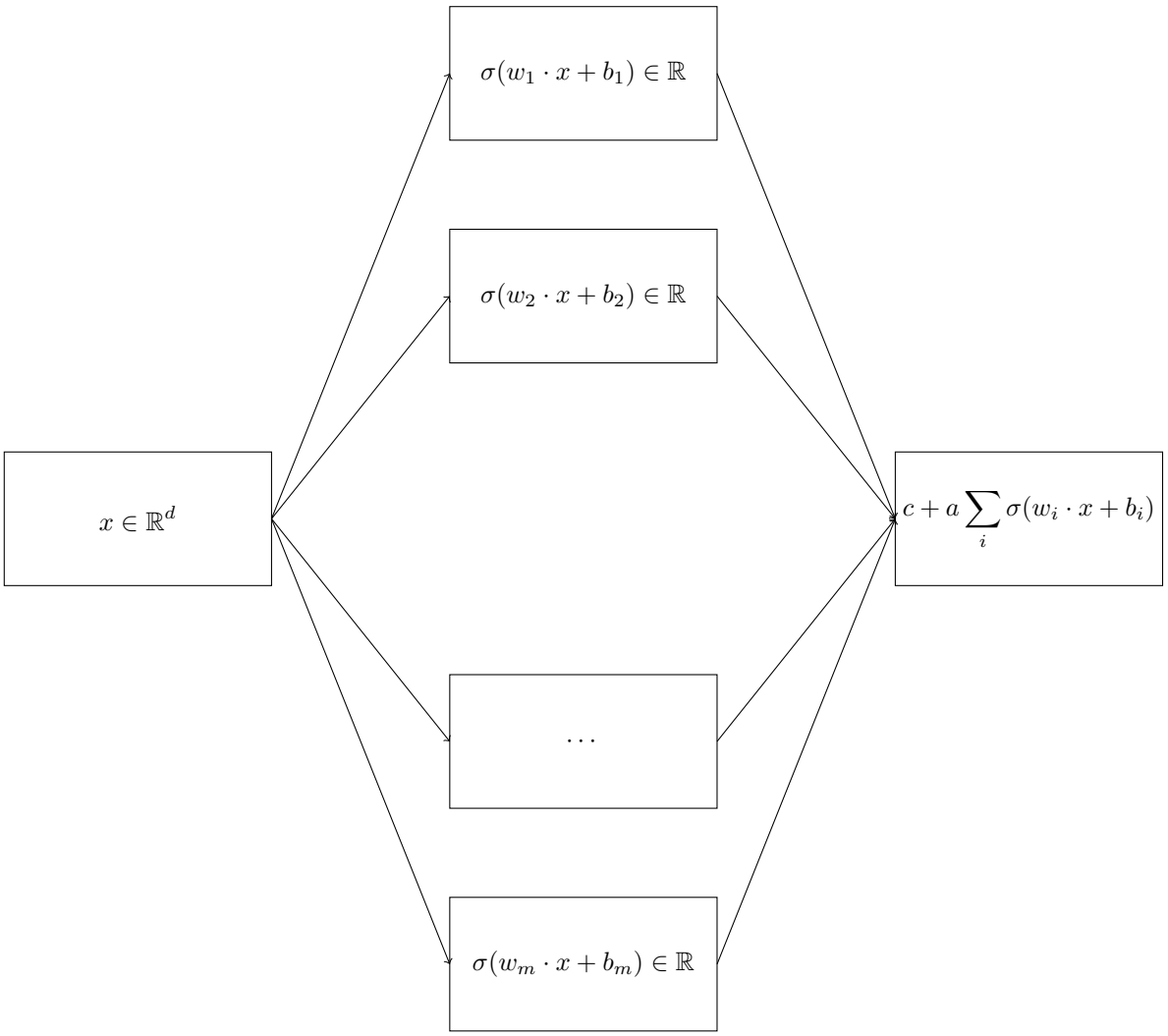


Figure 1: A two-layer neural network of width m

2 Preliminaries

This section introduces the mathematical framework we consider in this paper to relate two-layer neural networks and high-dimensional Poisson equations.

2.1 Problem setting

The following Poisson equation is considered on $\Omega := [0,1]^d$ ($d \in \mathbb{N}$) with Neumann boundary condition : 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 $\bar{a}$ involved in the definition of the energy writes :*

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

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

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

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 ans so is 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,

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

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,

$$\|\sigma' - \sigma'_\tau\|^2_{L^2(\mathbb{R})} = \int_{-1/\tau}^{1/\tau} (\sigma'(x) - \sigma'_\tau(x))^2 dx \leq \frac{4}{\tau}.$$

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 $[0,1]$. Note however that neither $\sigma_{H,\tau}$ nor σ_H are homogeneous (in contrast to the activation functions considered in [2, 3]). Notice also that a direct corollary of Lemma 2 is that there exists a constant $C > 0$ such that for all $\tau > 0$,

$$\|\sigma_H - \sigma_{H,\tau}\|_{H^1(\mathbb{R})} \leq \frac{C}{\sqrt{\tau}} \tag{9}$$

We will also use the fact that there exists a constant $C > 0$ such that for all $\tau > 0$,

$$\|\sigma_{H,\tau}\|_{L^\infty(\mathbb{R})} \leq C,\ \|\sigma'_{H,\tau}\|_{L^\infty(\mathbb{R})} \leq C,\ \|\sigma''_{H,\tau}\|_{L^\infty(\mathbb{R})} \leq C\tau \text{ and } \|\sigma'''_{H,\tau}\|_{L^\infty(\mathbb{R})} \leq C\tau^2. \tag{10}$$

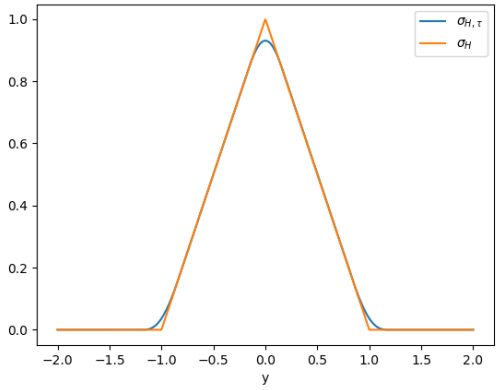


Figure 2: The hat activation function and its regularization ($\tau = 4$)

2.3 Spectral Barron space

We introduce 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

$$\forall k = (k_1, \dots, k_d) \in \mathbb{N}^d,\ \forall x := (x_1, \cdots, x_d) \in \Omega, \quad \phi_k(x_1, \dots, x_d) := \prod_{i=1}^{d} \cos(\pi k_i x_i). \tag{11}$$

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 [15] as follows :

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

$$\mathcal{B}^s(\Omega) := \left\{ u \in L^1(\Omega) : \sum_{k\in\mathbb{N}^d} (1 + \pi^s |k|_1^s)|\hat{u}(k)| < +\infty \right\} \tag{12}$$

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

$$\|u\|_{\mathcal{B}^s(\Omega)} := \sum_{k\in\mathbb{N}^d} (1 + \pi^s |k|_1^s)|\hat{u}(k)|. \tag{13}$$

By [15, 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:*

$$\mathcal{F}_{\chi,m}(B) := \left\{ c + \sum_{i=1}^{m} a_i \chi(w_i \cdot x + b_i) : c, a_i, b_i \in \mathbb{R},\ w_i \in \mathbb{R}^d,\ |c| \le 2B, |w_i| = 1, |b_i| \le 1, \sum_{i=1}^{m} |a_i| \le 4B \right\} \tag{14}$$

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 :

$$\|u - \tilde{u}_m\|_{H^1(\Omega)} \le \frac{C\|u\|_{\mathcal{B}(\Omega)}}{\sqrt{m}}. \tag{15}$$

where for both items, 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 [15, 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, we assume that any parameter vector $\theta = (c, a, w, b)$ takes values in the neural network parameter set

$$\Theta := \mathbb{R} \times \mathbb{R} \times S_{\mathbb{R}^d}(1) \times [-\sqrt{d} - 2, \sqrt{d} + 2], \tag{16}$$

with $S_{\mathbb{R}^d}(1)$ the unit sphere of $\mathbb{R}^d$. In addition, for all $r > 0$, we denote by

$$K_r := [-2r, 2r] \times [-4r, 4r] \times S_{\mathbb{R}^d}(1) \times [-\sqrt{d} - 2, \sqrt{d} + 2]. \tag{17}$$

The particular choice of the domain value of the parameter b, namely $[-\sqrt{d} - 2, \sqrt{d} + 2]$ will be made clear in the following. Moreover, let $\mathcal{P}_2(\Theta)$ (respectively $\mathcal{P}_2(K_r)$) denote the set of probability measures on Θ (respectively on K_r) with finite second-order moments.

Let us make the following remark.

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 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. In addition, the sequence $(\mu_m)_{m \in \mathbb{N}^*}$ is uniformly (with respect to 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. *[15] 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

$$\forall \theta := (c, a, w, b) \in \Theta, \ \forall x \in \Omega, \quad \Phi_\tau(\theta; x) := c + a \sigma_{H,\tau}(w \cdot x + b) \tag{18}$$

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

$$\forall \theta := (c, a, w, b) \in \Theta, \ \forall x \in \Omega, \quad \Phi_\infty(\theta; x) := c + a \sigma_H(w \cdot x + b). \tag{19}$$

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

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

  where the last inequality comes from (10).

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 $\mathcal{P}_2(K_r)$, 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 compactness of $\mathcal{P}_2(K_r)$ (as K_r is compact) 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*

$$\mu^\star_{\tau,r} \in \underset{\mu\in\mathcal{P}_2(\Theta)}{\operatorname{argmin}} \mathcal{E}_{\tau,r}(\mu). \tag{20}$$

For large values of τ and $r = d\|f\|_{\mathcal{B}^0(\Omega)}$, solutions of (20) 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 (20) with $\tau = m$ and $r = d\|f\|_{\mathcal{B}^0(\Omega)}$, it holds that:*

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

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 (15) 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 (20) 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 :

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

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 $\bar{a}$, we obtain that

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

□

2.5 Main results

In this section, we find a solution to 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 the convergence necessarily holds towards a global minimizer. In all the sequel, we fix an a prioori chosen value of $\tau > 0$.

2.5.1 Well-posedness

First, we introduce the concept of gradient curve which formally writes for $r > 0$:

$$\forall t \geq 0,\ \frac{d}{dt}\mu^r(t) = -\nabla\, \mathcal{E}_{\tau,r}(\mu^r(t)). \tag{21}$$

Equation (21) has no mathematical sense since the space $\mathcal{P}_2(\Theta)$ is not a Hilbert space and consequently, the gradient of $\mathcal{E}_{\tau,r}$ is not available in a classical sense. Nevertheless $\mathcal{P}_2(\Theta)$ being an Alexandrov space, it has a differential structure which allows to define gradients properly. The careful reader wishing to understand this structure can find a complete recap of all useful definitions and properties of Alexandrov spaces in Appendix A.

Before exposing our main results of well-posedness, we recall the basic definition of local slope [16]. In the sequel, we denote by $\mathcal{P}_2(K_r)$ the set of probability measures on Θ with support included in K_r.

Definition 4. *At every $\mu \in \mathcal{P}_2(K_r)$, the local slope writes :*

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

which may be infinite.

In Section 3.1, we prove two theorems; the first one states the existence and the uniqueness of the gradient curve with respect to $\mathcal{E}_{\tau,r}$ when $r < \infty$.

