Convergence and Geometric properties of Gradient Descent in the training of Deep Residual Neural Networks

Raphaël Barboni [*][†]
raphael.barboni@ens.fr

September 20, 2021

Abstract

This work studies the problem of the convergence of gradient descent in the training of Residual Neural Networks in the infinite depth limit where the model is defined by the integration of an ODE driven by a parametrized residual term. We present analytical tools for studying such continuous models in a supervised learning setting. Thereby, we derive functional inequalities of Polyak-Lojasiewicz type for the associated loss and recover local convergence results that already exist for Neural Networks with finite depth.

Observing that the parametrization we use defines a structure of *Reproducing Kernel Hilbert Space (RKHS)* on the space of residual transformations, we leverage tools from the study of groups of Diffeomorphisms in order to redefine our models as the action of diffeomorphisms on the inputs. This allows us to study a model acting on densities by push-forward action, for which we derive functional inequalities of Kurdyka-Lojasiewicz type. Finally, studying the linearization around identity of this model establishes a connection with a class of dissipative PDEs interpreted as gradient flows on the space of densities provided with a right-invariant metric.

Keywords: Neural Networks, Residual Networks, Neural ODE, Polyak-Lojasiewicz inequalities, Overparametrized models, Implicit bias, Right-invariant metric, Flow of diffeomorphisms

1 Introduction

Residual Networks (ResNet) [30] are Artificial Neural Network architectures for which each layer consists in the the addition of a *skip connection* and a *residual term.* A generic example for such architecture, with L layers, is:

$$\begin{array}{rcl} F(W,x) & = & W^{(L)}x^{(L)}, \\ x^{(l+1)} & = & x^{(l)} + f(W^{(l)}, x^{(l)}) \quad \forall l \in [\![0, L-1]\!], \\ x^{(0)} & = & x, \end{array}$$

where $x \in \mathbb{R}^d$ is the input data, $W = (W_l)_{0 \leq l \leq L}$ is the parameter of the model and $f : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^d$ is a transformation called *residual term.* Those are typically *Multi-Layer Perceptrons (MLP)*, *fully connected layers*, *convolutional layers* or several such layers

[*] Département Mathématiques et Applications - ENS Ulm

[†] École Normale Supérieure Paris-Saclay - MSc. Mathématiques, Vision, Apprentissage

stacked in a row. The addition, at each layer's output, of the preceding layer's output is a *skip connection.* Note that such skip connections force the dimension of each layer's output to stay constant, so that in practice, architectures alternate between residual layers with skip connections and *intermediary layers* where the dimension of the outputs can be modified.

ResNets are considered in a *supervised learning* setting where one is provided with a training set of input data $(x^i)_{1\leq i\leq N}$ and objective data $(y^i)_{1\leq i\leq N} \in (\mathbb{R}^d)^N$. Given a loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, "training" the model consists in finding a parameter W minimizing the (regularized) *Empirical Risk*:

$$L(W) := \sum_{i=1}^{N} \ell(F(W, x^i), y^i) + \lambda \mathcal{R}(W),$$

where $\mathcal{R}(W)$ is a regularizing term and λ is a positive constant. In our work, λ will be set to zero in order to study the implicit regularization induced by the optimization method, which will be a *Gradient Descent* or a *Gradient Flow* on the parameters.

Overparametrization The presence of skip connections in ResNet architectures allows to deal with the problem of vanishing gradient encountered in other "deep" architectures [52] and doing so to use models with increasingly more layers and parameters [60]. The performances of residual networks, and more generally the performances of overparametrized models, have drawn the attention of the Machine learning community as they seem to be efficiently trained by "vanilla" first order optimization methods such as gradient descent. In the past years, several works addressed the problem of proving the convergence of gradient descent in the training of Deep Neural Networks [19, 18, 1, 63, 62, 44], generally concluding that a polynomial dependency of the width of intermediary layers w.r.t. the number of data N and the number of layers L is sufficient to ensure convergence.

An important argument is that along the gradient descent, these models can be approximated by their linearization at initialization, therefore acting as linear models [31, 35]. Following the work of [38, 37], we will analyse our model in the light of functional inequalities of Polyak-Lojasiewicz type (see [10] and references therein), exhibiting conditions on the loss at initialization to transition towards this "linear regime".

Continuous models Similarities between ResNet architectures and discrete numerical schemes motivated the introduction of continuous models [15, 56] as well as a new approaches inspired from the theory of dynamical systems and optimal control [14, 57, 21]. The models considered here are continuous models, thought to be the limiting models for ResNet whose depth L tends to infinity:

$$F(W, x) = x_1, \quad \text{with} \left\{ \begin{array}{rcll} \frac{d}{dt}x_t & = & f(W_t, x_t) & \forall t \in [0, 1] \\ x_0 & = & x, & \end{array} \right.$$

where $x \in \mathbb{R}^d$ is the input data, $W = (W_t)_{t\in[0,1]}$ is the parameter of the model and $f : \mathbb{R}^m \to \mathbb{R}^d$ is the residual term.

Studying those continuous models brings the hope of understanding the behavior of very deep ResNet architectures, something which is for now much misunderstood. Whereas proofs of convergence of gradient descent in the training of discrete models abound, there are only few for continuous ones. Moreover, their mathematical analysis has interest on its own as it is different from the one of discrete models and is in some way more elegant.

We considered in this work a simple set of residual terms consisting in the composition of a fixed non-linearity with a trained linear layer, observing that it amounts to consider vector-fields belonging to a *Reproducing Kernel Hilbert Space (RKHS)*. This establishes a link with the theory of *Large Deformation Diffeomorphic Metric Matching (LDDMM)* where the flow of such vector fields is used in order to solve Image Registration problems [53, 54, 59]. The range of the model's possible outputs corresponding to a set of diffeomorphic deformations of the inputs, the model thus belongs to the class of "reversible" models [13]. The idea was for example used in [50] for building models of *Normalizing Flows* [33] in a generative modeling perspective.

1.1 Contributions

Our main contribution is to show that the overparametrized setting and the proof scheme exposed in [38] can be extended to the framework of continuous models [15, 56]. Doing so we recover for continuous ResNets already existing results for the convergence of gradient descent in the training of discrete models. In particular, we show that the transition towards a "linear regime" with constant *Neural Tangent Kernel* [31] happens for a sufficiently small initial loss. An important drawback of our results is that they only deal with identity (i.e. $W^0 = 0$) initialization whereas most of the literature consider random initializations.

For linear models, the threshold condition for convergence in Theorem 3.1 is numerically explicit with respect to the parameters of the problem and can be enforced in a way that is independent to the number of data points N. This is an improvement with respect to lots of results from the literature [19, 18, 1, 63, 62, 44] but is most likely due to the linear structure of the considered problem.

For models with non-linear residual terms, assumptions on the regularity and the expressivity of these residual terms allow to recover similar convergence results in Theorem 5.2. In this case, the condition for convergence should generally depend on the number of data points. Results already exist in the literature for the convergence of gradient descent in the training of discrete ResNets with non-linear residual terms [18, 1, 62, 37]. However, unlike for the linear model, those do not pass to the limit of infinite depth as they assume at least a polynomial dependence of the intermediary layers' width w.r.t. the depth L of the network. In that sense our result brings a new contribution.

Using a parametrization of the residual terms with vector fields belonging to an abstract RKHS allows us to make an interesting connection with tools from Image Registration [59, 53, 8]. The models we study implement diffeomorphic transformations of the inputs and belong in particular to the class of "reversible" models such as *Normalizing Flows*. The use of such an RKHS parametrization was already suggested in [47] and [50].

Finally, we took interest in the limiting problem where N tends towards infinity by studying models acting on densities by pushforward action of diffeomorphisms [50]. In this setting we show convergence results for unregularized density matching in Property 6.4 and Property 6.5, with possible applications to generative modeling. Studying the linearization of this model around identity, we explored a link with the study of a certain class of dissipative PDEs. Those evolution equations can be interpreted as gradient flows for a right invariant metric and, unlike gradient flows for the Wasserstein metric, the asymptotic behavior of their solutions is not well understood yet. We argue that this could be a first step towards a general proof for the convergence of gradient descent in the training of continuous ResNets.

2 Convergence of overparametrized models and NTK analysis

A common feature of Residual Networks models is their large number of trainable parameters compared to the number of data parameters they are trained on, up to 100 times more in [60]. In a broader picture, they take place in the actual context where machine learning models have a growing number of parameters, up to trillions of parameters recently [24]. An intriguing phenomenon which is not fully understood, is the capability of simple, first-order, optimization methods such as Gradient Descent or Stochastic Gradient Descent to efficiently train such models so as to achieve zero training loss. We will in this section give an informal explanation based on [38].

In the next section, our machine learning models will describe the action of a parameter $v \in \mathbb{R}^m$ on an input object x, belonging to an input space $\mathcal{N}$, resulting in an output object y, belonging to a real d-dimensional riemannian manifold $\mathcal{M}$. This action is defined as a function of (v, x) with value in the output space $\mathcal{M}$:

$$F : \left\{ \begin{array}{ccc} \mathbb{R}^m \times \mathcal{N} & \to & \mathcal{M} \\ (v, x) & \mapsto & F(v, x) =: y. \end{array} \right.$$

We will assume F to be differentiable w.r.t. v and will call y the output object.

Those models belong to the class of parametric models. Provided a differentiable loss function $\ell : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$, we will be interested in the optimization task:

$$\text{Find} \quad v^* \in \underset{v \in \mathbb{R}^m}{\arg\min}\, L(v) := \sum_{i=1}^{N} \ell(F(v; x^i); y^i) \tag{1}$$

for a family of labeled input data $(x^i)_{1 \leq i \leq N} \in \mathcal{N}^N$ and objective data $(y^i)_{1 \leq i \leq N} \in \mathcal{M}^N$.

However, we are here only interested in an optimization over the variable v with fixed input and output data. These can be omitted to rewrite the model as a function of the parameters with values in the cartesian product of the output space $\tilde{\mathcal{M}} := \mathcal{M}^N$:

$$\tilde{F} : \left\{ \begin{array}{ccc} \mathbb{R}^m & \to & \tilde{\mathcal{M}} \\ v & \mapsto & (F(v; x^i))_{1 \leq i \leq N}. \end{array} \right.$$

We say that the model is *overparametrized* whenever $m \gg dN = \dim(\tilde{\mathcal{M}})$.

Considering the loss function $\tilde{\ell} : \tilde{y} = (\tilde{y}^i)_{1 \leq i \leq N} \in \tilde{\mathcal{M}} \mapsto \sum_{i=1}^{N} l(\tilde{y}^i; y^i)$, Equation (1) can be restated as:

$$\text{Find} \quad v^* \in \underset{v \in \mathbb{R}^m}{\arg\min}\, L(v) = \tilde{\ell}(\tilde{F}(v)). \tag{2}$$

Remark 2.1 (Implicit Regularization)
Note that we will not consider here any regularization term on the loss L. Whereas one generally adds a regularization term to the training loss in order to ensure good generalization properties once the model is trained, we are here interested in the implicit regularization that is induced by the optimization method.