Theorem 4. *For all $\mu_0 \in \mathcal{P}_2(K_r)$, there exists a unique locally Lipschitz gradient curve $\mu^r : \mathbb{R}_+ \to \mathcal{P}_2(K_r)$ which is also a curve of maximal slope with respect to the upper gradient $|\nabla^-\, \mathcal{E}_{\tau,r}\,|$. Moreover, for almost all $t \geq 0$, there exists a vector field $v_t^r \in L^2(\Theta;\, d\mu^r(t))^{d+3}$ such that*

$$\int_\Theta \|v_t^r\|^2\, d\mu^r(t) = \|v_t^r\|^2_{L^2(\Theta;\, d\mu^r(t))} < +\infty \tag{22}$$

and :

$$\begin{cases} \partial_t \mu^r(t) + \operatorname{div}(v_t^r \mu^r(t)) = & 0 \\ \mu^r(0) = & \mu_0 \\ \mu^r(t) \in & \mathcal{P}_2(K_r). \end{cases} \tag{23}$$

In the second theorem, we focus on the case when $r = +\infty$ for which we formally take the limit of gradient curves $(\mu^r)_{r>0}$ as r goes to infinity. Introducing the following quantities, the definition of which will be made precise below :

$$\begin{cases} \phi_\mu(\theta) := & d\,\mathcal{E}\,|_{P_\tau(\mu)}(\Phi_\tau(\theta;\cdot)), \\ v_\mu(\theta) := & \nabla_\theta \phi_\mu(\theta), \end{cases}$$

and $\mathbf{P}$ the projection on the tangent bundle of Θ the precise definition of which is given in Definition 5, the following theorem is proved.

Theorem 5. *For all μ_0 compactly supported, there exists a curve $\mu : \mathbb{R}_+ \to \mathcal{P}_2(\Theta)$ such that :*

$$\begin{cases} \partial_t \mu(t) + \operatorname{div}((-\mathbf{P} v_{\mu(t)})\mu(t)) = 0 \\ \mu(0) = \mu_0 \end{cases} \tag{24}$$

and for almost all $t \geq 0$:

$$\int_\Theta |\mathbf{P} v_{\mu(t)}|^2 \, d\mu(t) = \|\mathbf{P} v_{\mu(t)}\|^2_{L^2(\Theta; d\mu(t))} < +\infty.$$

Moreover, the solution satisfies :

$$\forall t \geq 0, \mu(t) = \chi(t) \# \mu_0$$

with $\chi : \mathbb{R}_+ \times \Theta \to \Theta$ solution to

$$\begin{cases} \partial_t \chi(t;\theta) = -\mathbf{P} v_{\mu(t)}(\theta) \\ \chi(0;\theta) = \theta. \end{cases}$$

In Remark 6, we argue why proving the existence and uniqueness of a gradient curve for $\mathcal{E}_{\tau,\infty}$ is not reachable. This is why $\mu : \mathbb{R}_+ \to \mathcal{P}_2(\Theta)$ is described as a limiting gradient curve and not a gradient curve itself in Theorem 5.

2.5.2 Link with neural network

Our motivation for considering the analysis presented in the previous section is that we can link the learning phase of a neural network with the optimization procedure given by gradient curves defined above. Indeed, let $m > 0$ be an integer. A two-layer neural network u with $\sigma_{H,\tau}$ as activation function can always be written as :

$$u = \frac{1}{m} \sum_{i=1}^{m} \Phi_\tau(\theta_i, \cdot) \tag{25}$$

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

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

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

$$\nabla_{\theta_i} \mathcal{F}(\theta_1, \cdots, \theta_m) = \frac{1}{m} \nabla_\theta \phi_\mu(\theta_i)$$

where :

$$\mu := \frac{1}{m} \sum_{i=1}^{m} \delta_{\theta_i} \in \mathcal{P}_2(\Theta). \tag{26}$$

As a consequence, a gradient descent of $\mathcal{F}$ in the sense that, for all $1 \leq i \leq m$,

$$\begin{cases} \frac{d}{dt}\theta_i(t) = -m\mathbf{P}\nabla_{\theta_i} \mathcal{F}(\theta_1(t), \cdots, \theta_m(t)) \\ \theta_i(0) = \theta_{i,0}, \end{cases}$$

which is equivalent to the gradient curve of $\mathcal{E}_{\tau,+\infty}$ with initial condition given by

$$\mu_{0,m} := \frac{1}{m} \sum_{i=1}^{m} \delta_{\theta_{i,0}}. \tag{27}$$

Theorem 6. *Let $\mu_0 \in \mathcal{P}_2(\Omega)$ compactly supported, $(\mu_{0,m})_{m\in\mathbb{N}^*}$ be such that for all $m \in \mathbb{N}^*$, $\mu_{0,m}$ is of the form (27) for some $(\theta_{i,0})_{1\leq i\leq m} \subset \mathrm{Supp}(\mu_0)$ and $\lim_{m\to+\infty} W_2(\mu_{0,m}, \mu_0) = 0$.*

Let $\mu : \mathbb{R}_+ \to \mathcal{P}_2(\Theta)$ and $\mu_m : \mathbb{R}_+ \to \mathcal{P}_2(\Theta)$ be the gradient curves constructed in Theorem 5 associated respectively to the initial conditions $\mu(0) = \mu_0$ and $\mu_m(0) = \mu_{0,m}$. Then for all $T > 0$, there exists a constant $C_T > 0$ such that

$$\sup_{0\leq t\leq T} W_2(\mu(t), \mu_m(t)) \leq C_T W_2(\mu_0, \mu_{0,m}).$$

This theorem is proved in Section 3.2.

2.5.3 Convergence

Our convergence result towards a global optimum is based on the following hypothesis on the initial measure μ_0 :

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

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

Under such hypothesis, one gets a result of convergence in the spirit of a previous work from Bach and Chizat [2] :

Theorem 7. *If μ_0 satisfies Hypothesis 1 and $\mu(t)$ converges towards $\mu^\star \in \mathcal{P}_2(\Theta)$ as t goes to infinity, then $\mu^\star$ is optimal for Problem 1.*

This theorem is proved in Section 3.3.

3 Gradient curve

This section is dedicated to the proof of the two theorems stated in Section 2.5.

3.1 Well-posedness

3.1.1 Proof of Theorem 4

Let us fix some value of $r > 0$ in this section. In the following, $C > 0$ will denote an arbitrary constant which does not depend on τ and r. Let $\mathfrak{P}$ be the set of geodesics of Θ *ie* the set of absolutely continuous curves $\pi : [0, 1] \to \Theta$ such that for all $t_1, t_2 \in [0, 1]$, $d(\pi(t_1), \pi(t_2)) = d(\pi(0), \pi(1))|t_1 - t_2|$. Besides, it holds that for all $0 \leq t \leq 1$, we have $|\dot\pi(t)| = d(\pi(0), \pi(1))$.

For all $s \in [0, 1]$, we define the application map $e_s : \mathfrak{P} \to \Theta$ 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. *[17, Proposition 2.10] For all $\mu, \nu \in \mathcal{P}_2(\Theta)$ and any geodesic $\kappa : [0, 1] \to \mathcal{P}_2(\Theta)$ between them (i.e. such that $\kappa(0) = \mu$ and $\kappa(1) = \nu$) in the W_2 sense, there exists $\Pi \in \mathcal{P}_2(\mathfrak{P})$ such that :*

$$\forall t \in [0, 1],\ \kappa(t) = e_t \# \Pi.$$

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. In addition, it holds that $\gamma := (e_0, e_1)\#\Pi$ is then an optimal transport plan between μ and ν for the quadratic cost, i.e. $W_2(\mu,\nu)^2 = \int_{\Theta\times\Theta} |\theta - \widetilde\theta|^2 \, d\gamma(\theta, \widetilde\theta)$.*

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 $\mathcal{P}_2(K_r)$. Moreover, there exists a constant $C_{r,\tau} > 0$ such that for all $\mu, \nu \in \mathcal{P}_2(K_r)$, $\gamma \in \Gamma(\mu, \nu)$ with support included in $K_r \times K_r$:*

$$\left| \mathcal{E}_{\tau,r}(\nu) - \mathcal{E}_{\tau,r}(\mu) + \int_{\Theta^2} v_\mu(\theta) \cdot (\tilde{\theta} - \theta) d\gamma(\theta, \tilde{\theta}) \right| \leq C_{r,\tau} c_2(\gamma) \tag{28}$$

with

$$c_2(\gamma) := \int_{\Theta^2} (\theta - \tilde{\theta})^2 \, d\gamma(\theta, \tilde{\theta}),$$

and

$$v_\mu(\theta) := \nabla_\theta \phi_\mu(\theta) \tag{29}$$

where for all $\theta \in K_r$,

$$\begin{aligned} \phi_\mu(\theta) &:= \langle \nabla_x P_\tau(\mu), \nabla_x \Phi_\tau(\theta; \cdot) \rangle_{L^2(\Omega)} - \langle f, \Phi_\tau(\theta; \cdot) \rangle_{L^2(\Omega)} + \int_\Omega P_\tau(\mu)(x) dx \times \int_\Omega \Phi_\tau(\theta; x) dx \\ &= d\,\mathcal{E}\,|_{P_\tau(\mu)}(\Phi_\tau(\theta; \cdot)). \end{aligned} \tag{30}$$

The properness and coercivity are easy to prove and left to the reader. Before proving the differentiability property of $\mathcal{E}_{\tau,r}$, we will need the following auxiliary lemma.

Lemma 5. *There exists a constant $C > 0$ such that for all $\tau > 0$ and all $\theta \in \Theta$, we have*

$$\begin{aligned} \|\Phi_\tau(\theta; \cdot)\|_{L^\infty(\Omega)} &\leq C|\theta|, \\ \|\nabla_x \Phi_\tau(\theta; \cdot)\|_{L^\infty(\Omega)} &\leq C|\theta|, \\ \|\nabla_\theta \Phi_\tau(\theta; \cdot)\|_{L^\infty(\Omega)} &\leq C|\theta|, \\ \|H_\theta \Phi_\tau(\theta; \cdot)\|_{L^\infty(\Omega)} &\leq C|\theta|\tau, \\ \|\nabla_x \nabla_\theta \Phi_\tau(\theta; \cdot)\|_{L^\infty(\Omega)} &\leq C|\theta|\tau, \\ \|\nabla_x H_\theta \Phi_\tau(\theta; \cdot)\|_{L^\infty(\Omega)} &\leq C|\theta|\tau^2, \end{aligned}$$

where for all $\theta \in \Theta$ and $x \in \Omega$, $H_\theta \Phi_\tau(\theta; x)$ denotes the Hessian of Φ_τ with respect to the variable θ at the point $(\theta, x) \in \Theta \times \Omega$.