In order to minimize the loss L, we will be considering optimization by continuous gradient flow or by discrete gradient descent which we define below:

Definition 2.1 (Gradient Flow)
Given $v \in \mathbb{R}^m$, the gradient flow of L starting at v is defined as the solution of the ODE:

$$\begin{cases} \frac{d}{d\tau} v^\tau &= -\nabla L(v^\tau) \\ v^0 &= v. \end{cases}$$

Definition 2.2 (Gradient Descent)
Given $v \in \mathbb{R}^m$ and a step size $\eta > 0$, the gradient descent of L starting at v is defined as the sequence:

$$\begin{cases} v^{k+1} &= v^k - \eta \nabla L(v^k) \quad \textit{for every } k \geq 0 \\ v^0 &= v. \end{cases}$$

2.1 Neural Tangent Kernel analysis

In order to study the convergence properties of the gradient descent, [38] outlined the importance of studying the conditioning of the *Neural Tangent Kernel (NTK)*, which represents the metric implicitly induced by the model's parametrization on $\tilde{\mathcal{M}}$. The concept of NTK was first introduced by [31] and then broadly studied in the literature of Neural Networks [18, 35, 37].

Indeed, in the problem of training the (parametric) model $\tilde{F}$, one optimizes its parameters, the variable $v \in \mathbb{R}^m$, in order to minimize a loss which is defined on $\tilde{\mathcal{M}}$. This raises the question of whether this optimization method is efficient or not. One can answer this question by considering a small variation δv of the parameters induced by the gradient flow:

$$\delta v = -\nabla L(v) = -D\tilde{F}(v)^\top \nabla \tilde{\ell}(\tilde{F}(v)).$$

This induces a variations on the output through the model:

$$\delta \tilde{y} := D\tilde{F}(v)\delta v = -D\tilde{F}(v) D\tilde{F}(v)^\top \nabla \tilde{\ell}(\tilde{y}),$$

with $\tilde{y} := \tilde{F}(v)$.

Therefore, the Neural Tangent Kernel can be defined as this new metric appearing on the tangent space of $\tilde{\mathcal{M}}$:

$$K(v) := D\tilde{F}(v) D\tilde{F}(v)^\top.$$

Thus, if $K(v)$ becomes degenerate then the optimization is not efficient whereas if $K(v)$ has large eigenvalues the gradient flow is very efficient in order to minimize the loss. In fact, uniform conditioning of the NTK over the entire parameter space implies the convergence of the gradient flow towards a global minima of the loss [38], which can explain the efficiency of gradient descent in the training of overparametrized models as whenever $m \gg dN = \dim(\tilde{\mathcal{M}})$, the NTK $K(v) = D\tilde{F}(v) D\tilde{F}(v)^\top$ is a positive definite form.

However, this definition of the NTK does not provide an efficient way to instantiate it in practice, in particular for complex models such as Deep Neural Networks, where the differential $D\tilde{F}$ can be hard to evaluate. For example, for the continuous models we will consider in the following, the space of parameters will typically be an infinite dimensional functional space such as an L^2 space. Therefore we will not directly evaluate the NTK but rather rely on functional properties which implicitly express its conditioning.

3 Linear Residual Networks

Linear Neural Networks are Neural Networks whose layers consist only of matrix vector products. Even though the class of functions that can be represented by such a model only consists of linear transformations and is thus very restricted, its mathematical analysis is interesting to carry out in the first place precisely because the model's linear structure allows to rely on well-known linear algebra and simplifies the proofs.

Given a data matrix $X \in \mathbb{R}^{d \times N}$ and a set of matrix parameter $W = (W_l)_{1 \leq l \leq L}$ respectively of size $n_l \times n_{l-1}$, the output of a Linear Neural Network with L layers parametrized by W is given by:

$$F(W) = BW_L W_{L-1} ... W_1 AX,$$

with (optional) fixed dimension scaling matrices $A \in \mathbb{R}^{n_0 \times d}, B \in \mathbb{R}^{d' \times N_L}$.

Analogously, a Linear Residual Network is defined as a ResNet whose residual terms consist of linear transformations. It can be considered as a Linear Neural Network whose parameters are defined around the identity matrix I_q:

$$F(W) = B\left(I_q + \frac{1}{L}W_L\right) ... \left(I_q + \frac{1}{L}W_1\right) AX, \tag{3}$$

where the compatibility condition for intermediary dimensions is $n_0 = ... = n_L = q$, and where the factor $1/L$ is a rescaling factor.

Our model of interest here is a continuous linear model which corresponds to the infinite depth limit $L \to +\infty$ of the preceding:

Definition 3.1 (Linear-FlowResNet)
Given a data matrix $X \in \mathbb{R}^{d \times N}$ and (optional) dimension scaling matrices $A \in \mathbb{R}^{q \times d}$, $B \in \mathbb{R}^{d' \times q}$, we define for any $W \in L^2([0,1], \mathbb{R}^{q \times q})$ the output of the Linear-FlowResNet model by:

$$F(W) := BU_1 AX$$

with U the unique absolutely continuous solution to the forward problem*:*

$$\begin{cases} U_0 &= I_q \\ \dot{U}_t &= W_t U_t. \end{cases} \tag{4}$$

W will therefore be called control parameter*.*

Proof. Note that existence and uniqueness of an absolutely continuous solution of Equation (4) is implied by the Caratheodory existence theorem (cf. Section A). □

Remark 3.1
On the space of matrices we consider the Frobenius norm, noted $\|.\|$. The norm on the control parameter space will be the one induced by the Frobenius norm:

$$\|W\|_{L^2} := \left(\int_0^1 \|W_t\|^2 dt\right)^{1/2}.$$

Spectral norm will be noted $\|.\|_2$ and singular values $\sigma_i(.)$.

We will study this model in a *supervised learning* setting where we want to fit data labels living in $\mathbb{R}^{d'}$. That is given an objective data matrix $Y \in \mathbb{R}^{d' \times N}$ and a control parameter W, we will consider the quadratic loss:

$$L(W) := \frac{1}{2}\|F(W) - Y\|_F^2 = \frac{1}{2}\|BU_1AX - Y\|_F^2 \tag{5}$$

and the corresponding Empirical Risk Minimization problem:

$$\text{Find} \quad W^* \in \underset{W \in L^2([0,1], \mathbb{R}^{q \times q})}{\arg\min} \quad L(W). \tag{ERM}$$

3.1 Convergence result for the linear model

Linear convergence properties for the discrete model Equation (3) have already been extensively studied [7, 58, 63, 38]. However, extending the proof to our continuous model requires a different kind of analysis as our parameters live in the infinite dimensional functional space $L^2([0,1]\,, \mathbb{R}^{q\times q})$. At the end, we recover a result that is equivalent to the one of [63] for discrete linear models:

Theorem 3.1 (Local convergence of linear FlowResNet)
Let r be the rank of the data matrix X and define L^* as the square distance between the data labels and the vector space spanned by achievable outputs:

$$L^* := \inf_{U \in \mathbb{R}^{q \times q}} \frac{1}{2}\|BUAX - Y\|^2 = \frac{1}{2}\|BU^*AX - Y\|^2.$$

Assume that at initialization $W^0 := 0$ there exists a radius $R \geq 0$ such that:

$$\frac{\sigma_{max}(A)\sigma_{max}(B)\sigma_{max}(X)}{\sigma_{min}(A)^2\sigma_{min}(B)^2\sigma_r(X)^2}\left[L(W^0) - L^*\right]^{1/2} \leq \frac{Re^{-3R}}{4}. \tag{6}$$

Then for a sufficiently small step size η, gradient descent converges towards a global optimum of the loss in linear time. For every discrete time step $k \geq 0$:

$$\left[L(W^k) - L^*\right] \leq \left(1 - \eta e^{-2R}\sigma_{min}(B)^2\sigma_{min}(A)^2\sigma_r(X)^2\right)^k \left[L(W^0) - L^*\right].$$

Moreover, the gradient descent dynamic is bounded by R: $\|W^k\| \leq R$ for every $k \geq 0$.

Remark 3.2
Be careful not to be misled ! The notation U^* does not mean that there always exist a control parameter W^* satisfying $U_1(W^*) = U^*$. Orthogonal projection of the data emphasizes the fact that we can here benefit from the problem linear structure to have a better a priori estimate on the range of expectable results. This will not be the case for non-linear models.

Remark 3.3 (Convergence of Stochastic Gradient Descent)
Though it won't be discussed here, general theory for overparametrized models satisfying local Polyak-Lojasiewicz inequalities of Property 3.1 also shows a similar convergence result for a training with Stochastic Gradient Descent [38]. We observed this convergence in practice in Section 9.

Theorem 3.1 expresses a behavior of "local" convergence in the sense that the gradient descent converges at a linear rate towards a global minimum of the loss under the condition that it has been initialized sufficiently close from such a global minimum. The condition ensuring this convergence in Equation (6), thus corresponds to a threshold between two kinds of behavior:

- ◇ Equation (6) is not satisfied and one cannot tell anything about convergence,
- ◇ the loss at initialization is sufficiently small and there is convergence at a linear rate.

A limiting behavior is when $L(W^0) \to 0$. Then, in Theorem 3.1, implicit regularization induced by gradient descent allows to bound the dynamic in a ball of radius $R \to 0$ and, all along the gradient descent, the model is well approximated by its first order development at initialization:

$$F(W^k) \simeq F(W^0) + DF(W^k)(W^k - W^0)$$

This "linear regime" or "kernel regime" has been well studied in the literature [31, 37, 35].

Remark 3.4
Note that the condition in Equation (6) *can be enforced by embedding the data in a space of sufficiently high dimension q. Such a method has already been introduced in [20] in order to remove topological impossibilities, [61] proved its universality. In our case, we show that such embedding allows to break symmetries and reduces the "geometric complexity" of the problem.*

Introducing for every $n \geq 0$ the matrices $A_n := (I_d, \dots, I_d)^\top \in \mathbb{R}^{nd\times d}$, $B_n := (I_d, 0, \dots, 0) \in \mathbb{R}^{d\times nd}$ and the subspace of admissible results of the flow $E_n := \{(Y^\top * ... *)^\top\} \subset \mathbb{R}^{nd\times N}$ one can define the embedded data $\tilde{X}_n = A_n X$ and its orthogonal projection $\Pi_{E_n}^{\perp} \tilde{X}_n = (Y^\top X^\top ... X^\top)^\top$. Then the embedding dimension n as no impact on the euclidean distance:

$$\|\tilde{X}_n - \Pi_{E_n}^{\perp}\tilde{X}_n\|^2 = \|X - Y\|^2 = 2L(W^0).$$

However, taking the point of view of linear deformations of the space, the "geometric" distance is lowered. Calculating the unoriented angle:

$$\begin{aligned}\left(\tilde{X}_n, \Pi_{E_n}^{\perp}\tilde{X}_n\right) &= \arccos\left(\frac{\langle \tilde{X}_n, \Pi_{E_n}^{\perp}\tilde{X}_n\rangle}{\|\tilde{X}_n\|\|\Pi_{E_n}^{\perp}\tilde{X}_n\|}\right) \\ &= \arccos\left(\frac{\langle X, Y\rangle + (n-1)\|X\|^2}{(n\|X\|^2)^{1/2}\left(\|Y\|^2 + (n-1)\|X\|^2\right)^{1/2}}\right) \\ &\xrightarrow[n\to+\infty]{} 0.\end{aligned}$$

Remark 3.5 ("Lazy" or "NTK" regime ?)
The condition in Equation (6) *can be read in terms of transition towards a "lazy training" regime as described in [17]. Indeed, the criterion used by the authors for a generic model F is:*

$$\kappa(W^0) := \|F(W^0) - Y\|\frac{\|D^2F(W^0)\|}{\|DF(W^0)\|^2} \ll 1.$$

An upper bound for the l.h.s. in our case can precisely be shown to be:

$$\kappa(W^0) \leq \frac{\sigma_{max}(A)\sigma_{max}(B)\sigma_{max}(X)}{\sigma_{min}(A)^2\sigma_{min}(B)^2\sigma_r(X)^2}\left[L(W^0) - L^*\right]^{1/2}.$$

Therefore, our result seems to exhibit a threshold between the "lazy regime" where linear convergence is achieved and a "rich regime" where we can find examples of non-convergence (see Section 8). However, the model can't be described by the mechanisms presented in [17] where transition towards the "lazy regime" is achieved through rescaling of the parameters. Here, the l.h.s. in Equation (6) *is asymptotically invariant by rescaling.*

Instead, our model's training dynamic should be compared to the ones observed in [31, 38, 37] where near constancy of the neural tangent kernel is implied by an important scale difference between DF and its differential at initialization:

$$\frac{\|D^2F(W^0)\|}{\|DF(W^0)\|^2} \leq \frac{\sigma_{max}(A)\sigma_{max}(B)\sigma_{max}(X)}{\sigma_{min}(A)^2\sigma_{min}(B)^2\sigma_r(X)^2} \ll 1.$$

However, in contrast to the analysis of [37], for our continuous model the Hessian D^2F does not vanish because of an increase in the dimension of the control parameter (which is infinite anyway) but because of an increase in the dimension q of the embedding space. This argues in the sense that, properly rescaled, "depth" plays no role in the benefits of overparametrization.

Remark 3.6 (Independence w.r.t. the number of data points)
Consider that the input data points are drawn independently from a multivariate gaussian distribution and that the objective data points are all set to 0. *Then the condition in Equation* (6) *does not depend on the number of data points N.*

Indeed, if $X_N = (1+x_i^j)_{1\leq i\leq d}^{1\leq j\leq N}$ with i.i.d. random variables $x_i^j \sim \mathcal{N}(0,1)$ and $Y_N = 0 \in \mathbb{R}^{d'\times N}$, one has asymptotically almost surely $rank(X_N) = d$ and:

$$\frac{1}{1+d}\sigma_{\max}(X_N)^2 \sim \sigma_d(X_N)^2 \sim \frac{1}{d}L(F(W^0) \sim N.$$

This independence is an improvement with respect to lots of results ([19, 18] for example) but is due to the specific structure of our linear model.

3.2 Local Polyak-Lojasiewicz property

In contrast with classical proofs for the convergence of gradient descent in the optimization of parametric models, the proof of Theorem 3.1 will here not rely on convexity assumptions for the loss landscape. In fact, as pointed out in [38], because of overparametrization, the loss landscape is typically not convex, with a non-discrete infinite set of global minima, typically a $(m-d)$-dimensional manifold. Instead, one can rely on a local Polyak-Lojasiewicz property [39], which is a functional property satisfied by the loss L. This kind of property is weaker than convexity, but still allows to recover convergence results for the gradient descent.

Property 3.1 (Local Polyak-Lojasiewicz property)
Consider the loss $L : L^2([0,1], \mathbb{R}^{q\times q}) \to \mathbb{R}_+$ defined in Equation (5). *Then there exists*

strictly positive continuous functions $m, M : \mathbb{R}_+ \to \mathbb{R}_+$ such that for every control parameter W:

$$\begin{array}{rcl} \|\nabla L(W)\|^2 & \geq & m(\|W\|_{L^2})\,[L(W) - L^*]\,, \\ \|\nabla L(W)\|^2 & \leq & M(\|W\|_{L^2})\,[L(W) - L^*]\,. \end{array} \tag{7}$$

We say that L satisfies a local Polyak-Lojasiewicz property.

Note that this property prevents the existence of spurious (bounded) local minima of the loss function. However it has a local nature as m and M typically become degenerate as the control parameter W tends towards $+\infty$, which makes this property different from classical Polyak-Lojasiewicz or Kurdyka-Lojasiewicz type inequalities (see [10]). In particular, in constrat with strict convexity, it is not sufficient to prove unconditional linear convergence (see Section 8) of gradient descent.

In the following, considering the Linear-FlowResNet model of Definition 3.1, we show that Property 3.1 is satisfied by the quadratic loss L defined in Equation (5). We first express the gradient of the loss with respect to the control parameter W using *adjoint sensitivity method*:

Lemma 3.1
The gradient of the loss L defined in Equation (5) *w.r.t. the control parameter W is given by:*

$$\nabla L(W) = -PU^\top, \tag{8}$$

where the adjoint variable *P is given the unique solution of the* backward problem*:*

$$\left\{ \begin{array}{rcl} P_1 & = & -B^\top B\,(U_1 - U^*)\,AXX^\top A^\top \\ \dot{P}_t & = & -W_t^\top P_t. \end{array} \right. \tag{9}$$

Proof. Introducing the adjoint variable P, the Lagrangian of Problem ERM writes:

$$\begin{aligned} \mathcal{L}(U, P, W) &= \frac{1}{2}\|BU_1AX - Y\|^2 + \int_0^1 \langle P_t, \dot{U}_t - W_tU_t\rangle dt \\ &= L^* + \frac{1}{2}\|B\,(U_1 - U^*)\,AX\|^2 + \int_0^1 \langle P_t, \dot{U}_t - W_tU_t\rangle dt \\ &= \frac{1}{2}\|BU_1AX - Y\|^2 + [\langle P_t, U_t\rangle]_0^1 - \int_0^1 \left[\langle \dot{P}_t, U_t\rangle + \langle P_tU_t^\top, W_t\rangle\right] dt. \end{aligned}$$

Therefore, adjoint sensitivity theory gives the gradient of the loss w.r.t. W as:

$$\nabla L(W) = \nabla_W \mathcal{L}(U, P, W) = -P_tU_t^\top,$$

with P given by extremal condition:

$$\nabla_P \mathcal{L}(U, P, W) = 0 \iff \left\{ \begin{array}{rcl} P_1 & = & -B^\top B\,(U_1 - U^*)\,AXX^\top A^\top \\ \dot{P}_t & = & -W_t^\top P_t. \end{array} \right.$$

□

The next step consists in exhibiting basic a priori estimates for solutions of the forward and backward problem of Equations (4) and (9):

Lemma 3.2
Let U and P be the solutions to the forward and backward problem for the control parameter W, then for every $t \in [0,1]$:

$$\sqrt{q}e^{-\int_0^t \|W_s\|ds} \leq \|U_t\| \leq \sqrt{q}e^{\int_t^1 \|W_s\|ds}, \tag{10}$$

$$e^{-\int_0^t \|W_s\|ds} \leq \sigma_{\min}(U_t) \leq \sigma_{\max}(U_t) \leq e^{\int_0^t \|W_s\|ds}, \tag{11}$$

$$e^{-\int_t^1 \|W_s\|ds}\|P_1\| \leq \|P_t\| \leq e^{\int_t^1 \|W_s\|ds}\|P_1\|. \tag{12}$$

Proof. Let us begin with the first inequality. $\|U\|^2$ is absolutely continuous as is U and therefore is almost everywhere differentiable with:

$$\frac{d}{dt}\|U_t\|^2 = 2\langle W_t U_t, U_t\rangle.$$

Therefore we have for every $t \in [0,1]$:

$$\begin{aligned} \|U_t\|^2 &= q + 2\int_0^t \langle W_s U_s, U_s\rangle ds \\ &\leq q + 2\int_0^t \|W_s\|\|U_s\|^2 ds \\ &\geq q - 2\int_0^t \|W_s\|\|U_s\|^2 ds. \end{aligned}$$

Leading to the first line of inequalities using Grönwall lemma both forward and backward in time. The proof of the third line of inequalities is the same.

For the second line of inequalities, let us consider a vector $x \in \mathbb{R}^q$ with unit norm. Then $\|Ux\|^2$ is an absolutely continuous function with derivative :

$$\frac{d}{dt}\|U_t x\|^2 = 2\langle W_t U_t x, U_t x\rangle \leq 2\|W_t\|\|U_t x\|^2.$$

Therefore using Grönwall lemma leads us to:

$$\|U_t x\| \leq e^{\int_0^t \|W_s\|ds},$$

which implies the desired upper bound.

For the lower bound, observe that as long as $\|U_t x\|^2$ does not vanish, $1/\|Ux\|^2$ is absolutely continuous with derivative:

$$\frac{d}{dt}\left[\frac{1}{\|U_t x\|^2}\right] = -\frac{2\langle W_t U_t x, U_t x\rangle}{\|U_t x\|^4} \leq \frac{2\|W_t\|}{\|U_t x\|^2}.$$

This leads to:

$$\|U_t x\|^{-2} \leq e^{\int_0^t \|W_s\|ds}.$$

Therefore it never vanishes, giving the result by inverting both terms in the inequality and taking the infimum over unit vectors x. □

Remark 3.7
Note that this second line of inequalities implies that (U_t) is a flow on the space of linear isomorphisms. This is to be compared with the results of Section 5 where we will consider a model generating flows of diffeomorphisms of $\mathbb{R}^d$.

We now have enough tools to prove Property 3.1:

Proof of the local PL property. We begin with the upper bound:

$$\begin{aligned}\|P_t U_t^\top\|^2 &\leq \sigma_{\max}(U_t)^2 \|P_t\|^2 \\ &\leq e^{2\int_0^1 \|W_s\| ds} \|P_1\|^2 \\ &\leq e^{2\|W\|_{L^2}} \|P_1\|^2,\end{aligned}$$

where we used Lemma 3.2 in the second inequality and Cauchy-Schwarz inequality in the third line.

Then by definition of the backward problem in Equation (9), noting $V := B(U_1 - U^*)AX$:

$$\begin{aligned}\|P_1\|^2 &= \|B^\top V A^\top X^\top\|^2 \\ &\leq \sigma_{\max}(A)^2 \sigma_{\max}(B)^2 \sigma_{\max}(X)^2 \|V\|^2 \\ &= 2\sigma_{\max}(A)^2 \sigma_{\max}(B)^2 \sigma_{\max}(X)^2 \left[L(W) - L^*\right].\end{aligned}$$

Therefore, we proved the desired upper-bound with constant:

$$M(\|W\|_{L^2}) := 2\sigma_{\max}(A)^2 \sigma_{\max}(B)^2 \sigma_{\max}(X)^2 e^{2\|W\|_{L^2}}. \tag{13}$$

The proof for the lower-bound is the same:

$$\begin{aligned}\|P_t U_t^\top\|^2 &\geq \sigma_{\min}(U_t)^2 \|P_t\|^2 \\ &\geq e^{-2\|W\|_{L^2}} \|P_1\|^2 \\ &\geq e^{-2\|W\|_{L^2}} \sigma_{\min}(A)^2 \sigma_{\min}(B)^2 \sigma_r(X)^2 \left[L(W) - L^*\right],\end{aligned}$$

where in the third inequality we used the fact that if X is of rank r their exists a full rank matrix $\tilde{X} \in \mathbb{R}^{d \times r}$ such that $XX^\top = \tilde{X}\tilde{X}^\top$ and $\sigma_r(X) = \sigma_{\min}(\tilde{X})$. Therefore, for any matrix M:

$$\begin{aligned}\|MX\|^2 &= Tr(M^\top M X X^\top) \\ &= Tr(M^\top M \tilde{X}\tilde{X}^\top) \\ &= \|M\tilde{X}\|^2 \geq \sigma_{\min}(X)^2 \|M\|^2.\end{aligned}$$

Therefore, we proved the desired lower-bound with constant:

$$m(\|W\|_{L^2}) := 2\sigma_{\min}(A)^2 \sigma_{\min}(B)^2 \sigma_r(X)^2 e^{-2\|W\|_{L^2}}. \tag{14}$$

□

3.3 Proof of Theorem 3.1

Considering the gradient flow of the loss function L, the lower-bound and the upper-bound in the local PL property allow the control of both the evolution of the loss and the evolution of the parameter along the gradient dynamic. Therefore, assuming Property 3.1 is in fact sufficient to recover the result of Theorem 3.1 for the gradient flow.