Proof. Let $\theta = (c, a, w, b) \in \Theta$. It then holds that, for all $x \in \Omega$,

$$\begin{cases} \dfrac{\partial \Phi_\tau(\theta; x)}{\partial c} = 1 \\ \dfrac{\partial \Phi_\tau(\theta; x)}{\partial a} = \sigma_{H,\tau}(w \cdot x + b) \\ \dfrac{\partial \Phi_\tau(\theta; x)}{\partial w} = ax\sigma'_{H,\tau}(w \cdot x + b) \\ \dfrac{\partial \Phi_\tau(\theta; x)}{\partial b} = a\sigma'_{H,\tau}(w \cdot x + b). \end{cases} \tag{31}$$

This expression yields the first desired inequality. In addition, the nonzero terms of the Hessian matrix read as:

$$\begin{cases} \dfrac{\partial^2 \Phi_\tau(\theta;x)}{\partial a \partial w} = \sigma'_{H,\tau}(w\cdot x+b)x \\ \dfrac{\partial^2 \Phi_\tau(\theta;x)}{\partial a \partial b} = \sigma'_{H,\tau}(w\cdot x+b) \\ \dfrac{\partial^2 \Phi_\tau(\theta;x)}{\partial^2 w} = a\sigma''_{H,\tau}(w\cdot x+b)xx^T \\ \dfrac{\partial^2 \Phi_\tau(\theta;x)}{\partial w \partial b} = a\sigma''_{H,\tau}(w\cdot x+b)x \\ \dfrac{\partial^2 \Phi_\tau(\theta;x)}{\partial^2 b} = a\sigma''_{H,\tau}(w\cdot x+b). \end{cases} \tag{32}$$

From these expressions, together with (10), we easily get that, for all $\theta \in K_r$,

$$\|H_\theta \Phi_\tau(\theta;\cdot)\|_{L^\infty(\Omega)} \leq Cr\tau,$$

for some constant $C>0$ independent of θ, r and τ. Moreover, for all $x\in\Omega$,

$$\nabla_x \Phi_\tau(\theta;x) = aw\sigma'_{H,\tau}(w\cdot x+b), \tag{33}$$

which implies that

$$\begin{cases} \dfrac{\partial \nabla_x \Phi_\tau(\theta;x)}{\partial c} = 0 \\ \dfrac{\partial \nabla_x \Phi_\tau(\theta;x)}{\partial a} = w\sigma'_{H,\tau}(w\cdot x+b) \\ \dfrac{\partial \nabla_x \Phi_\tau(\theta;x)}{\partial w} = a\sigma'_{H,\tau}(w\cdot x+b)I_d + axw^T\sigma''_{H,\tau}(w\cdot x+b) \\ \dfrac{\partial \nabla_x \Phi_\tau(\theta;x)}{\partial b} = aw\sigma''_{H,\tau}(w\cdot x+b). \end{cases} \tag{34}$$

This implies then that

$$\|\nabla_\theta \nabla_x \Phi_\tau(\theta;\cdot)\|_{L^\infty(\Omega)} \leq Cr\tau.$$

Moreover, it then holds, using again (10), that for all $\theta\in K_r$,

$$\|H_\theta \nabla_x \Phi_\tau(\theta;\cdot)\|_{L^\infty(\Omega)} \leq Cr\tau^2,$$

for some constant $C>0$ independent of θ, r and τ.

□

The following corollary is also a prerequisite for the proof of Proposition 3.

Corollary 1. *There exists a constant $C_\tau>0$ and a constant $C_{r,\tau}>0$ such that for all $\mu,\nu\in\mathcal{P}_2(K_r)$:*

$$\|P_\tau(\mu)\|^2_{H^1(\Omega)} \leq C_\tau \int_\Theta |\theta|^2 d\mu(\theta), \tag{35}$$

and

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

Proof. From Lemma 5 we immediately obtain that, for all $\tau, r>0$, there exists a constant $C_{\tau,r}>0$ such that for all $\theta_1,\theta_2\in K_r$,

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

The corollary immediately follows from that fact.

□

Now we are able to prove Proposition 3.

Proof. First, we focus on the proof of (28)-(30). 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 $N(u) := \frac{1}{2}\bar{a}(u, u)$ for all $u \in H^1(\Omega)$ and $M_\tau(\theta, \tilde{\theta}; x) := \int_0^1 (\tilde{\theta} - \theta)^T H_\theta \Phi_\tau(\theta + t(\tilde{\theta} - \theta); x)(\tilde{\theta} - \theta)(1 - t)dt$. Using Lemma 5, there exists a constant $C > 0$ independent on r and τ such that:

- $\forall x \in \Omega,\ \forall \theta, \tilde{\theta} \in K_r,\ |M_\tau(\theta, \tilde{\theta}; x)| \leq Cr\tau|\theta - \tilde{\theta}|^2$,
- $\forall x \in \Omega,\ \forall \theta, \tilde{\theta} \in K_r,\ |\nabla_x M_\tau(\theta, \tilde{\theta}; x)| \leq Cr\tau^2|\theta - \tilde{\theta}|^2$.

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

$$0 \leq N(u) \leq C\|u\|^2_{H^1(\Omega)}. \tag{36}$$

Thus, for $\mu, \nu \in \mathcal{P}_2(K_r)$ and $\gamma \in \Gamma(\mu, \nu)$ supported in K_r^2, it holds that:

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

Using standard derivation integral theorems, a bound on M_τ is available :

$$\begin{aligned} \Big\| \int_{K_r^2} M_\tau(\theta, \tilde{\theta}; \cdot) d\gamma(\theta, \tilde{\theta}) \Big\|^2_{H^1(\Omega)} &= \Big\| \int_{K_r^2} M_\tau(\theta, \tilde{\theta}; \cdot) d\gamma(\theta, \tilde{\theta}) \Big\|^2_{L^2(\Omega)} + \Big\| \int_{K_r^2} \nabla_x M_\tau(\theta, \tilde{\theta}; \cdot) d\gamma(\theta, \tilde{\theta}) \Big\|^2_{L^2(\Omega)} \\ &\leq \int_{K_r^2} \|M_\tau(\theta, \tilde{\theta}; \cdot)\|^2_{L^2(\Omega)} d\gamma(\theta, \tilde{\theta}) + \int_{K_r^2} \|\nabla_x M_\tau(\theta, \tilde{\theta}; \cdot)\|^2_{L^2(\Omega)} d\gamma(\theta, \tilde{\theta}) \\ &\leq C(r^2\tau^2 + r^2\tau^4) \int_{\Theta^2} |\tilde{\theta} - \theta|^4 d\gamma(\theta, \tilde{\theta}) \\ &\leq C(r^4\tau^2 + r^4\tau^4) \int_{\Theta^2} |\tilde{\theta} - \theta|^2 d\gamma(\theta, \tilde{\theta}) \\ &= C(r^4\tau^2 + r^4\tau^4) c_2(\gamma), \end{aligned}$$

where we used Jensen inequality to get the first inequality and Lemma 4 to get the last inequality. Using Corollary 1 and the uniform continuity of $d\,\mathcal{E}$, it holds :

$$\left| d\,\mathcal{E}\,|_{P_\tau \mu} \left(\int_{K_r^2} M_\tau(\theta, \tilde{\theta}; \cdot) d\gamma(\theta, \tilde{\theta}) \right) \right| \leq C_{\tau,r} c_2(\gamma).$$

Moreover, using similar calculations, it holds that

$$\begin{aligned}
\left\| \int_{K_r^2} \nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta)\, d\gamma(\theta,\tilde\theta) \right\|^2_{H^1(\Omega)} &= \left\| \int_{K_r^2} \nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta)\, d\gamma(\theta,\tilde\theta) \right\|^2_{L^2(\Omega)} \\
&\quad + \left\| \int_{K_r^2} \nabla_x\nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta)\, d\gamma(\theta,\tilde\theta) \right\|^2_{L^2(\Omega)}, \\
&\le \int_{K_r^2} \left\| \nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta) \right\|^2_{L^2(\Omega)} d\gamma(\theta,\tilde\theta) \\
&\quad + \int_{K_r^2} \left\| \nabla_x\nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta) \right\|^2_{L^2(\Omega)} d\gamma(\theta,\tilde\theta), \\
&\le C(r^2+r^2\tau^2)\int_{\Theta^2} |\tilde\theta-\theta|^2 d\gamma(\theta,\tilde\theta) \\
&\le C(r^2+r^2\tau^2)c_2(\gamma).
\end{aligned}$$

Hence, together with the previous bounds and (36), we easily obtain that there exists a constant $C_{r,\tau}>0$ such that for all $\mu,\nu\in\mathcal{P}_2(K_r)$, it holds that

$$\left| \mathcal{E}_{\tau,r}(\nu) - \mathcal{E}_{\tau,r}(\mu) + d\,\mathcal{E}\,|_{P_\tau(\mu)} \left(\int_{K_r^2} \nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta-\theta) d\gamma(\theta,\tilde\theta) \right) \right| \le C_{r,\tau} c_2(\gamma). \tag{37}$$

Now we focus on the first order term and by Fubini and standard integral derivation theorem, we obtain that:

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

where

$$v_\mu(\theta) := \nabla_\theta \phi_\mu(\theta) \quad \gamma-\text{almost everywhere}, \tag{38}$$

with

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

Note that (38) is equivalent to

$$v_\mu(\theta) := \nabla_\theta \phi_\mu(\theta) \quad \mu-\text{almost everywhere},$$

as v_μ only depends on θ.

□

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

Proposition 4. *For all $\tau, r > 0$, there exists $\lambda_{\tau,r} > 0$ such that for all $\mu, \nu \in \mathcal{P}_2(K_r)$ with associated geodesic $\kappa(t) := e_t\#\Pi$ given by Proposition 2, the functional $[0,1] \ni t \mapsto \frac{d}{dt}\left(\mathcal{E}_{\tau,r}(\kappa(t))\right)$ is $-\lambda_{\tau,r}$-Lipschitz.*

Proof. First of all, one has to check that for all $t \in [0,1]$, $\kappa(t) \in \mathcal{P}_2(K_r)$. 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).

Let $t, s \in [0,1]$ and define $\alpha(t,s) := (e_t, e_s)\#\Pi \in \Gamma(\kappa(t), \kappa(s))$. By (37), it holds that

$$\left|\mathcal{E}_{\tau,r}(\kappa(s)) - \mathcal{E}_{\tau,r}(\kappa(t)) + \int_{\Theta^2} d\,\mathcal{E}\,|_{P_\tau(\kappa(t))}\Big(\nabla_\theta \Phi_\tau(\theta;\cdot)\cdot(\tilde\theta - \theta)\Big) d\alpha(t,s)(\theta,\tilde\theta)\right| \leq C_{r,\tau} c_2(\alpha(t,s)),$$

which reads equivalently as

$$\begin{aligned}
&\left|\frac{\mathcal{E}_{\tau,r}(\kappa(s)) - \mathcal{E}_{\tau,r}(\kappa(t))}{s-t} - \int_{\mathfrak{P}} d\,\mathcal{E}\,|_{P_\tau(\kappa(t))}\Big(\nabla_\theta \Phi_\tau(\pi(t);\cdot)\cdot\Big(\frac{\pi(s)-\pi(t)}{s-t}\Big)\Big) d\Pi(\pi)\right| \\
\leq & C_{r,\tau}\frac{1}{|s-t|}\int_{\Theta^2} |\theta - \tilde\theta|^2\, d\alpha(t,s)(\theta,\tilde\theta) \\
= & C_{r,\tau}\frac{1}{|s-t|}\int_{\mathfrak{P}} |\pi(t) - \pi(s)|^2\, d\Pi(\pi) \\
= & C_{r,\tau}|s-t|\int_{\mathfrak{P}} |\pi(1) - \pi(0)|^2\, d\Pi(\pi) \\
\leq & C_{r,\tau}|s-t|,
\end{aligned}$$

where the value of the constant $C_{r,\tau}$ only depends on r and τ. Letting s go to t and using the dominated convergence theorem, one concludes that $[0,1] \ni t \mapsto \mathcal{E}_{\tau,r}(\kappa(t))$ is differentiable with derivative equal to :

$$h(t) := \frac{d}{dt}\left(\mathcal{E}_{\tau,r}(\kappa(t))\right) = \int_{\mathfrak{P}} d\,\mathcal{E}\,|_{P_\tau(\kappa(t))}\Big(\nabla_\theta \Phi_\tau(\pi(t);\cdot)\cdot\Big(\frac{d}{dt}\pi(t)\Big)\Big) d\Pi(\pi).$$