However, as it is the way optimization is performed in numerical applications we will in this section detail the proof of convergence for gradient descent with a discrete step-size η. In particular, this forces us to show a supplementary smoothness property in order to ensure that this discrete dynamic is a good approximation of the continuous gradient flow:

Lemma 3.3 (Smoothness)
There exists a positive function $\mathbf{C} : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$*, non-decreasing w.r.t. both of its variables, such that for two control parameters* W*,* W' *with* $\|W\|_{L^2}, \|W'\|_{L^2} \leq R$*, with a certain radius* $R \geq 0$ *the following inequality holds:*

$$\|\nabla L(W) - \nabla L(W')\|_{L^2} \leq \mathbf{C}(R, L(W))\|W - W'\|_{L^2}. \tag{15}$$

Proof. We note U and U' the solutions of the forward problem and P, P' the solutions of the backward problem for control parameters W and W' respectively. We also note $\Delta U_t := U_t' - U_t$ and $\Delta P_t := P_t' - P_t$ for every time $t \in [0, 1]$.

We have:

$$\begin{aligned}\frac{d}{dt}\Delta U_t &= W_t' U_t' - W_t U_t \\ &= W_t' \Delta U_t + (W_t' - W_t)U_t.\end{aligned}$$

Therefore, as soon as $\Delta U \neq 0$:

$$\begin{aligned}\frac{d}{dt}\|\Delta U_t\| &= \frac{1}{\|\Delta U_t\|}\left[\langle W_t'\Delta U_t, \Delta U_t\rangle + \langle (W_t' - W_t)U_t, \Delta U_t\rangle\right] \\ &\leq \|W_t'\|\|\Delta U_t\| + \sigma_{\max}(U_t)\|W_t' - W_t\| \\ &\leq \|W_t'\|\|\Delta U_t\| + \|W_t' - W_t\| e^{\int_0^t \|W\|}.\end{aligned}$$

Then using Grönwall's inequality:

$$\begin{aligned}\|\Delta U_t\|_F &\leq e^{\int_0^t \|W_s'\| ds}\int_0^t \|W_s' - W_s\| e^{\int_0^s (\|W_r\| - \|W_r'\|)dr} ds \\ &\leq e^{\int_0^t (\|W_s\| + \|W_s'\|)ds}\int_0^t \|W_s' - W_s\| ds \\ &\leq e^{2R}\|W' - W\|_{L^2}.\end{aligned}$$

The same idea works for ΔP :

$$\frac{d}{dt}\Delta P_t = -(W_t')^\top \Delta P_t - (W_t' - W_t)^\top P_t.$$

Therefore:

$$\|\Delta P_t\| \leq \|\Delta P_1\| e^{\int_t^1 \|W_s'\| ds} + e^{\int_t^1 (\|W_s\| + \|W_s'\|)ds}\int_t^1 \|W_s' - W_s\| ds.$$

Then using the initial condition in the backward problem and the bound on ΔU:

$$\begin{aligned}\|\Delta P_1\| &= \|B^\top B \Delta U_1 A X X^\top A^\top\| \\ &\leq \sigma_{\max}(A)^2 \sigma_{\max}(B)^2 \sigma_{\max}(X)^2 \|\Delta U_1\| \\ &\leq \sigma_{\max}(A)^2 \sigma_{\max}(B)^2 \sigma_{\max}(X)^2 e^{2R}\int_0^1 \|W_s' - W_s\| ds,\end{aligned}$$

which gives for any time $t \in [0, 1]$:

$$\|\Delta P_t\| \leq e^{2R}\left[1 + \sigma_{\max}(A)^2 \sigma_{\max}(B)^2 \sigma_{\max}(X)^2 e^R\right]\|W' - W\|_{L^2}.$$

Finally recalling the upper-bound on U' and P:

$$\sigma_{\max}(U'_t) \leq e^{\int_0^t \|W'_s\| ds} \leq e^R,$$
$$\|P_t\| \leq \sqrt{2}\sigma_{\max}(A)\sigma_{\max}(B) \left[L(W) - L^*\right]^{1/2} \sigma_{\max}(X)e^R,$$

one can derive the result as:

$$\begin{aligned}
\|\nabla L(W') - \nabla L(W)\|_{L^2}^2 &= \int_0^1 \|P'_t(U'_t)^\top - P_t U_t^\top\|^2 dt \\
&\leq \int_0^1 \|P_t(U'_t - U_t)^\top + (P'_t - P_t)(U'_t)^\top\|^2 dt \\
&\leq \int_0^1 \left(\|P_t\|\|\Delta U_t\| + \sigma_{\max}(U_t)\|\Delta P_t\|\right)^2 dt \\
&\leq \mathbf{C}(R, L(W))^2 \|W' - W\|_{L^2}^2.
\end{aligned}$$

Finally, we found found the constant:

$$\begin{aligned}
\mathbf{C}(R, L) := &\sqrt{2}e^{3R}\sigma_{\max}(A)\sigma_{\max}(B)\sigma_{\max}(X)\left[L - L^*\right]^{1/2} \\
&+ e^{3R}\left(1 + \sigma_{\max}(A)^2\sigma_{\max}(B)^2\sigma_{\max}(X)^2 e^R\right).
\end{aligned}$$

□

As a consequence, ∇L is locally-uniformly Lipshitz, thus the hessian of the loss is defined almost everywhere and for $\|W\|, \|W'\| \leq R$:

$$L(W') - L(W) \leq \langle \nabla L(W), W' - W \rangle_{L^2} + \frac{1}{2}\mathbf{C}(R, L(W))\|W' - W\|_{L^2}^2.$$

Finally, one can prove Theorem 3.1:

Proof of Theorem 3.1. Consider a gradient descent step size $\eta > 0$ whose value is to be defined. The considered dynamic reads for every discrete time step $k \geq 0$:

$$W^{k+1} = W^k - \eta \nabla L(W^k)$$

Assume that the condition of Equation (6) is verified at identity initialization $W^0 = 0$ for some $R \geq 0$. We will prove the result by induction on $k \in \mathbb{N}$.

The result trivially holds for $k = 0$. Then for $k \geq 0$, since for every $l \leq k$, $\|W^l\|_{L^2} \leq R$ and $L(W^l) \leq L(W^0)$:

$$\begin{aligned}
L(W^{k+1}) - L(W^k) &\leq -\eta\|\nabla L(W^k)\|_{L^2}^2 + \frac{\eta^2 \mathbf{C}}{2}\|\nabla L(W^k)\|_{L^2}^2 \\
&\leq -\frac{\eta}{2}\|\nabla L(W^k)\|_{L^2}^2 \\
&\leq -\frac{\eta m(R)}{2}\left[L(W^k) - L^*\right],
\end{aligned}$$

where $\mathbf{C} = \mathbf{C}(R, L(W^0))$ and we chose $\eta \leq 1/\mathbf{C}$. Whence by induction:

$$L(W^{k+1}) - L^* \leq (1 - \frac{\eta m(R)}{2})^{k+1}\left[L(W^0) - L^*\right]$$

which is the desired result considering the form of $m(R)$ in Equation (14).

Moreover:

$$\begin{aligned}
\|W^{k+1}\|_{L^2} &= \| \sum_{l=0}^{k} \eta \nabla L(W^l) \|_{L^2} \\
&\leq \eta \sum_{l=0}^{k} M(R)^{1/2} \left[L(W^k) - L^* \right]^{1/2} \\
&\leq \eta \sum_{l=0}^{k} M(R)^{1/2} \left[L(W^0) - L^* \right]^{1/2} \left(1 - \frac{\eta m(R)}{2} \right)^{l/2} \\
&\leq \eta M(R)^{1/2} \left[L(W^0) - L^* \right]^{1/2} \frac{1}{1 - \sqrt{1 - \eta m(R)/2}} \\
&\leq 4 \frac{M(R)^{1/2}}{m(R)} \left[L(W^0) - L^* \right]^{1/2},
\end{aligned}$$

where we used the inequality $1/(1 - \sqrt{1-x}) \leq 2/x$ for $0 < x \leq 1$.

Then, recalling the form of $m(R)$ and $M(R)$ in Equation (14) and Equation (13), we recover $\|W^{k+1}\|_{L^2} \leq R$ because:

$$\begin{aligned}
& \frac{\sigma_{max}(A)\sigma_{max}(B)\sigma_{max}(X)}{\sigma_{min}(A)^2 \sigma_{min}(B)^2 \sigma_r(X)^2} \left[L(W^0) - L^* \right]^{1/2} \leq \frac{Re^{-3R}}{4} \\
\iff \quad & 4 \frac{M(R)^{1/2}}{m(R)} \left[L(W^0) - L^* \right]^{1/2} \leq R,
\end{aligned}$$

which is exactly Equation (6), thus concluding the proof. □

4 Non-linear Residual Networks

As the network model of the previous section can only generate a linear deformation of the inputs, it is of limited interest for concrete applications. In this section we will consider non-linear models, which means that the output is a non-linear transformation of the input. This is more inline with the way Residual Networks are used in practice. Indeed, in applications, each layer of a Neural Network includes a non-linear transformation, often applied component-wise, such as ReLU or Sigmoid. For a L layer ResNet, a generic model writes

$$\begin{aligned}
F((W,U);x) &= W^{(L)} x^{(L)}, \\
x^{(l+1)} &= x^{(l)} + W^{(l)} \sigma(U^{(l)} x^{(l)}) \quad \text{for } l \in [\![0, l-1]\!], \\
x^{(0)} &= x,
\end{aligned} \tag{16}$$

where $x \in \mathbb{R}^d$ is the data input, $(W^{(l)})_l$ and $(U^{(l)})_l$ is a family of parameters with $W^{(l)} \in \mathbb{R}^{d \times m_l}$, $U^{(l)} \in \mathbb{R}^{m_l \times d}$ for $0 \leq l < L$ and $W^L \in \mathbb{R}^{d' \times d}$. $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linearity applied component-wise on vector-valued inputs.