To conclude, one has the decomposition :

$$\begin{aligned}
|h(t) - h(s)| \leq & \left|\int_{\mathfrak{P}} d\,\mathcal{E}\,|_{P_\tau(\kappa(t))}\Big((\nabla_\theta \Phi_\tau(\pi(t);\cdot) - \nabla_\theta \Phi_\tau(\pi(s);\cdot))\cdot\Big(\frac{d}{dt}\pi(t)\Big)\Big) d\Pi(\pi)\right| \\
& + \left|\int_{\mathfrak{P}} (d\,\mathcal{E}\,|_{P_\tau(\kappa(t))} - d\,\mathcal{E}\,|_{P_\tau(\kappa(s))})\Big(\nabla_\theta \Phi_\tau(\pi(s);\cdot)\cdot\Big(\frac{d}{dt}\pi(t)\Big)\Big) d\Pi(\pi)\right| \\
& + \left|\int_{\mathfrak{P}} d\,\mathcal{E}\,|_{P_\tau(\kappa(s))}\Big(\nabla_\theta \Phi_\tau(\pi(s);\cdot)\cdot\Big(\frac{d}{dt}\pi(t) - \frac{d}{dt}\pi(s)\Big)\Big) d\Pi(\pi)\right|.
\end{aligned} \tag{39}$$

Recalling (39), denoting $\alpha := (e_0, e_1)\#\Pi$ and using the previous estimates, we obtain that, for all

$t, s \in [0,1]$,

$$
\begin{aligned}
|h(t)-h(s)| &\leq C_{r,\tau}\Big(\|P_\tau(\kappa(t))\|_{H^1(\Omega)} \int_{\mathfrak{P}} |\pi(t)-\pi(s)|\Big|\frac{d}{dt}\pi(t)\Big| d\Pi(\pi) \\
&+ \|P_\tau(\kappa(t)) - P_\tau(\kappa(s))\|_{H^1(\Omega)} \int_{\mathfrak{P}} |\pi(s)|\Big|\frac{d}{dt}\pi(t)\Big| d\Pi(\pi) \\
&+ \|P_\tau(\kappa(s))\|_{H^1(\Omega)} \int_{\mathfrak{P}} |\pi(s)|\Big|\frac{d}{dt}\pi(t) - \frac{d}{dt}\pi(s)\Big| d\Pi(\pi)\Big) \\
&\leq C_{r,\tau}\Big(|t-s|\|P_\tau(\kappa(t))\|_{H^1(\Omega)} \int_{\mathfrak{P}} |\pi(1)-\pi(0)|^2 d\Pi(\pi) \\
&+ \|P_\tau(\kappa(t)) - P_\tau(\kappa(s))\|_{H^1(\Omega)} \int_{\mathfrak{P}} \sup_{u\in[0,1]} |\pi(u)||\pi(1)-\pi(0)| d\Pi(\pi) \\
&+ |t-s|\|P_\tau(\kappa(s))\|_{H^1(\Omega)} \int_{\mathfrak{P}} \sup_{u\in[0,1]} |\pi(u)| \sup_{u\in[0,1]} \left|\frac{d^2\pi(u)}{dt^2}\right| d\Pi(\pi)\Big) \\
&\leq C_{r,\tau}\left(|t-s|\left(\sqrt{\int_{\Theta^2} |\theta|^2 d\kappa(t)(\theta)} + \sqrt{\int_{\Theta^2} |\theta|^2 d\kappa(s)(\theta)}\right)(1+c_2(\alpha)) + W_2(\kappa(t),\kappa(s))c_2(\alpha)\right)
\end{aligned}
$$

where we have used Lemma 4 to get the second inequality and the fact that $\sup_{u\in[0,1]} \left|\frac{d^2\pi(u)}{dt^2}\right|$ is uniformly bounded (since the curvature of Θ is bounded) to get the last one. We also have the following estimates:

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

$$\int_\Theta |\theta|^2 d\kappa(t)(\theta) \leq C(1+r^2).$$

- Moreover,

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

This allows us to conclude that :

$$|h(t)-h(s)| \leq C_{r,\tau}(1+c_2(\alpha))|t-s|.$$

As the measure α is supported in K_r^2, we get :

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

for some $\lambda_{\tau,r} > 0$, which yields the desired result. □

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 2. *There exists a constant $C_\tau > 0$ such that for all $r > 0$, all $\mu \in \mathcal{P}_2(K_r)$ and $\theta \in \Theta$:*

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

Proof. This can be proved combining (35), (31) and (34). The rest is just basic computations and left to the reader. □

An important consequence of Proposition 4 is that $\mathcal{E}_{\tau,r}$ is $(-\lambda_{\tau,r})$-convex along geodesics. Now we are able to prove Theorem 4.

Proof of Theorem 4. 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 [18, Theorem 5.9, 5.11] to get the existence and the uniqueness of a gradient curve $\mu^r : \mathbb{R}_+ \to \mathcal{P}_2(K_r)$ in the sense of [18, Definition 5.8]. Being a gradient curve, it is also a curve of maximal slope in the sense of [16, Definition 1.3.2]. Note that in [18], the space on which the probability measures are defined (here this is Θ) is supposed to be compact. This is not a problem here since the domain of the functional $\mathcal{E}_{\tau,r}$ is reduced to probability measures whose support is included in K_r which is compact and geodesically convex.

The existence of the vector field v_t^r for almost all $t \geq 0$ is given by the absolute continuity of the curve $[0,1] \ni t \mapsto \mu^r(t)$ (because it is a gradient curve) and by [19, Proposition 2.5].

□

The work is not finished here since we do not have any knowledge about the velocity field v_t^r and the well-posedness result is proved only for $\mathcal{E}_{\tau,r}$ with $r < \infty$. In the following sections, we prove that this velocity field can be related to $v_{\mu^r(t)}$ and use a bootstrap argument to prove an existence result for the gradient curve of $\mathcal{E}_{\tau,+\infty}$.

3.1.2 Identification of the vector field v_t^r

In the following, we denote by $T\Theta$ the tangent bundle of Θ, i.e.

$$T\Theta := \bigcup_{\theta \in \Theta} \{\theta\} \times T_\theta \Theta,$$

where $T_\theta\Theta$ is the tangent space to Θ at θ. It is easy to check that for all $\theta = (c, a, w, b) \in \Theta$, it holds that $T_\theta\Theta = \mathbb{R} \times \mathbb{R} \times \mathrm{Span}\{w\}^\perp \times \mathbb{R}$, where $\mathrm{Span}\{w\}^\perp$ is the subspace of $\mathbb{R}^d$ containing all d-dimensional vectors orthogonal to w.

We also introduce the operators G and S_h for $0 < h \leq 1$ as follows :

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

and

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

The next lemma concerns the local behaviour of couplings along a curve of maximal slope $\mu^r : \mathbb{R}_+ \to \mathcal{P}_2(K_r)$. In the following, for any $\mu, \nu \in \mathcal{P}_2(\Theta)$, we denote by $\Gamma_o(\mu, \nu)$ the set of optimal transport plans between μ and ν in the sense of the quadratic cost. In other words, for all $\gamma \in \Gamma_o(\mu, \nu)$, it holds that $W_2^2(\mu, \nu) = \displaystyle\int_{\Theta \times \Theta} |\theta - \widetilde{\theta}|^2 \, d\gamma(\theta, \widetilde{\theta})$.

Lemma 6. *Let $\mu^r : \mathbb{R}_+ \to \mathcal{P}_2(K_r)$ be a solution to (23) and for all $0 < h \leq 1$, let $\Pi_h \in \mathcal{P}_2(\mathfrak{P})$ such that $\gamma_h := (e_0, e_1)\#\Pi_h \in \Gamma_o(\mu^r(t), \mu^r(t+h))$ (i.e. satisfying the condition of Proposition 2 with $\mu = \mu^r(t)$ and $\nu = \mu^r(t+h)$). Then, for almost all $t \geq 0$, it holds that*

$$\lim_{h \to 0} (S_h \circ G)\#\Pi_h = (i \times v_t^r)\#\mu^r(t) \ \text{ in } \ \mathcal{P}_2(T\Theta),$$

where $(v_t^r)_{t \geq 0}$ is given by Theorem 4, and $i : \Theta \to \Theta$ is the identity map.

Moreover,

$$\lim_{h \to 0} \frac{W_2^2(\mu^r(t+h), \exp(hv_t^r)\#\mu^r(t))}{h^2} = 0,$$

where $\exp(hv_t^r) : \Theta \ni \theta \mapsto \exp_\theta(hv_t^r(\theta))$.

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

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

for η smooth compactly supported in $\mathbb{R}_+$. Taking η as an approximation of the characteristic function of $[t, t+h]$, owing to the fact that μ^r is locally Lipschitz and passing to the limit, one gets :

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

Passing to the limit as h goes to 0, one gets the differentiability almost everywhere of $\mathbb{R}_+ \ni t \mapsto \int_\Theta \phi \, d\mu^r(t)$ and :

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

For all $0 < h \leq 1$, let us introduce $\nu_h := (S_h \circ G) \# \Pi_h$ and let ν_0 be an accumulation point of $(\nu_h)_{0<h\leq 1}$ with respect to the narrow convergence on $\mathcal{P}_2(T\Theta)$.

Then, it holds that

$$\begin{aligned}
\frac{\int_\Theta \phi \, d\mu^r(t+h) - \int_\Theta \phi \, d\mu^r(t)}{h} &= \frac{1}{h} \int_{\Theta^2} (\phi(\tilde{\theta}) - \phi(\theta)) \, d\gamma_h(\theta, \tilde{\theta}) \\
&= \frac{1}{h} \int_{\mathfrak{P}} (\phi(\pi(1)) - \phi(\pi(0))) \, d\Pi_h(\pi) \\
&= \frac{1}{h} \int_{T\Theta} (\phi(\exp_\theta(v)) - \phi(\theta)) \, dG \# \Pi_h(\theta, v) \\
&= \frac{1}{h} \int_{T\Theta} (\phi(\exp_\theta(hv)) - \phi(\theta)) \, d(S_h \circ G) \# \Pi_h(\theta, v) \\
&= \int_{T\Theta} \nabla_\theta \phi(\theta) \cdot v \, d\nu_h(\theta, v) \\
&+ \int_{T\Theta} R_h(\theta, v) \, d\nu_h(\theta, v) \\
&\underset{h \to 0}{\longrightarrow} \int_{T\Theta} \nabla_\theta \phi(\theta) \cdot v d\nu_0(\theta, v),
\end{aligned}$$

where $R_h(\theta, v) := \dfrac{\phi(\exp_\theta(hv)) - \phi(\theta)}{h} - \nabla_\theta \phi(\theta) \cdot v$ is bounded by $C(\phi)|v|^2 h$ ($\phi \in C_c^\infty(\Theta)$ and the euclidean curvature in Θ is uniformly bounded; see [20, Chapter 8] for the definition of euclidean curvature). Actually, to get the last limit, we need the following arguments detailed below :

- For the first term, $\nabla \phi(\theta) \cdot v$ is quadratic in (θ, v) and consequently the passage to the limit is allowed.
- For the second one,