Remark 4.1

The term x^l is called a skip connection *and the term $W^l \sigma(U^l x^l)$ is the* residual term, *which is here a* Multi-Layer Perceptron (MLP) *with one hidden layer. Note that in practice, one can consider a broad variety of transformations as residual terms.* Convolutional layers *are for example particularly used for applications in imaging [30].*

We will consider a simplification by only training the second linear layer of each residual term. This leads us to the following continuous model which corresponds to the infinite depth limit of the previous one:

Definition 4.1 (Non-linear FlowResNet)
Consider a function $\varphi : \mathbb{R}^d \to \mathbb{R}^q$. Then for a control parameter $W \in L^2([0,1], \mathbb{R}^{d\times q})$ and a data input $x \in \mathbb{R}^d$ we define the Non-linear FlowResNet model's output as:

$$F(W, x) := x_t,$$

where x_1 is defined as the time one solution of the forward problem:

$$\begin{cases} \dot{x}_t &= W_t \varphi(x_t) \\ x_0 &= x. \end{cases} \tag{17}$$

Remark 4.2
One could have defined the continuous model in an other way by choosing to put the matrix-vector product inside the non-linearity in that manner:

$$\dot{x}_t = \varphi(W_t x_t),$$

or even by using two matrix-vector product placed before and after the non-linearity:

$$\dot{x}_t = W_t \varphi(U_t x_t).$$

We are aware that this is a considerable restriction of our model as by definition the feature map φ is fixed and non-trainable whereas the addition of a trained hidden layer would have allowed optimization between a range of different feature maps. However, we motivate our choice by the fact that in our definition the driving term in the forward ODE Equation (17) *is linear w.r.t. the control parameter which simplifies the analysis. In particular the gradient of the loss won't depend on W itself.*

As for the linear model we will consider the non-linear model in a supervised learning setting, trying to solve the following Empirical Risk Minimization problem:

$$\text{Find} \quad W^* \in \underset{W \in L^2([0,1],\mathbb{R}^{d\times q})}{\arg\min} \quad L(W), \tag{ERM}$$

where the quadratic loss L is defined for a family of input data $(x^i)_{1\leq i\leq N} \in (\mathbb{R}^d)^N$ and a family of labeled output data $(y^i)_{1\leq i\leq N} \in (\mathbb{R}^d)^N$ as:

$$L(W) = \frac{1}{2}\sum_i \|F(W, x^i) - y^i\|^2. \tag{18}$$

4.1 Convergence result for the non-linear model

The problem of convergence of gradient descent towards a global optimizer for a ResNet model with non-linear residual terms has been addressed by several authors in the literature. [19, 18] showed convergence of randomly initialized ResNets at a linear rate as long as the width m of the intermediary layers is sufficiently big. The same result was also proved by [37] in a more general setting, relating the proof to the phenomenon of constancy of the Neural Tangent Kernel [31]. However both works need polynomial or exponential dependency of the width m w.r.t. the number of layers (or depth) L suggesting that the result won't scale to our continuous model.

Still, we show a convergence result that is similar to the one of the previous section by considering the following (strong) assumption:

Assumption 4.1 (N-universality)
Let us consider a function $\varphi : \mathbb{R}^d \to \mathbb{R}^q$. For a family of N data points $(x^i)_{1\leq i\leq N} \in (\mathbb{R}^d)^N$ we define the kernel matrix $\mathbb{K}$ of φ at $(x^i)_i$ by the symmetric block matrix:

$$\mathbb{K}((x_i)_{1\leq i\leq N}) := \big(\langle \varphi(x^i), \varphi(x^j)\rangle I_d\big)_{1\leq i,j\leq N} \, .$$

We say that φ is N-universal if for every family of N two-by-two disjoint data points $(x^i)_{1\leq i\leq N}$ their exist constants $\lambda, \Lambda > 0$ such that:

$$\lambda \leq \lambda_{\min}(\mathbb{K}) \leq \lambda_{\max}(\mathbb{K}) \leq \Lambda,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denotes respectively the smallest and the greatest eigenvalue.

Remark 4.3
A necessary condition for φ to be N-universal is in particular $q \geq N$. Indeed, it needs $(\varphi(x^i))_{1\leq i\leq N}$ to be independent for every family of two-by-two disjoint points $(x^i)_{1\leq i\leq N}$. This shows that, because of the lack of a specific structure on the set of generated deformations such as in the linear case, convergence generally depends on the number of data points.

On the other hand, Assumption 4.1 can always be ensured by choosing a function φ taking value in a sufficiently high dimensional feature space. *More precisely, it can be ensured by choosing a feature space of dimension $q = \mathcal{O}(N^d)$, with N the number of data points and d their dimension. We recover thereby a result that is similar to the ones already existing for discrete models [18, 1, 62], a major difference being that because of the continuous form of our model, the dimension of the feature space does not have to depend one the depth L of the network.*

Indeed, consider the polynomial feature map of degree n:

$$\varphi : \left\{ \begin{array}{ccc} \mathbb{R}^d & \to & \mathbb{R}^{q(d,n)} \\ x & \mapsto & (x^\alpha)_{0\leq|\alpha|\leq n}, \end{array} \right.$$

with for every $\alpha \in \mathbb{N}^d$, $x^\alpha = x_1^{\alpha_1}...x_d^{\alpha_d}$ and *$|\alpha| = \alpha_1 + ... + \alpha_d$. The dimension of the feature space is therefore given by:*

$$q(d,n) := \#\{\alpha \in \mathbb{N}^d \mid |\alpha| \leq n\} = \sum_{k=0}^{n} \binom{k+d-1}{d-1} = \mathcal{O}_{n\to+\infty}(n^d).$$

Then consider a family of N two-by-two disjoint data points $(x^i)_{1\leq i\leq N} \in (\mathbb{R}^d)^N$. The positivity of the kernel matrix $\mathbb{K}$ is equivalent to the linear independence of the family $(\varphi(x^i))_{1\leq i\leq N}$ in $\mathbb{R}^{d(\alpha+1)}$. Therefore, consider $p_1, ..., p_N$ real numbers such that:

$$p_1\varphi(x^1) + ... + p_N\varphi(x^N) = 0.$$

Then for every exponent $0 \leq |\alpha| \leq n$:

$$p_1(x^1)^\alpha + ... + p_N(x_k^N)^\alpha = 0,$$

and by summing linear combinations of this equality, we have that for every multivariate polynomial $P \in \mathbb{R}[X_1, ..., X_d]$ of degree at most n:

$$p_1 P(x^1) + ... + p_N P(x^N) = 0.$$

Because the points are two by disjoints, for each index $i \geq 2$ there exists a coordinate k_i such that $x^i_{k_i} \neq x^1_{k_i}$. Then consider the polynomial of degree $N-1$:

$$P(X_1, ...X_d) = \prod_{i=2}^{N} \frac{(X_{k_i} - x^i_{k_i})}{(x^1_{k_i} - x^i_{k_i})},$$

such that $P(x^i) = 0$ for every $i \geq 2$, but $P(x^1) = 1$, implying $p_1 = 0$. Choosing $n = N-1$ allows to consider such a polynomial and we recover $p_2 = ... = p_N = 0$ in the same manner, i.e. the kernel matrix $\mathbb{K}$ is positive definite.

Because it is a corollary of Theorem 5.2 we state the obtained result in an informal manner:

Theorem 4.1
Suppose that the input data points $(x^i)_{1 \leq i \leq N}$ are two-by-two disjoint and that the non-linear activation function φ is N-universal. Then their exist a threshold constant $L^0 > 0$ such that if the loss at identity initialization $W^0 := 0$ satisfies:

$$L(W^0) \leq L^0,$$

then gradient flow (or gradient descent with a sufficiently small step-size) converges towards 0 *a zero training loss global optimum with a linear convergence rate.*

5 General model : Deep ResNet with RKHS parametrization

In the preceding sections, the forward ODEs Equations (4) and (17) can be viewed as transport equations driven by a time-dependent vector field which, given the non-linearity φ, belongs to the space $V = \{v : x \in \mathbb{R}^d \mapsto W\varphi(x),\ W \in \mathbb{R}^{d \times q}\}$. Considering the Frobenius norm on matrices then endows V with a structure of (vector-valued) *Reproducing Kernel Hilbert Space (RKHS)* with the scalar-product:

$$\forall v, v' \in V, \quad \langle v, v' \rangle_V = \langle W, W' \rangle_F.$$

Note that φ is the feature map associate to V and the associated kernel is:

$$\forall x, x' \in \mathbb{R}^d \quad K(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathbb{R}^q} I_d,$$

which corresponds to V being a RKHS of finite dimension. In that setting, we saw that local convergence could be ensured as soon as the feature map satisfies Assumption 4.1, requiring to consider a feature space whose dimension increases (at least linearly) with the number of data points N. In order to recover a convergence result for arbitrary large N, one therefore needs to consider as a feature space an infinite dimensional Hilbert space, whose associated RKHS is then infinite dimensional (see Section 7).

These considerations motivate the introduction of a new (theoretical) model where the transport vector-field is chosen freely in an abstract, possibly infinite-dimensional, RKHS. We only imposes for such RKHS to be *admissible* as defined in Assumption 5.1.

Definition 5.1 (RKHS-FlowResNet)
Let V be an admissible RKHS of vector-field over $\mathbb{R}^d$. Then for a control parameter $v \in L^2([0,1], V)$ and a data input $x \in \mathbb{R}^d$, the model RKHS-FlowResNet's output is defined as:

$$F(v,x) := x_1,$$

where x_1 is the time-one solution of the following forward problem:

$$\begin{cases} \dot{x}_t &= v_t(x_t) \\ x_0 &= x. \end{cases} \tag{19}$$

Remark 5.1
In this model, the space of control parameters is $L^2([0,1], V)$, the space of squared integrable time-dependent vector-fields of V, endowed with the hilbertian structure:

$$\forall v, v' \in L^2([0,1], V), \quad \langle v, v' \rangle_{L^2} = \int_0^1 \langle v, v' \rangle_V dt.$$

As before we will consider this model in a supervised learning setting. For families of data inputs $(x^i)_{1 \leq i \leq N} \in (\mathbb{R}^d)^N$ and data outputs $(y^i)_{1 \leq i \leq N} \in (\mathbb{R}^d)^N$ we will address the question of convergence of gradient descent (or gradient flow) for the minimization of the Empirical Risk associated with the quadratic loss:

$$\text{Find} \quad v^* \in \underset{v \in L^2([0,1],V)}{\arg\min} \frac{1}{2} \sum_{i=1}^{N} \|F(v; x^i) - y^i\|^2. \tag{ERM}$$

Remark 5.2
As for the linear model in Section 3, one could consider to put fixed dimension expansion matrices A, B at the input and the output of our model, probably leading to similar results where the condition for convergence can be enforced by embedding the data in sufficiently high dimension. However, as it does not change the proof and for the sake of simplicity we chose not to do so.

In the case where A is fixed and B is trained, [44] used a similar inductive proof to show uniform conditioning of the NTK and convergence at a linear rate for sufficiently wide networks. Such a proof can be easily adapted to our model with similar results.

5.1 Right invariant metric and flow of diffeomorphisms

Using a parametrization of the residual terms of our model with an abstract RKHS opens an interesting connection with mathematical tools concerning the study of groups of diffeomorphisms and already used for solving Image Registration problems [59, 8].

From now on we will assume that the following regularity assumption is satisfied by the RKHS V.

Assumption 5.1 (Admissible RKHS)
Let V be an RKHS of vector-field over $\mathbb{R}^d$. We say that V is admissible *if it can be continuously embedded in $W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$, the space of bounded vector-fields with bounded differential, i.e. there exist a constant $\kappa > 0$ such that:*

$$\forall v \in V, \quad \|v\|_\infty + \|dv\|_\infty \leq \kappa \|v\|_V.$$

We will note the embedding as $V \hookrightarrow W^{1,\infty}$.