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

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

As a consequence,

$$\int_{T\Theta} \nabla_\theta \phi(\theta) \cdot v d\nu_0(\theta, v) = \int_\Theta \nabla_\theta \phi(\theta) \cdot v_t^r(\theta) \, d\mu^r(t)(\theta)$$

which is no more than (by disintegration) :

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

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

$$\operatorname{div}((\tilde{v_t} - v_t^r)\mu^r(t)) = 0.$$

In addition, as $T\Theta \ni (\theta, v) \mapsto |v|^2$ is positive and lower semicontinuous and as for almost all $t \geq 0$ we have that $\lim_{h\to 0} \dfrac{W_2(\mu^r(t), \mu^r(t+h))}{h} = |(\mu^r)'|(t)$ (as μ^r is locally Lipschitz):

$$\begin{aligned}
\int_\Theta \int_{T_\theta\Theta} |v|^2 \, d\nu_{0,\theta}(v) \, d\mu^r(t)(\theta) &\leq \liminf_{h\to 0} \int_{T\Theta} |v|^2 \, d\nu_h(\theta, v) \\
&= \liminf_{h\to 0} \frac{1}{h^2} \int_{T\Theta} |v|^2 \, dG\#\Pi_h(\theta, v) \\
&= \liminf_{h\to 0} \frac{1}{h^2} \int_{\mathfrak{P}} |\dot{\pi}(0)|^2 \, d\Pi_h(\pi) \\
&= \liminf_{h\to 0} \frac{1}{h^2} \int_{\Theta^2} d(\theta, \tilde{\theta})^2 \, d\gamma_h(\theta, \tilde{\theta}) \\
&= \liminf_{h\to 0} \frac{W_2^2(\mu^r(t), \mu^r(t+h))}{h^2} \\
&= |(\mu^r)'|^2(t).
\end{aligned} \tag{40}$$

As a consequence and by Jensen inequality,

$$\|\tilde{v_t}\|^2_{L^2(\Theta; d\mu^r(t))} \leq \int_\Theta \int_{T_\theta\Theta} |v|^2 \, d\nu_{0,\theta}(v) \, d\mu^r(t)(\theta) \leq |(\mu^r)'|^2(t) = \|v_t^r\|^2_{L^2(\Theta; d\mu^r(t))}. \tag{41}$$

By [19, Lemma 2.4], one gets $\tilde{v}_t = v_t^r$. Reconsidering (41), one gets the equality case in Jensen inequality *ie* :

$$\int_\Theta |\tilde{v_t}(\theta)|^2 \, d\mu^r(t)(\theta) = \int_\Theta \int_{T_\theta\Theta} |v|^2 d\nu_{h,\theta}(v) \, d\mu^r(t)(\theta),$$

and as a consequence $\nu_{0,\theta} = \delta_{v_t^r(\theta)}$, $\mu^r(t)$-almost everywhere in Θ. In addition,

$$\lim_{h\to 0} (S_h \circ G)\#\Pi_h = (i \times v_t^r)\#\mu^r(t),$$

in the sense of the narrow convergence. The convergence of the v moment is given by (40)-(41) where inequalities can be replaced by equalities (as $\tilde{v}_t = v_t^r$) and the lim inf can be replaced by a lim as $\lim_{h\to 0} \dfrac{W_2(\mu^r(t), \mu^r(t+h))}{h} = |(\mu^r)'|(t)$ exists :

$$\int_\Theta \int_{T_\theta\Theta} |v|^2 d\nu_{0,\theta}(v) \, d\mu^r(t)(\theta) = \lim_{h\to 0} \int_{T\Theta} |v|^2 d\nu_h(\theta, v). \tag{42}$$

For the θ moment, it is more obvious as for all $0 < h \leq 1$:

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

and

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

Consequently,

$$\int_{T\Theta} |\theta|^2 d\nu_0(\theta, v) = \lim_{h\to 0} \int_{T\Theta} |\theta|^2 d\nu_h(\theta, v). \tag{43}$$

With (42)-(43), the convergence of moments is asserted. The narrow convergence combined with the convergence of moments gives the convergence in $\mathcal{P}_2(\Theta)$ and the proof of the first part of the

lemma is finished.

For the second part, it holds that $(\exp(hv_t^r) \times i)\#\gamma_h$ belongs to $\Gamma(\exp(hv_t^r)\#\mu^r(t), \mu^r(t+h))$. Hence,

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

where we have used the boundedness of the euclidean curvature of the manifold Θ in the last inequality and the fact that $\nu_h \to (i \times v_t^r)\#\mu^r(t)$, which was proved earlier. Hence the desired result. □

We now introduce the projection operator on the manifold Θ :

Definition 5. *For all θ in Θ, the orthogonal projection on the tangent space of Θ is given by the operator $\mathbf{P}_\theta : \mathbb{R}^{d+3} \to T_\theta\Theta$. The operator $\mathbf{P} : L^1_{\mathrm{loc}}(\Theta;\mathbb{R}^{d+3}) \to L^1_{\mathrm{loc}}(\Theta;\mathbb{R}^{d+3})$ denotes the corresponding projection on vector fields, i.e. for all $X \in L^1_{\mathrm{loc}}(\Theta;\mathbb{R}^{d+3})$, $(\mathbf{P}X)(\theta) := \mathbf{P}_\theta X(\theta)$ for almost all $\theta \in \Theta$.*

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 $\mathrm{Supp}(\mu^r(t)) \subset K_{r-\delta}$, then the velocity field v_t^r in (23) is equal to $-\mathbf{P}v_{\mu^r(t)}$ $\mu^r(t)$-almost everywhere.*

Proof. On the one hand, for $\gamma_h := (e_0, e_1)\#\Pi_h \in \Gamma_o(\mu^r(t), \mu^r(t+h))$, by Proposition 3 and the fact that for all $t \geq 0$, $\mu^r(t) \in \mathcal{P}_2(K_r)$:

$$\left|\mathcal{E}_{\tau,r}(\mu^r(t+h)) - \mathcal{E}_{\tau,r}(\mu^r(t)) - \int_{\Theta^2} v_{\mu^r(t)}(\theta)\cdot(\tilde{\theta}-\theta)\, d\gamma_h(\theta,\tilde{\theta})\right| \leq C_{r,\tau} W_2(\mu^r(t),\mu^r(t+h))^2,$$

which is equivalent to

$$\left|\frac{\mathcal{E}_{\tau,r}(\mu^r_{t+h}) - \mathcal{E}_{\tau,r}(\mu^r(t))}{h} - \int_{T\Theta} v_{\mu^r(t)}(\theta)\cdot\frac{\exp_\theta(hv)-\theta}{h}\, d(S_h\circ G)\#\Pi_h(\theta,v)\right| \leq C_{r,\tau}\frac{1}{h} W_2(\mu^r(t),\mu^r(t+h))^2.$$

Then, one can use the decomposition :

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

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 $\mathbb{R}_+ \ni t \to \mathcal{E}_{\tau,r}(\mu^r(t))$ almost everywhere and for almost all $t \geq 0$:

$$\frac{d}{dt}\left[\mathcal{E}_{\tau,r}(\mu^r(t))\right] = \int_\Theta v_{\mu^r(t)}(\theta)\cdot v_t^r(\theta)\, d\mu^r(t)(\theta).$$

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

- First, $v \cdot v_{\mu^r(t)}(\theta)$ is at most quadratic in (θ, v) which is given by Corollary 2.

- Second, it holds that $|v_{\mu^r(t)}(\theta) \cdot R_h(\theta, v)| \leq Cr|\theta|h|v|^2$ by Corollary 2 and consequently :

$$\begin{aligned}\left|\int_{T\Theta} v_{\mu^r(t)}(\theta) \cdot R_h(\theta, v)\, d(S_h \circ G)\#\Pi_h(\theta, v)\right| &\leq C_r h \int_{T\Theta} |\theta||v|^2\, d(S_h \circ G)\#\Pi_h(\theta, v) \\ &\leq C_r h \int_{T\Theta} |v|^2\, d(S_h \circ G)\#\Pi_h(\theta, v) \\ &\leq C_r h \frac{W_2(\mu^r(t), \mu^r(t+h))^2}{h^2}\end{aligned}$$

  where we used the fact that Π_h is supported in K_r in its first variable to get the second inequality. The last term converges to zero since $(\mu_r(t))_t$ is local Lipschitz.

Next as $\mathbf{P}v_t^r = v_t^r$, it holds that:

$$\frac{d}{dt}\left[\mathcal{E}_{\tau,r}(\mu^r(t))\right] = \int_{\Theta^2} \mathbf{P}v_{\mu^r(t)}(\theta) \cdot v_t^r(\theta)\, d\mu^r(t)(\theta). \tag{44}$$

On the other hand, consider the curve $\tilde{\mu_h} : \mathbb{R}_+ \to \mathcal{P}_2(\Theta)$ satisfying :

$$\forall t \geq 0, \quad \tilde{\mu}_h(t) := \exp(-h\mathbf{P}v_{\mu^r(t)})\#\mu^r(t).$$

As $\mathrm{Supp}(\mu^r(t)) \subset K_{r-\delta}$, there exists a small time interval around zero such that $\tilde{\mu}_h(t)$ is in $\mathcal{P}_2(K_r)$ for $h > 0$ small enough. So, with $\gamma_h := (i \times \exp(-h\mathbf{P}v_{\mu^r(t)}))\#\mu^r(t) \in \Gamma(\mu^r(t), \tilde{\mu}_h(t))$,

$$\left|\mathcal{E}_{\tau,r}(\tilde{\mu}_h(t)) - \mathcal{E}_{\tau,r}(\mu^r(t)) - \int_{\Theta^2} \mathbf{P}v_{\mu^r(t)}(\theta) \cdot (\tilde{\theta} - \theta)\, d\gamma_h(\theta, \tilde{\theta})\right| \leq C_{r,\tau} W_2^2(\mu^r(t), \tilde{\mu}_h(t))$$

and it holds that

$$\int_{\Theta^2} \mathbf{P}v_{\mu^r(t)}(\theta) \cdot (\tilde{\theta} - \theta)\, d\gamma_h(\theta, \tilde{\theta}) = h \int_{\Theta^2} \mathbf{P}v_{\mu^r(t)}(\theta) \cdot \frac{\exp_\theta(-h\mathbf{P}v_{\mu^r(t)}(\theta)) - \theta}{h}\, d\mu^r(t)(\theta).$$

Hence,

$$\frac{\mathcal{E}_{\tau,r}(\tilde{\mu}_h(t)) - \mathcal{E}_{\tau,r}(\mu^r(t))}{W_2(\tilde{\mu}_h(t), \mu^r(t))} = \frac{h}{W_2(\tilde{\mu}_h(t), \mu^r(t))} \int_{\Theta^2} \mathbf{P}v_{\mu^r(t)}(\theta) \cdot \frac{\exp_\theta(-h\mathbf{P}v^\tau_{\mu^r(t)}(\theta)) - \theta}{h}\, d\mu^r(t)(\theta) + o_h(1)$$

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 to the fact that $\limsup_{h\to 0} \frac{W_2(\tilde{\mu}_h(t), \mu^r(t))}{h} \leq \|\mathbf{P}v_{\mu^r(t)}\|_{L^2(\Theta; d\mu^r(t))}$, we obtain that

$$|\nabla^- \mathcal{E}_{\tau,r}|(\mu^r(t)) \geq \|\mathbf{P}v_{\mu^r(t)}\|_{L^2(\Theta; d\mu^r(t))}. \tag{45}$$

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

$$\begin{aligned}\frac{d}{dt}\left[\mathcal{E}_{\tau,r}(\mu^r(t))\right] = \int_\Theta \mathbf{P}v_{\mu^r(t)}(\theta) \cdot v_t^r(\theta)\, d\mu^r(t)(\theta) &\leq -\frac{1}{2}\|v_t^r\|_{L^2(\Theta; d\mu^r(t))} - \frac{1}{2}|\nabla^- \mathcal{E}_{\tau,r}|^2(\mu^r(t)) \\ &\leq -\frac{1}{2}\|v_t^r\|^2_{L^2(\Theta; d\mu^r(t))} - \frac{1}{2}\|\mathbf{P}v_{\mu^r(t)}\|^2_{L^2(\Theta; d\mu^r(t))}\end{aligned}$$

where we have used (45). As a consequence,

$$\int_\Theta \left(\frac{1}{2}(\mathbf{P}v_{\mu^r(t)})^2(\theta) + \frac{1}{2}|v_t^r(\theta)|^2 - \mathbf{P}v_{\mu^r(t)}(\theta) \cdot v_t^r(\theta)\right) d\mu^r(t)(\theta) \leq 0$$

and

$$v_t^r = -\mathbf{P}v_{\mu^r(t)} \quad \mu^r(t)\text{-a.e.}$$

□

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 $\chi^r : \mathbb{R}_+ \times \Theta \to \Theta$ be the flow associated to the velocity field $-\mathbf{P}v_{\mu^r(t)}$:*

$$\begin{cases} \partial_t \chi^r(t) = -\mathbf{P}v_{\mu^r(t)} \\ \chi^r(0;\theta) = \theta. \end{cases}$$

Then χ^r is uniquely defined, continuous, and for all $t \geq 0$, $\chi^r(t)$ is Lipschitz on K_r. Moreover, as long as $\mathrm{Supp}(\mu^r(t)) \subset K_{r-\delta}$ for some $\delta > 0$:

$$\mu^r(t) = \chi^r(t)\#\mu_0.$$

Proof. This is a direct consequence of the fact that $v_t^r = -\mathbf{P}v_{\mu^r(t)} = -\mathbf{P}\nabla_\theta \phi_{\mu^r(t)}$ is C^∞. □

Next lemma relates the curve $[0,1] \ni h \mapsto \exp(hv_t^r)\#\mu^r(t)$ with $\nabla_- \mathcal{E}_{\tau,r}(\mu^r(t))$. This will be useful later to prove that the velocity field characterizes the gradient curve.

Lemma 7. *For all $\mu \in \mathcal{P}_2(\Theta)$ with $\mathrm{Supp}(\mu) \subset K_{r-\delta}$ for some $\delta > 0$, the map $\nu : [0,1] \ni h \mapsto \exp(-h\mathbf{P}v_\mu/\|\mathbf{P}v_\mu\|_{L^2(\Theta;d\mu)})\#\mu$ is differentiable at $h = 0$. Moreover, it holds that*

$$\nu'(0) = \nabla_- \mathcal{E}_{\tau,r}(\mu)/|\nabla_- \mathcal{E}_{\tau,r}|(\mu).$$

Proof. First, we claim that $|\nabla_- \mathcal{E}_{\tau,r}(\mu)| = \|\mathbf{P}v_\mu(\theta)\|_{L^2(\Theta;d\mu)}$. In order to prove it, take an arbitrary unit speed geodesic $[0,1] \ni s \mapsto (e_s)\#\Pi$ starting at μ for which there exists a time interval around zero such that $(e_s)\#\Pi$ belongs to $\mathcal{P}_2(K_r)$. As a consequence, one can write for all $s > 0$ sufficiently small :

$$\left| \mathcal{E}_{\tau,r}((e_s)\#\Pi) - \mathcal{E}_{\tau,r}(\mu) + \int_{\Theta^2} v_\mu(\theta) \cdot (\tilde{\theta} - \theta)\, d(e_0, e_s)\#\Pi(\theta, \tilde{\theta}) \right| \leq C_{r,\tau} W_2^2(\mu, (e_s)\#\Pi).$$

with

$$\int_{\Theta^2} v_\mu(\theta) \cdot (\tilde{\theta} - \theta)\, d(e_0, e_s)\#\Pi(\theta, \tilde{\theta}) = \int_{T\Theta} v_\mu(\theta) \cdot (\exp_\theta(sv) - \theta)\, dG\#\Pi(\theta, v).$$

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

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

Note that, to get the last equation, we need to prove that for all s sufficiently small the function $\eta(s) : T\Theta \ni (\theta, v) \mapsto v_\mu(\theta) \cdot \dfrac{\exp_\theta(sv) - \theta}{s}$ is uniformly integrable with respect to $G\#\Pi$. In fact, this is given by Corollary 2 and the uniform curvature bound on Θ giving $|\eta(s)|(\theta, v) \leq Csr|\theta||v|^2$. As the term $Cr|\theta||v|^2$ is integrable with respect to the measure $G\#\Pi$ (recall that it has finite second-order moments and is supported in K_r in the θ variable), we have the desired uniform integrability property.

Moreover, by Cauchy-Schwartz inequality:

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

where the last equality comes from :

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

The last equality is derived from the fact that $[0,1] \ni s \mapsto (e_s)\#\Pi$ is a unit speed geodesic. To conclude, we have proved that for all unit speed geodesic $(\alpha, 1) \in C_\mu(\mathcal{P}_2(K_r))$

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

which by [18, Lemma 4.3], asserts that :

$$|\nabla_- \mathcal{E}_{\tau,r}|(\mu) \leq \|\mathbf{P}v_\mu\|_{L^2(\Theta;d\mu)}. \tag{46}$$

Aside that, let $h > 0$:

$$\begin{aligned} W_2^2(\nu(h),\nu(0)) &\leq \int_\Theta d^2\left(\exp_\theta\left(-h\mathbf{P}v_\mu(\theta)/\|\mathbf{P}v_\mu^\tau\|_{L^2(\Theta;d\mu)}\right),\theta\right)\, d\mu(\theta) \\ &\leq h^2 \int_\Theta d^2\left(\exp_\theta\left(-\mathbf{P}v_\mu(\theta)/\|\mathbf{P}v_\mu\|_{L^2(\Theta;d\mu)}\right),\theta\right)\, d\mu(\theta) \\ &= h^2, \end{aligned}$$

and

$$\limsup_{h\to 0} \frac{W_2(\nu(h),\nu_0)}{h} \leq 1. \tag{47}$$

Moreover as $\mathrm{Supp}(\mu) \subset K_{r-\delta}$, v_μ is bounded in $L^\infty(K_r)$ by Corollary 2 and for a small time interval around zero $\nu(h) \in \mathcal{P}_2(K_r)$. Consequently, as h goes to 0,

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

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

$$\lim_{h\to 0} \frac{\mathcal{E}_{\tau,r}(\nu(h)) - \mathcal{E}_{\tau,r}(\mu)}{h} = -\|\mathbf{P}v_\mu^\tau(\theta)\|_{L^2(\Theta;d\mu)}. \tag{48}$$

Additionally, with (47) :

$$\limsup_{h\to 0} \frac{\mathcal{E}_{\tau,r}(\nu(h)) - \mathcal{E}_{\tau,r}(\mu)}{W_2(\nu(h),\nu(0))} \leq -\|\mathbf{P}v_\mu\|_{L^2(\Theta;d\mu)}. \tag{49}$$

To conclude :

- With (49) and (46), the claim is proved :

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

- Owing to this, (47) and (49) the curve $[0,1] \ni h \mapsto \nu(h)$ is differentiable at $h = 0$ by [18, 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 without support limitation

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

Proof of Theorem 5. Let :

- $r_0 > 0$ be such that $\mathrm{Supp}(\mu_0) \subset K_{r_0}$,
- $\mu^r : \mathbb{R}_+ \ni t \mapsto \mu^r(t)$ the gradient curve associated to $\mathcal{E}_{\tau,r}$ for $r > r_0$.

By Corollary 2, it holds that $|v_{\mu^r(t)}(\theta)| \leq Cr|\theta|$ for all $t \geq 0$. Hence, for all $\theta \in K_{r_0}$, $|\chi^r(t;\theta)| \leq r_0 e^{Crt}$ for all time $t \in \left[0, T_r := \frac{1}{Cr}\log\left(\frac{r+r_0}{2r_0}\right)\right]$ and $\mathrm{Supp}(\mu^r(t)) \subset K_{(r+r_0)/2} \subset K_r$. By the definition of the gradient curve :

$$\forall t \in [0,T_r],\ (\mu^r)'(t) = \nabla_-\,\mathcal{E}_{\tau,r}(\mu^r(t)) = g'(0) \tag{50}$$

with $g : [0,1] \ni h \mapsto \exp\left(-\mathbf{P}v_{\mu^r(t)}h\right)$, by Lemma 7. Note that the right hand side of last equation does not depend explicitly on r but on $\mu^r_\cdot$.

We construct the curve $\mu : [0,T_r] \to \mathcal{P}_2(\Theta)$ as follows:

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

This is well-defined since by uniqueness of the gradient curve with respect to $\mathcal{E}_{\tau,r}$, $\mu^{r_1}(t) = \mu^{r_2}(t)$ on $[0, \min(T_{r_1}, T_{r_2})]$ for $r_0 < r_1 \leq r_2$. Defining for all $n \in \mathbb{N}^*$

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

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 μ^r derived in Theorem 4 and Proposition 6. □

Remark 6. *We make here two important remarks:*

- *We did not prove the existence of a gradient curve with respect to $\mathcal{E}_{\tau,\infty}$ because this functional is not proved to be convex along geodesics and it is impossible to define gradients without such an assumption.*
- *The uniqueness of a solution to (24) is out of the scope of this article. To prove it, one should link (24) and the support condition to prove that locally in time, a solution to (24) coincides with the unique gradient curve of $\mathcal{E}_{\tau,r}$ for some $r > 0$ large enough.*

3.2 Link with backpropagation in neural network

Here, we give a proof of Theorem 6.

Proof of Theorem 6. Returning back to the proof of Theorem 5 and for all time $T > 0$, one can find $r > 0$ large enough such that μ, μ_m coincide with gradient curves on $[0,T]$ with respect to $\mathcal{E}_{\tau,r}$ starting from μ_0 and $\mu_{0,m}$ respectively. As gradient curves with respect to $\mathcal{E}_{\tau,r}$ verifies the following semigroup property [18, Theorem 5.11]

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

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

3.3 Convergence of the measure towards the optimum

In the following, a LaSalle's principle argument is invoked in order to prove Theorem 7. 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 functional $\mathcal{E}_\tau$ to the set of signed finite measures on Θ, denoted by $\mathcal{M}(\Theta)$.