Then the solution $(x_t)_{t\in[0,1]}$ of the forward problem in Equation (19) can be interpreted as the flow of the time-varying vector field v starting from $x_0 = x$. General theory for the well-posedness of such transport equations has been broadly studied (see [2] for a review). In our case the admissibility assumption $V \hookrightarrow W^{1,\infty}$ imposes to only work with "regular enough" vector-fields:

Property 5.1
Let $v \in L^2([0,1], V)$. Then there exists a unique flow map $\Phi^v : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ so that for any $x \in \mathbb{R}^d$, $\Phi^v_t(x)$ is the unique solution of the ODE:

$$\left\{ \begin{array}{lcl} \frac{d}{dt}\Phi^v_t(x) & = & v_t(\Phi^v_t(x)) \\ \Phi^v_0(x) & = & x. \end{array} \right. \tag{20}$$

Therefore, taking the point of view of the global deformation of the space, we can restate our model in a new way. Setting the control parameter $v \in L^2([0,1], V)$ the output of the RKHS-FlowResNet model F of Definition 5.1 is given, for any input $x \in \mathbb{R}^d$, by:

$$F(v, x) = \Phi^v_1(x),$$

with Φ^v provided by Property 5.1. In fact, in this setting, the ODE defining Φ^v can even be given a rigorous meaning without taking its evaluation into consideration. We will here simply note that Φ^v formally satisfies:

$$\left\{ \begin{array}{lcl} \frac{d}{dt}\Phi^v & = & v_t \circ \Phi^v_t \\ \Phi^v_0 & = & Id. \end{array} \right.$$

That is, Φ^v is the *time-one flow* of the time-varying vector field v.

Considering the set of time one flows Φ^v_1 for $v \in V$:

$$\mathcal{G}_V = \{\Phi^v_1 \mid \partial_t \Phi^v_t = v_t \circ \Phi^v_t,\ \Phi^v_0 = Id,\ v \in L^2([0,1], V)\},$$

A. Trouvé showed this is a group of $\mathcal{C}^1$-diffeomorphisms for the left-action by composition, V being its set of infinitesimal generators [53]. Furthermore, the scalar structure of V provides $\mathcal{G}_V$ with a local metric for which it is invariant by the right-action by composition:

$$\forall v, v' \in V,\ \forall \Phi \in \mathcal{G}_V, \quad \langle v \circ \Phi, v' \circ \Phi \rangle_{T_\Phi \mathcal{G}_V} = \langle v, v' \rangle_{T_{Id}\mathcal{G}_V} = \langle v, v' \rangle_V.$$

The associated right-invariant metric structure is:

$$\forall \Phi_1, \Phi_0 \in \mathcal{G}_V, \quad d(\Phi_1, \Phi_0)^2 = d(\Phi_1 \circ \Phi_0^{-1}, Id)^2 = \inf\{\int_0^1 \|v_t\|_V^2 dt \mid \Phi_1 = \Phi^v_1 \circ \Phi_0\}.$$

In particular, $\mathcal{G}_V$ is proven to be complete for this metric [53, 59].

The work of [11] gives a tighter analysis when the RKHS belongs to the class of Sobolev spaces H^s. In this case, the admissibility assumption imposes $s > d/2 + 1$ and the flow map belongs to the space of Sobolev diffeomorphisms $\mathcal{D}^s(\mathbb{R}^d)$ defined as:

$$\begin{array}{lcl} \mathcal{D}^s(\mathbb{R}^d) & = & \{\Phi \in Id + H^s \mid \Phi \text{ bijective},\ \Phi^{-1} \in Id + H^s\} \\ & = & \{\Phi \in Id + H^s \mid \Phi \in Diff^1(\mathbb{R}^d)\} \\ & = & \{\Phi \in Id + H^s \mid det(D\Phi(x)) > 0, \forall x \in \mathbb{R}^d\}. \end{array} \tag{21}$$

[11] proves the equivalence of these definitions and proves that the space $\mathcal{D}^s$ can be provided with a structure of Hilbert manifold. This is an improvement w.r.t. the work of Trouvé [53] as it is shown that $\mathcal{D}^s$ is geodesically complete for the right-invariant metric in the sense that there always exists a minimizing geodesic between two diffeomorphisms in the same connected component.

Remark 5.3
Recall that the considered manifold $\mathcal{D}^s$ is of infinite dimension. Thus the Hopf-Rinow theorem does not apply and there is no equivalence between metric and geodesic completeness (see [23]).

Finally, the following result concerning the regularity of the flow map will be useful:

Theorem 5.1 (Theorem 4.4 of [11])
Let $s > d/2+1$ and $v \in L^2([0,1], H^s)$. Then v has a $\mathcal{D}^s$-valued flow $\Phi^v \in \mathcal{C}([0,1], \mathcal{D}^s(\mathbb{R}^d))$. Moreover, the flow map:

$$Fl : \left\{ \begin{array}{ccc} L^2([0,1], H^s) & \to & \mathcal{C}([0,1], \mathcal{D}^s(\mathbb{R}^d)) \\ v & \mapsto & \Phi^v \end{array} \right.$$

is continuous.

5.2 Convergence results

We will consider in the following a fixed admissible RKHS V with kernel K. As in the non-linear case an other *universality* assumption is needed in order to recover a local-convergence property:

Assumption 5.2 (Universality of V)
We say that the RKHS V is universal if for every family of two-by-two disjoint points $(x^i)_{1\le i\le N}$ the kernel matrix $\mathbb{K}$ defined by block as:

$$\mathbb{K} = (K(x^i, x^j))_{i,j}$$

is positive definite. We will further assume that there exists constants $\lambda, \Lambda > 0$, λ non-increasing w.r.t. $\min_{i,j} \|x^i - x^j\|$, such that:

$$\lambda(\min_{1\le i,j\le N} \|x^i - x^j\|) \le \lambda_{\min}(\mathbb{K}) \le \lambda_{\max}(\mathbb{K}) \le \Lambda.$$

An equivalent definition is that for every family of vector $(p^i)_{1\le i\le N}$ there exists some $v \in V$ such that:

$$\forall i \in [\![1, N]\!], \quad v(x^i) = p^i.$$

Remark 5.4
This definition forces V to be infinite dimensional. However a broad class of commonly used kernels satisfy this assumption, such as Gaussian kernel, Laplacian kernels, Sobolev kernels, ...

Assuming the universality assumption is satisfied, we prove that a local PL property is satisfied for the Empirical Risk of Problem ERM with the RKHS-FlowResNet model of Definition 5.1:

Property 5.2 (Local Polyak-Lojasiewicz property)
Assume that the RKHS V is admissible and universal. Then there exists positive functions $m, M : \mathbb{R}_+ \to \mathbb{R}_+^*$ such that for every control parameter $v \in L^2([0,1], V)$ it holds:

$$\begin{array}{rcl} \|\nabla L(v)\|_{L^2} & \leq & M(\|v\|_{L^2})L(v), \\ \|\nabla L(v)\|_{L^2} & \geq & m(\|v\|_{L^2})L(v). \end{array} \tag{22}$$

Using the same technique as in the linear case (c.f. [38]), the local PL property allows to prove local convergence of the gradient flow starting from identity initialization:

Theorem 5.2 (Local convergence of the gradient flow)
Assume that the RKHS V is admissible and universal. Assume furthermore that there exists a radius $R > 0$ satisfying at identity initialization $v^0 := 0$:

$$\sqrt{L(v^0)} \leq \sqrt{\frac{m(R)}{M(R)}} R, \tag{23}$$

where m, M are the constants of Property 5.2.

Then the gradient flow starting at v^0 converges towards a global optimum of the loss function with a linear convergence rate. More precisely for every time $\tau \geq 0$ of the gradient flow:

$$L(v^\tau) \leq e^{-c(R)\tau} L(v^0 = 0).$$

Moreover the control parameter is bounded by R along the flow:

$$\|v^\tau\|_{L^2} \leq R, \quad \forall \tau \geq 0.$$

Remark 5.5
To prove the same result for a discrete gradient descent requires an improved regularity of the loss and in particular an upper-bound on the spectral norm of its Hessian. Therefore one would need a stronger regularity assumption on V, such as for example $V \hookrightarrow W^{2,\infty}(\mathbb{R}^d, \mathbb{R}^d)$.

5.3 Proof of Property 5.2

For $x, p \in \mathbb{R}^d$ we will use the notation $\delta_x^p \in V^*$ for the linear form consisting of the scalar product between p and the vector field evaluated at x:

$$\forall v \in V, \quad \langle \delta_x^p, v \rangle = \langle p, v(x) \rangle.$$

By abuse of notation, K will denote both the kernel associated to V and the isometry $K : V^* \to V$, such that for every $x, p \in \mathbb{R}^d$:

$$K * \delta_x^p = K(., x)p.$$

The proof of Property 5.2 will be detailed into several lemmas, the first one expresses the form of the gradient for the quadratic loss associated to Problem ERM.

Lemma 5.1
Consider two input and objective point clouds $(x^i)_{1 \leq i \leq N}$ and $(y^i)_{1 \leq i \leq N}$ and the square loss

associated to Problem ERM. Then for a control parameter $v \in L^2([0,1], V)$ *the gradient of the loss reads:*

$$\nabla L(v) = -\sum_{i=1}^{N} K * \delta_{x^i}^{p^i} = -\sum_{i=1}^{N} K(., x^i)p^i, \tag{24}$$

where the adjoint variables p^i *are the time-one solutions of the backward problem:*

$$\begin{cases} \dot{p}_t^i &= -Dv_t(x_t^i)^\top p_t^i \\ p_1^i &= y^i - x_1^i. \end{cases} \tag{25}$$

Proof. The proof relies on the adjoint sensitivity method. The Lagrangian of the problem is given by:

$$\begin{aligned} \mathcal{L}((x^i),(p^i),v) &= \sum_{i=1}^{N} \left(\frac{1}{2}\|x_1^i - y^i\|^2 + \int_0^1 \langle p_t^i, \dot{x^i}_t - v_t(x_t^i)\rangle dt \right) \\ &= \sum_{i=1}^{N} \left(\frac{1}{2}\|x_1^i - y^i\|^2 + \left[\langle p_t^i, x_t^i\rangle\right]_0^1 - \int_0^1 \left[\langle \dot{p^i}_t, x_t^i\rangle + \langle p_t^i, v_t(x_t^i)\rangle\right] dt \right). \end{aligned}$$

Then Equation (25) is obtained by satisfying the first order condition w.r.t. the x coordinates:

$$\nabla_{x^i}\mathcal{L} = 0,$$

implying:

$$\begin{cases} \dot{p}_t^i &= -Dv_t(x_t^i)^\top p_t^i \\ p_1^i &= y^i - x_1^i. \end{cases}$$

Finally, Equation (24) is obtained by differentiating w.r.t. v:

$$\nabla L = \nabla_v \mathcal{L} = -\sum_{i=1}^{N} K * \delta_{x^i}^{p^i}.$$

□

The second lemma gives a geometric control on the deformations generated by the flow of a control parameter v:

Lemma 5.2
Let $v \in L^2([0,1], V)$ *and let* Φ^v *be the associated flow of diffeomorphisms. Then for every time* $0 \leq t \leq 1$*:*

$$\|D\Phi_t^v\|_\infty, \|(D\Phi_t^v)^{-1}\|_\infty \ \leq \ e^{\kappa\|v\|_{L^2}},$$

where the norms are operator / spectral norms.