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 $\mathcal{E}_\tau(\mu) = 0$. Taking $\mu_p := \delta_{(0,0,w,b)}$ with w, b taken arbitrary ěis sufficient to prove the desired result. Now, if $\mu(\Theta) \neq 0$, consider $\mu_p := T\#\left(\frac{\mu}{\mu(\Theta)}\right)$ 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 (18)-(19) 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 μ-measurable sets such that $P \cup N = \Theta$ and μ is non-negative (respectively non-positive) on P (respectively 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 if and only if $\phi_\mu(\theta) = 0$ for all $\theta \in \Theta$.*

Proof. Suppose $\mu \in \mathcal{P}_2(\Theta)$ optimal and let $\zeta \in L^1(\Theta; \mu)$. Then, for all $\nu := \zeta\mu + \nu^\perp \in \mathcal{M}(\Theta)$ (Lebesgue decomposition of ν with respect to μ with $\zeta \in L^1(\Theta; \mu)$) and owing to Lemma 8, as t goes to 0,

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

Hence as μ is optimal

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

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

$$\phi_\mu = 0 \ \mu\text{-almost everywhere}, \quad \phi_\mu = 0 \ \nu^\perp\text{-almost everywhere} \tag{51}$$

for all $\nu^\perp \perp \mu$. As ϕ_μ is continuous, this is equivalent to $\phi_\mu = 0$ everywhere in Θ. Indeed, let $\theta \in \Theta$. If θ belongs to $\text{Supp}(\mu)$, 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 $\theta_\varepsilon \underset{\varepsilon \to 0}{\longrightarrow} \theta$, using the continuity of ϕ_μ, we obtain $\phi_\mu(\theta) = 0$. If $\theta \notin \text{Supp}(\mu)$, then $\delta_\theta \perp \mu$ and necessarily, $\phi_\mu(\theta) = 0$. The reverse implication is trivial.

Conversely suppose now $\phi_\mu = 0$ everywhere in Θ and take $\nu \in \mathcal{P}_2(\Theta)$, then by previous computations and the convexity of $\mathcal{E}$ (slopes are increasing)

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

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 $\mathcal{P}_2(\Theta)$, it holds that*

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

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

Proof. Here we focus on v_μ, the proof for ϕ_μ being very similar. Considering (29)-(30), one can decompose v_μ as

$$v_\mu =: v_{\mu,1} + v_2 + v_{\mu,3}, \tag{52}$$

with

$$\begin{aligned} v_{\mu,1} &:= \nabla_\theta \left[\langle \nabla_x P_\tau(\mu), \nabla_x \Phi_\tau(\theta; \cdot)\rangle_{L^2(\Omega)}\right], \\ v_2 &:= \nabla_\theta \left[-\langle f, \Phi_\tau(\theta; \cdot)\rangle_{L^2(\Omega)}\right], \\ v_{\mu,3} &:= \nabla_\theta \left[\int_\Omega P_\tau(\mu)(x)dx \times \int_\Omega \Phi_\tau(\theta; x)dx\right]. \end{aligned}$$

Using standard integral derivation and Fubini theorems, it holds that for all $\gamma \in \Gamma_o(\mu, \nu)$,

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

Owing to (33)-(34), one gets

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

where $C(\tau)$ is a positive constant which only depends on τ, and where we used the Cauchy-Schwartz inequality. For the third term in the decomposition (52), one has :

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

Owing to (31), one gets :

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

where we used again the Cauchy-Schwartz inequality. Hence the desired result. □

Proposition 7. *Let $\mu \in \mathcal{P}_2(\Theta)$ such that there exists $\theta \in \Theta$, $\phi_\mu(\theta) \neq 0$. 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 under the form

$$\phi_\mu(\theta) =: a\psi_\mu(\omega) + cr_\mu.$$

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 there exists ω such that $\psi_\mu(\omega) \neq 0$ or $r_\mu \neq 0$. Suppose that $\psi_\mu \neq 0$ is non zero somewhere, the case for r_μ being similar. For all $\alpha \in \mathbb{R}$, we denote by

$$\begin{cases} A_\alpha^+ = \psi_\mu^{-1}(]0, +\infty[), \\ A_\alpha^- = \psi_\mu^{-1}(]-\infty, 0[). \end{cases}$$

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 $\partial A_{-\eta}^- = \psi_\mu^{-1}(\{-\eta\})$ is a $(d+1)-$orientable manifold on which $\nabla_\omega \psi_\mu$ is non zero. With our choice of activation function $\sigma_{H,\tau}$, it is easy to prove that $A_{-\eta}^-$ is a bounded set. Indeed, if b is large enough, then $\Omega \ni x \mapsto \sigma_{H,\tau}(w \cdot x + b)$ is zero and $\psi_\mu(w, b)$ is zero.

On $A_{-\eta}^-$, the gradient $\nabla_\omega \psi_\mu$ is pointing outward $A_{-\eta}^-$ and, denoting by n_{out} the outward unit vector to $A_{-\eta}^-$, there exists $\beta > 0$ such that $|\nabla_\omega \psi_\mu \cdot n_{\text{out}}| > \beta$ for on $\partial A_{-\eta}^-$, since this continuous function is nonzero on a compact set. Hence, defining :

$$A := \{(a, c, \omega) \in \Theta \mid \omega \in A_{-\eta}^-,\ a \geq 0\}$$

and owing to the fact that $v_\mu = (v_{\mu,a}, v_{\mu,c}, v_{\mu,\omega})$ with $v_{\mu,a} = \psi_\mu(\omega)$, $v_{\mu,c} = r_\mu$, $v_{\mu,\omega} = a\nabla_\omega \psi_\mu(\omega)$, it holds :

$$\begin{cases} v_{\mu,a} < \eta \text{ on } A \\ v_{\mu,\omega} \cdot n_{out} > \beta a \text{ on } \mathbb{R}_+ \times \mathbb{R} \times \partial A_{-\eta}^-. \end{cases} \tag{53}$$

By contradiction, suppose that $\mu(t_0)$ 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 :

$$|v_{\mu(t)}(\theta) - v_\mu(\theta)| \leq C(\tau, \mu)(1 + |\theta|^2)\varepsilon \tag{54}$$

and

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

One takes $\varepsilon := \sqrt{\dfrac{\eta}{2C(\tau,\mu)R}}$ where $R > 0$ satisfies :

$$(R-1)\mu(t_0)(A) > \int |\theta|^2 d\mu + \frac{\eta}{2C(\tau,\mu)R} \tag{55}$$

which exists since $\mu(t_0)(A) > 0$ by hypothesis. On the set $\{\theta \in A \mid 1 + |\theta|^2 \leq R\}$ and by (54), we have :

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

and so by (53) and the fact that $v_t = -v_{\mu(t)}$:

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

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 $W_2(\mu, \mu(t)) \leq \varepsilon$:

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

and this a contradiction with the condition (55) on R.

□

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

$$A := \{(a, c, \omega) \in \Theta \mid \omega \in A^-_{-\eta_1}\} \cup \{(a, c, \omega) \mid \omega \in A^+_{\eta_2}\} \tag{56}$$

where η_1, η_2 are strictly positive.

Lemma 11. *For all $\mu \in \mathcal{P}_2(\Theta)$, if $\psi_\mu < 0$ somewhere, there exists a strictly negative regular value $-\eta$ ($\eta > 0$) of ψ_μ.*

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

Theorem 8 (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 ϕ_μ and the image of critical points of ϕ_μ is Lebesgue negligible. As a consequence, there exists a point $o \in O$ which is a regular value of ϕ_μ. As $o \in O$, it is strictly negative and this finishes the proof of the lemma.

□

3.3.3 Convergence

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 :*

$$\phi_\mu(\theta) = c r_\mu$$

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 $\phi_\mu = c r_\mu$ where r_μ is a constant. □

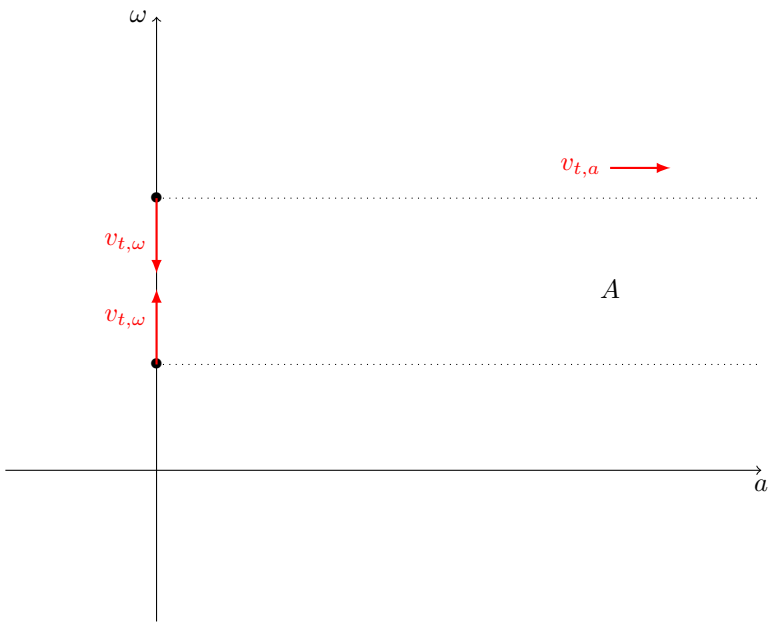


Figure 3: The escape of mass towards large values of a

In fact Hypothesis 1 is verified by the gradient curve $(\mu(t))_{t\geq 0}$ 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 [21]. With simple words, we enjoy the homotopy properties on the sphere to prove that the measure $\mu(t)$ keeps a large enough support.

Proof. For all $t \geq 0$, as $\mu(t) = (\chi(t))\#\mu_0$, we have [2, Lemma C.8] :

$$\text{Supp}(\mu(t)) = \overline{\chi(t)\,(\text{Supp}(\mu_0))}. \tag{57}$$

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, there 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 $\mu(t)$ by (57) 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 and prove Theorem 7.

Proof of Theorem 7. 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 :

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

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.
- ψ_μ is not identically zero and the set A defined in (56) is not empty and verifies :

$$A \subset \mathbb{R}^2 \times S_{\mathbb{R}^d(1)} \times [-\sqrt{d} - 2, \sqrt{d} + 2] \tag{58}$$

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 $(\mu(t))_{t\geq 0}$ 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 (58), 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 the potential of the proposed method.

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 :

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

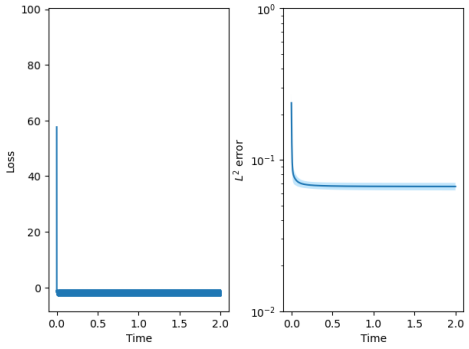
In higher dimension, we use the corresponding source term which is a tensor product of its one dimensional counterpart :

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

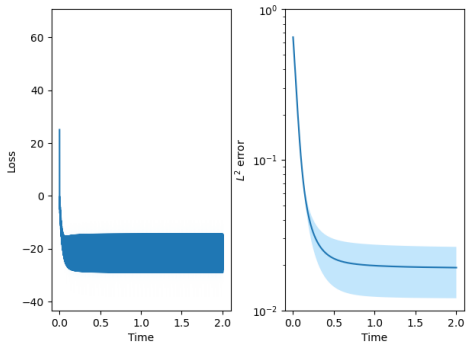
The code is written using python supplemented with Keras/Tensorflow framework. One should remember the following implementation facts :

- 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.

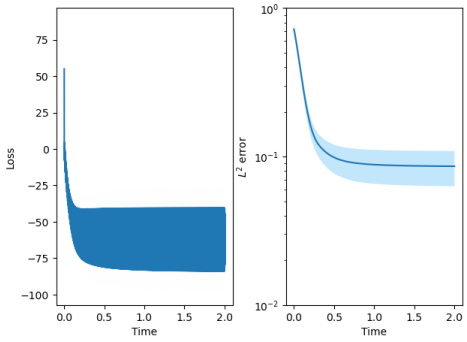
For $d = 1$ and a width $m = 1000$, the simulations are reported in Figure 4 for which very satisfactory results for $k = 1, 3$ are observed, the same conclusions hold for $d = 2$.



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

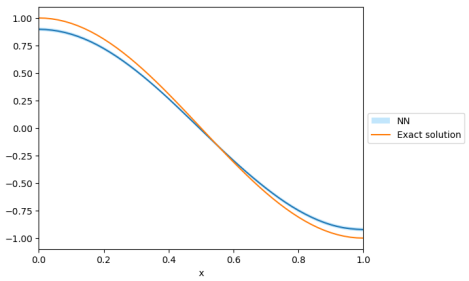


(b) The case $d = 1$ and $k = (3)$

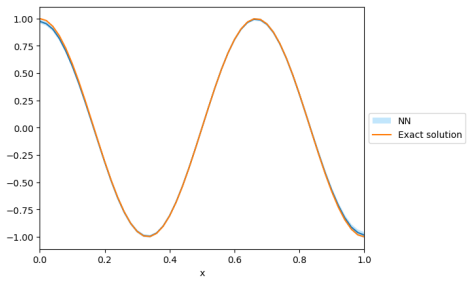


(c) The case $d = 1$ and $k = (5)$

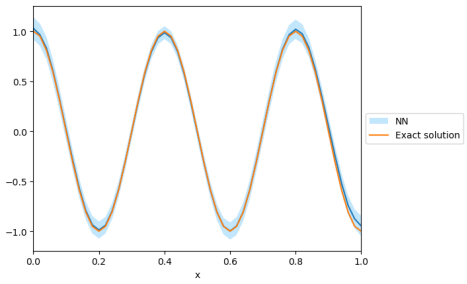
Figure 4: The effect of frequency on the approximation when $d = 1$ and $m = 1000$



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

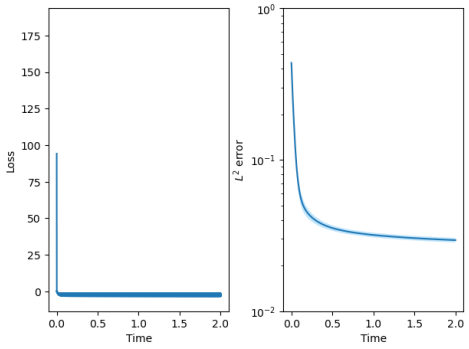


(b) The case $d = 1$ and $k = (3)$

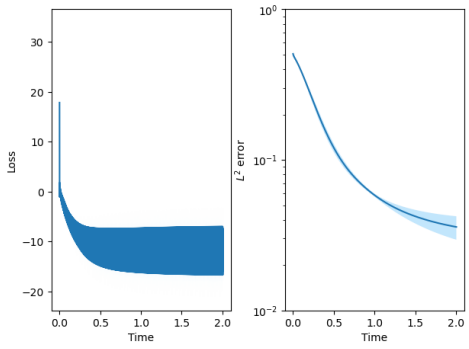


(c) The case $d = 1$ and $k = (5)$

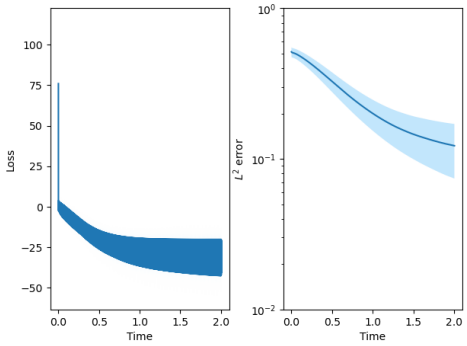
Figure 5: The numerical solutions when $d = 1$ and $m = 1000$



(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 some heuristic arguments for the present choice of scaling related to the learning rate :*

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

It is possible to write the learning scheme as follows :

$$\frac{\theta_{t+1} - \theta_t}{dt} = -\nabla_\theta \phi^n_{\mu_t^m}(\theta_t) \tag{59}$$

where :

$$\phi^n_{\mu_t^m}(\theta) := \frac{1}{nm} \sum_{i,j} \nabla\Phi(\theta_j, x_i) \cdot \nabla\Phi(\theta, x_i) - f(x_i)\Phi(\theta, x_i) + \left(\frac{1}{nm} \sum_{i,j} \Phi(\theta, x_i) \right)^2 \tag{60}$$

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

By analogy, equations (59)-(60) 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 :

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

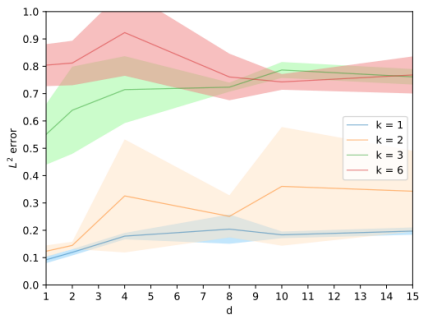
In practice, one can observe that if one takes $dt > O\left(\frac{1}{nm}\right)$ then the scheme diverges in the same way as a classic finite elements scheme.

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 $n, m = 10^2$.

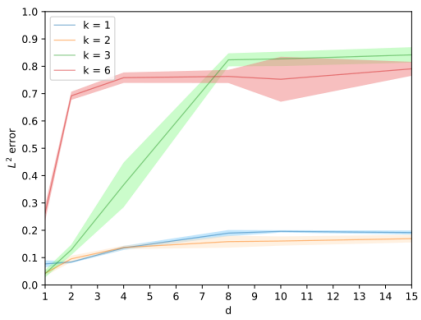
4.2 The effect of dimension

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 :

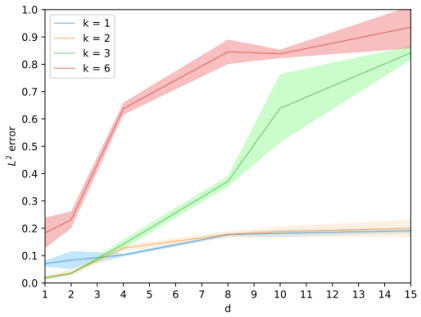
- For low frequency, the precision is not affected by dimension.
- At high frequency, performance are deteriorated as dimension increases.
- Having a larger neural network captures better high frequency modes up to a certain dimension.
- Variance increases with frequency but not with dimension.



(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 $(x_1, x_2, 0.5, \cdots, 0.5)$.

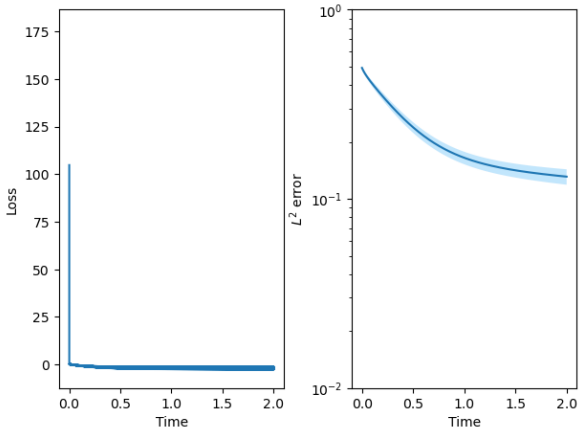


Figure 8: The case $d = 10$, $k = (1, 1, 0, \cdots, 0)$ and $m = 1000$

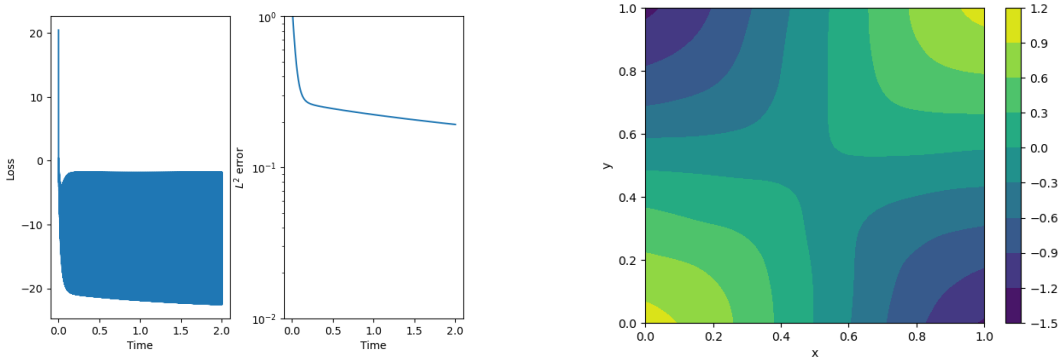
Finally we show an example where a lot of low frequencies are involved in the high dimensional regime :

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

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-layer 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 [18].

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. Roughly, Alexandrov spaces are spaces 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 :*

$$(\alpha, s) : \mathbb{R}_+ \ni t \mapsto \alpha(st)$$

the associated geodesic of velocity s. We then make the identification

$$"(\alpha, 1) = \alpha"$$

unit speed geodesic α.

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 with $\alpha(0) = \beta(0) =: a \in A$ and $s, t \geq 0$. 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,

$$\frac{1}{2st}\left(s^2 + t^2 - \sigma_a((\alpha, s), (\beta, t))\right) \tag{61}$$

depends neither on s nor on t.

The previous theorem is very important as it enables to introduce a notion of angle and scalar product :

Corollary 3. *One can define the local angle $\angle_a((\alpha, s), (\beta, t))$ between (α, s) and (β, 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 :

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

We then have the following definitions.

Definition 7. *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 relationship $\sigma_a = 0$ with respect to the distance σ_a.

The tangent cone, i.e. the set of geodesics departing from a at speed s, of the form (α, s) for some $(\alpha,1) \in \Sigma_a(A)$, is denoted by $C_a(A)$.

A major result from [18] is that if the underlying space A is Alexandrov and compact 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 McCann 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 8. *For a curve $(a_t)_{t\in\mathbb{R}}$ of A, it is said to be differentiable at $t=0$ if there exists $(\alpha,\tau) \in C_a(A)$ such that for all $(\alpha_i,1) \in \Sigma_a(A)$, $t_i \geq 0$ with $\lim_{i\to\infty} t_i = 0$, linking a_0 and a_{t_i} then :*

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

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 :

$$a_0' := (\alpha,\tau).$$

A.2 The notion of gradient

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

Definition 9. *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 [18, Lemma 4.3]). Before doing this, it is necessary to introduce the directional derivative :

Definition 10. *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 $a \in A$ such that $|\nabla_-\,\mathcal{E}\,|(a) < \infty$, there exists a unique direction $(\alpha,1) \in \Sigma_a(A)$ such that :*

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

This direction α is denoted by $\dfrac{\nabla_-\,\mathcal{E}(a)}{|\nabla_-\,\mathcal{E}\,|(a)}$, 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 11. *A Lipschitz curve $(a_t)_{t\geq 0}$ is said to be a gradient curve with respect to $\mathcal{E}$ 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 [18], results about existence and uniqueness of gradient curve on $\mathcal{P}_2(A)$ are given.

Acknowledgements

The authors acknowledge funding from the Tremplin-ERC Starting ANR grant HighLEAP (ANR-22-ERCS-0012).

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] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

[5] Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

[6] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[7] E Weinan, Jiequn Han, and Arnulf Jentzen. Algorithms for solving high dimensional pdes: from nonlinear monte carlo to machine learning. *Nonlinearity*, 35(1):278, 2021.

[8] Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

[9] Bruno Després. *Neural Networks and Numerical Analysis*, volume 6. Walter de Gruyter GmbH & Co KG, 2022.

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

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

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

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

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