Proof. By definition of the flow Φ^v one has for every $x \in \mathbb{R}^d$ and every time t:

$$\frac{d}{dt}\Phi_t^v(x) = v_t(\Phi_t^v(x)).$$

Differentiating this inequality with respect to x gives:

$$\frac{d}{dt}D\Phi_t^v(x) = Dv_t(\Phi_t^v(x))D\Phi_t^v(x).$$

Consider a unit vector $z \in \mathbb{S}^{d-1}$:

$$\begin{aligned}\frac{d}{dt}\|D\Phi_t^v(x)z\|^2 &= 2\langle Dv_t(\Phi_t^v(x))D\Phi_t^v(x)z, D\Phi_t^v(x)z\rangle\\ &\leq 2\|Dv_t\|_\infty\|D\Phi_t^v(x)z\|^2\\ &\leq 2\kappa\|v_t\|_V\|D\Phi_t^v(x)z\|^2,\end{aligned}$$

where we used the embedding $V \hookrightarrow W^{1,\infty}$ in the last inequality. Using Grönwall inequality one deduces:

$$\|D\Phi_t^v(x)z\|^2 \leq \|D\Phi_0^v(x)z\|^2 e^{2\kappa\int_0^t\|v_s\|_V ds} \leq e^{2\kappa\|v\|_{L^2}}.$$

Taking the supremum w.r.t. $z \in \mathbb{S}^{d-1}$ and then $x \in \mathbb{R}^d$ gives the desired upper-bound on $\|D\Phi_t^v\|_\infty$.

For the upper-bound on $\|(D\Phi_t^v)^{-1}\|_\infty$, one can consider a unit vector $z \in \mathbb{S}^{d-1}$ and write as long as $\|D\Phi_t^v(x)z\| > 0$:

$$\begin{aligned}\frac{d}{dt}\|D\Phi_t^v(x)z\|^{-2} &= 2\|D\Phi_t^v(x)z\|^{-4}\langle Dv_t(\Phi_t^v(x))D\Phi_t^v(x)z, D\Phi_t^v(x)z\rangle\\ &\leq 2\kappa\|v_t\|_V\|D\Phi_t^v(x)z\|^{-2}.\end{aligned}$$

Therefore $\|D\Phi_t^v(x)z\| > 0$ for every t and furthermore taking the infimum w.r.t. $z \in \mathbb{S}^{d-1}$:

$$\sigma_{\min}(D\Phi_t^v(x)) \geq e^{-\kappa\|v\|_{L^2}},$$

which gives the result by inverting $D\Phi_t^v$ and taking the supremum w.r.t. $x \in \mathbb{R}^d$.

□

A consequence is that, for any $x^1, x^2 \in \mathbb{R}^d$ and any $t \in [0,1]$, the distance between $\Phi_t^v(x^1)$ and $\Phi_t^v(x^2)$ is being controlled by:

$$\|x^1 - x^2\|e^{-\kappa\|v\|_{L^2}} \leq \|\Phi_t^v(x^1) - \Phi_t^v(x^2)\| \leq \|x^1 - x^2\|e^{+\kappa\|v\|_{L^2}}.$$

For a point cloud $(x^i)_{1\leq i\leq N}$, introducing, for every $t \in [0,1]$, $m_t := \min_{i,j}\|\Phi_t^v(x^i) - \Phi_t^v(x^j)\|$, we have:

$$m_t \geq m_0 e^{-\kappa\|v\|_{L^2}}.$$

Proof of the local PL property. We will note x, p for any pair of data point and adjoint variable solution of the forward and backward problem respectively.

First, notice that it is possible to control the evolution of the adjoint variable p. For $t \in [0,1]$:

$$\begin{aligned} \frac{d}{dt}\|p_t\|^2 &= -2\langle Dv_t(x_t)^\top p_t, p_t\rangle \\ &\leq 2\kappa\|v_t\|_V\|p_t\|^2 \\ &\geq -2\kappa\|v_t\|_V\|p_t\|^2, \end{aligned}$$

which gives:

$$\|p_1\|^2 e^{-2\kappa\|v\|_{L^2}} \leq \|p_t\|^2 \leq \|p_1\|^2 e^{2\kappa\|v\|_{L^2}}.$$

Then considering the set of trajectories (x_t^i) we can consider for every time $t \in [0,1]$ the kernel matrix $\mathbb{K}_t := (K(x_t^i, x_t^j))_{i,j}$. We already have:

$$\min_{1\leq i,j\leq N} \|x_t^i - x_t^j\| \geq e^{-\kappa\|v\|_{L^2}} \min_{1\leq i,j\leq N} \|x_0^i - x_0^j\|.$$

Therefore, because λ is non-increasing and Λ is a constant, along the trajectories it holds:

$$\lambda(m_0 e^{-\kappa\|v\|_{L^2}}) \leq \lambda_{\min}(\mathbb{K}_t) \leq \lambda_{\max}(\mathbb{K}_t) \leq \Lambda.$$

Using properties of the reproducing kernel and considering the form of the gradient in Equation (24), this gives for every time $t \in [0,1]$:

$$\begin{aligned} \|\nabla L(v)_t\|_V^2 = \|\sum_{i=1}^N K(.,x_t^i)p_t^i\|_V^2 &= \sum_{1\leq i,j\leq N} (p_t^i)^\top K(x_t^i, x_t^j)p_t^j \\ &= p_t^\top \mathbb{K}_t p_t \\ &\leq \Lambda\|p_t\|^2 \\ &\geq \lambda(m_0 e^{-\kappa\|v\|_{L^2}})\|p_t\|^2, \end{aligned}$$

where we use the notation $p := (p^i)_{1\leq i\leq N} \in \mathbb{R}^{Nd}$ for the stacked vectors. This gives the result using the control on $\|p_t\|^2$ and the initial condition:

$$\|p_1\|^2 = \sum_{i=1}^N \|p_1^i\| = 2L(v).$$

We found for the constants:

$$\begin{aligned} M(\|v\|_{L^2}) &:= 2\Lambda e^{2\kappa\|v\|_{L^2}}, \\ m(\|v\|_{L^2}) &:= 2\lambda(m_0 e^{-\kappa\|v\|_{L^2}})e^{-2\kappa\|v\|_{L^2}}, \end{aligned}$$

with $m_0 = \min_{i,j}\|x^i - x^j\|$. □

Remark 5.6
Like for the linear model, the bounding function m and M allow to precisely define a threshold for the transition towards a linear regime.

However, in contrast with the linear model, these functions are here hard to explicit and should depend both on the chosen kernel K and on the number of data points N.

6 A generative modeling perspective

We considered in the previous sections a finite dimensional supervised learning setting where the action of V is defined on a finite number of data points N. We saw that in this setting the gradient flow could converge towards a global minima of the Empirical Risk L, assuming a certain condition is satisfied at the initialization. In particular, this condition should generally depend on the number of data points N, raising the question of the behavior of our model as N tends towards infinity. In order to address this problem we took interest in the limiting model where V acts on functional spaces by transporting densities. The associated supervised learning problem is a density fitting problem which is addressed in the Machine Learning literature by Normalizing Flows models [33], with applications in generative modeling.

We will give here rather informal definitions, the precise definitions of the considered spaces depending on the considered problem.

6.1 Push-forward action and problem setting

Let U be a real differentiable manifold of dimension d and $\mathcal{M}(U)$ be a space of measures on U. Then for a diffeomorphism Φ of U, one defines his push-forward action on $\mathcal{M}(U)$ as:

$$\forall \mu \in \mathcal{M}(\mathcal{U}),\ \forall f \text{ test function}, \quad \langle \Phi_{\#}\mu, f\rangle = \langle \mu, f\circ\Phi\rangle.$$

For a control parameter $v \in L^2([0,1], V)$ such that $\Phi = \Phi_1^v$, the associated infinitesimal action writes as the *continuity equation*:

$$\left\{\begin{array}{rcl} \partial_t\mu_t + \nabla\cdot(\mu_t v_t) & = & 0 \\ \mu_0 & = & \mu. \end{array}\right. \tag{26}$$

This is the forward problem for the model defined as $F(v,\mu) := (\Phi_1^v)_{\#}\mu$.

The corresponding backward problem is a transport equation on the space of test functions:

$$\left\{\begin{array}{rcl} \partial_t I_t + \langle v_t, \nabla I_t\rangle & = & 0 \\ I_1 & = & -\partial\ell(\mu_1), \end{array}\right. \tag{27}$$

where ℓ is the loss of interest on the space of measures, classical examples being, among others, the *relative entropy* or the squared *Wasserstein distance.*

We will thus consider the optimization problem associated to the minimization of the loss $L(v) = \ell(F(v,\mu))$.

Property 6.1
Consider a loss functional ℓ on the space of measures. Then the gradient of L w.r.t. v writes:

$$\forall v \in L^2([0,1], V), \quad \nabla L(v) = K * \partial L(v) = -K * (\mu\nabla I), \tag{28}$$

with I the solution of Equation (27) *and K the kernel function associated to the RKHS V.*

Proof. The Lagrangian of the problem is defined as:

$$\begin{aligned}\mathcal{L}(\mu, I, v) &= \ell(\mu_1) + \int_0^1 \langle \partial_t\mu_t - \nabla\cdot(\mu_t v_t), I_t\rangle dt \\ &= \ell(\mu_1) + [\langle \mu_t, I_t\rangle]_0^1 - \int_0^1 \left[\langle \mu_t, \partial_t I_t\rangle + \int_U \langle v_t, \nabla I_t\rangle d\mu_t\right] dt.\end{aligned}$$

Then Equation (27) comes from the first order condition $\partial_\mu \mathcal{L} = 0$ and the gradient in Equation (28) comes from:

$$\nabla L = \nabla_v \mathcal{L} = K * (\mu \nabla I),$$

with the notation $\mu \nabla I \in V^*$ standing for the linear form on V:

$$\langle \mu \nabla I, v \rangle = \int_U \langle v, \nabla I \rangle d\mu.$$

□

6.2 Well-posedness of the continuity equation and regularity estimates

Equation (26) expresses the transport of the density μ along the flow of the time-varying vector-field v. There are many works in the literature addressing the problem of well-posedness of such so-called *continuity equations*, even with minimal regularity assumptions on v [2]. However, in the Sobolev setting $V = H^s$ with $s > d/2+1$, regularity Assumption 5.1 is satisfied and, recalling Property 5.1, we are provided with a flow map Φ which gives simple representation formulas:

Property 6.2
Let μ_0 be an input measure on $\mathbb{R}^d$, then the unique solution of transport Equation (26) *with initial data μ_0 is given by:*

$$\mu_t = (\Phi_t)_\# \mu_0,$$

with Φ the flow map corresponding to v and given by Property 5.1.

The same kind of representation formula holds for a sufficiently regular scalar function I_0, for example $I_0 \in W^{1,\infty}$. The unique solution to Equation (27) *is given by:*

$$I_t = I_0 \circ \Phi_t^{-1}.$$

This property gives in particular an explicit formula in the case where μ_0 is absolutely continuous with respect to the Lebesgue measure $\mathcal{L}$, with density ρ_0. Then μ_t is also absolutely continuous w.r.t. $\mathcal{L}$ for every time $t \in [0,1]$, with density:

$$\rho_t = \left| D(\Phi_t^{-1}) \right| \rho_0 \circ \Phi_t^{-1}.$$

The following lemmas give control over the H^s regularity of the solutions:

Lemma 6.1 (Lemma 2.2 in [11])
Let $s > d/2 + 1$ and $M, C > 0$. Then there exists a constant $C_s(M, C)$ such that for every $\Phi \in \mathcal{D}^s(\mathbb{R}^d)$ satisfying:

$$\inf_{x \in \mathbb{R}^d} \det(D\Phi(x)) > M \quad \textit{and} \quad \|\Phi - Id\|_{H^s} < C,$$

and every $f \in H^s$, we have:

$$\|f \circ \Phi\|_{H^s} \leq C_s(M, C) \|f\|_{H^s}.$$

Lemma 6.2
Let $s > d/2$. Then the product map defined as:

$$\begin{array}{ccc} H^s \times H^{s-1} & \rightarrow & H^{s-1} \\ (f, g) & \mapsto & (fg : x \mapsto f(x)g(x)) \end{array},$$

is a well-defined, bilinear continuous map. That is, for every $f \in H^s$, $g \in H^{s-1}$:

$$\|fg\|_{H^{s-1}} \leq C\|f\|_{H^s}\|g\|_{H^{s-1}}.$$

Remark 6.1
In the following, for a Hilbert space $\mathcal{H}$ we will use the shortcut notation $L^2(\mathcal{H}) = L^2([0,1]\,, \mathcal{H})$. When the considered Hilbert space is obvious we will not write it.

Also the dependency on $\|v\|_{L^2}$ will always be written $C(\|v\|_{L^2})$ for upper-bounds and $c(\|v\|_{L^2})$ for lower-bounds. Therefore, be aware that the precise value of these constants can change between two equations.

These two lemmas allow to derive a regularity result for the gradient of the loss L in our optimization problem:

Property 6.3 (Continuity of the gradient)
Let the initial measure μ_0 be absolutely continuous w.r.t. the Lebesgue measure with density $\rho_0 \in H^s$, $s > d/2 + 1$, and assume that the loss functional ℓ has a continuous differential $\partial\ell : H^{s-1} \rightarrow H^{s-1}$. Then there exists a constant $C(\|v\|_{L^2})$ such that for every control parameter $v \in L^2([0,1]\,, H^s)$:

$$\|\partial L(v)\|_{L^2(H^{s-2})} \leq C(\|v\|_{L^2})\|\rho_0\|_{H^s}.$$

Proof. Because ρ_0 is sufficiently regular, one can use the representation formula of Property 6.2 to get for every $t \in [0,1]$:

$$\mu_t = \rho_t d\mathcal{L} \quad \text{with} \quad \rho_t = \left|D(\Phi_t^{-1})\right| \rho_0 \circ \varphi_t^{-1}.$$

Thus using Lemmas 6.1 and 6.2 one can show that there exists a constant $C(\Phi_t)$ such that:

$$\|\rho_t\|_{H^{s-1}} \leq C(\Phi_t)\|\rho_0\|_{H^s}.$$

By using the continuity of the flow in Theorem 5.1 one can choose to express the dependency over Φ_t as a dependency over $\|v\|_{L^2}$. Thus:

$$\|\rho_t\|_{H^{s-1}} \leq C(\|v\|_{L^2})\|\rho_0\|_{H^s}.$$

Using the assumption that $\partial\ell$ is continuous we have:

$$\|I_1\|_{H^{s-1}} \leq C'\|\rho_1\|_{H^{s-1}} \leq C(\|v\|_{L^2})\|\rho_0\|^{H^s}.$$

Then using using the representation formula for the solution of the backward equation and the continuity of the composition gives:

$$\|I_t\|_{H^{s-1}} \leq C(\|v\|_{L^2})\|\rho_0\|_{H^s}.$$

Finally, using continuity of the product, we obtain the result:

$$\|\partial L(v)\|_{L^2(H^{s-2})} = \|\rho\nabla I\|_{H^{s-2}} \leq C(\|v\|_{L^2})\|\rho_0\|_{H^s}.$$

□

6.3 L^2 distance on the flat-torus

In order to have good equivalence properties between the different considered functional spaces, we choose to consider densities defined on a compact manifold such as the d-dimensional flat torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$. We will consider here a L^2 distance for the loss functional:

$$\forall \rho \in H^s(\mathbb{T}^d), \quad \ell(\rho) = \frac{1}{2}\|\rho - \rho^*\|^2_{L^2(\mathbb{T}^d)},$$

where $\rho^* \in H^s(\mathbb{T}^d)$ is the objective density to fit. Note that in particular ℓ admits a continuous differential $\partial\ell : \rho \mapsto (\rho - \rho^*)$.

In this setting, local PL properties (of a weaker form) can be obtained under mild regularity conditions:

Property 6.4 (Local Kurdyka-Lojasiewicz property)
Let $s > d/2+1$. Assume μ_0 admits a density $\rho_0 \in H^s$ with respect to the Lebesgue measure, that the objective density ρ^* is in H^s and that ρ_0 is uniformly lower-bounded on $\mathbb{T}^d$. Then there exist parameter depending constants $C(\|v\|_{L^2})$ and $c(\|v\|_{L^2})$ as well as some $\lambda, \mu > 0$ such that for every control parameter $v \in L^2((0,1], H^s)$:

$$\begin{array}{rcl} \|\nabla L(v)\|^2_{L^2} & \leq & C(\|v\|_{L^2})L(v)^{1-\mu}, \\ \|\nabla L(v)\|^2_{L^2} & \geq & c(\|v\|_{L^2})L(v)^{1+\lambda}. \end{array} \tag{29}$$

These are called local Kurdyka-Lojasiewicz *inequalities [10, 34].*

Proof. Part 1 : lower-bound

Using the results of Equation (28) and Property 6.3 we have that for every time $\tau \geq 0$ along the gradient flow:

$$\partial L(v)_t = \rho_t \nabla I_t \in H^{s-2}.$$

Then thanks to Gagliardo-Nirenberg interpolation inequalities (or using Parseval formula and Hölder inequality):

$$\|\rho\nabla I\|_{L^2} \leq \|\rho\nabla I\|_{H^{-s}}^{\frac{s-2}{2s-2}} \|\rho\nabla I\|_{H^{s-2}}^{\frac{s}{2s-2}}.$$

Thus setting $\lambda = 1 + 2/(s-2)$:

$$\begin{aligned} \|\partial L(v)\|^2_{V^*} = \|\rho\nabla I\|^2_{H^{-s}} &\geq \|\rho\nabla I\|_{L^2}^{2(1+\lambda)} \|\rho\nabla I\|_{H^{s-2}}^{-4(s-1)/s} \\ &\geq c(\|v\|_{L^2}, \|\rho_0\|_{H^s})\|\rho\nabla I\|_{L^2}^{2(1+\lambda)} \\ &\geq c(\|v\|_{L^2}, \|\rho_0\|_{H^s})\|\nabla I\|_{L^2}^{2(1+\lambda)} \inf_{\mathbb{T}^d} \rho^{2+2\lambda}, \end{aligned}$$

where the second inequality comes from the the upper-bound over $\|\rho\nabla I\|_{H^{s-2}}$ in Property 6.3. Thus, because ρ_0 is uniformly lower-bounded by $m > 0$ and thanks to the representation formula $\rho_t = |D(\Phi_t)^{-1}|\rho_0 \circ \Phi_t^{-1}$, we have for every time $t \in [0,1]$:

$$\inf_{\mathbb{T}^d} \rho_t \geq m e^{-\kappa\|v\|_{L^2}} = c(\|v\|_{L^2}).$$

Then using Poincaré inequality on $\mathbb{T}^d$, there exists a constant c such that:

$$\|\nabla I\|_{H^{-s}}^2 \geq c\|I\|_{L^2}^2,$$

which gives, by recalling that $I_t = I_1 \circ \Phi_1 \circ \Phi_t^{-1}$ and by making a change of variables:

$$\begin{aligned}\|I_t\|_{L^2} &\geq c(\|v\|_{L^2})\|I_1\|_{L^2} \\ &= c(\|v\|_{L^2})L(v)^{1/2},\end{aligned}$$

so that finally:

$$\|\partial L(v)\|_{V^*}^2 \geq c(\|v\|_{L^2})L(v)^{1+\lambda}.$$

Part 2 : upper-bound
We do not detail the proof for the upper-bound which uses similar arguments.

□

Note that the local KL inequalities of Property 6.4 guarantee the absence of spurious local minima as well as the convergence of the gradient flow towards a global minima for bounded dynamics. Indeed assume one has uniformly $c(\|v^\tau\|_{L^2}) \geq c > 0$ for every $\tau \geq 0$, then:

$$\frac{d}{d\tau}L(v^\tau) = -\|\nabla L(v^\tau)\|_{L^2}^2 \leq -cL(v)^{1+\lambda},$$

so that for every $\tau \geq 0$:

$$L(v^\tau) \leq \left[L(v^0)^{-\lambda} + \lambda c\tau\right]^{-1/\lambda} \xrightarrow[\tau\to+\infty]{} 0.$$

However, an inductive proof can no longer guarantee the local convergence of the gradient flow. Indeed assume one has uniformly $C(\|v^\tau\|_{L^2}) \leq C < +\infty$, the control over $\|v^\tau\|_{L^2}$ becomes:

$$\begin{aligned}\|v^\tau\|_{L^2} &\leq \|v^0\|_{L^2} + \int_0^\tau \|v^s\|_{L^2} ds \\ &\leq \|v^0\|_{L^2} + \int_0^\tau C^{1/2}\left[L(v^0)^{-\lambda} + \lambda cs\right]^{-(1-\mu)/2\lambda} ds.\end{aligned}$$

Therefore one would need $\lambda < 1/2$ in order for the r.h.s. to be bounded, and here $\lambda = 1 + 2/(s-2) > 1$.

Nevertheless, this is an expected behaviour. Indeed, consider an objective density such that $\rho^* = 0$ on an open set $U \subset \mathbb{T}^d$ with non-empty interior. Then the lower-bound $\inf \rho_1 \geq e^{-K\|v\|_{L^2}} \inf \rho_0$ imposes $\|v^\tau\|_{L^2} \to +\infty$ at convergence: the generated deformation Φ needs to tear the space in order to make mass disappear on U which behavior can only be obtained in the limit where Φ is no-longer a diffeomorhism.

This supports the fact that the weakening of the PL property into a KL property is inherent to the infinite dimensional setting.

6.4 Relative entropy on the flat torus

We consider as loss functional the relative entropy defined for every $\rho \in H^s$ by:

$$\ell(\rho) = \mathbf{H}(\rho||\rho^*) = \begin{cases} \int_{\mathbb{T}^d} \frac{d\rho}{d\rho^*} \log(\frac{d\rho}{d\rho^*}) d\rho^* & \text{if } d\rho << d\rho^* \\ +\infty & \text{otherwise,} \end{cases}$$

where ρ^* is the objective density. Then the functional derivative of ℓ is given by:

$$\partial\ell(\rho) = \log(\frac{d\rho}{d\rho^*}),$$

that we will assume to be continuous, at least on $\{\rho \in H^s | d\rho << d\rho^*\}$.

We will make the following assumption on the objective density ρ^*:

Assumption 6.1 (Log-Sobolev inequality (see [4]))
We say that ρ^* satisfies the log-sobolev inequality if for every density ρ:

$$\int_{\mathbb{T}^d} \rho \log(\rho/\rho^*) dx \leq 2 \int_{\mathbb{T}^d} \left|\nabla \sqrt{\rho/\rho^*}\right|^2 \rho^* dx.$$

Or in terms of relative entropy $\mathbf{H}$ and relative information $\mathbf{I}$:

$$\mathbf{H}(\rho||\rho^*) \leq \frac{1}{2}\mathbf{I}(\rho||\rho^*).$$

Remark 6.2
Note that we could consider densities defined on any real finite-dimensional compact manifold where log-sobolev inequality results exist.

As for the L^2 loss, one can show the existence of local KL-type inequalities for the relative entropy loss, ensuring the absence of spurious local minima:

Property 6.5 (Local KL property)
Let $s > d/2 + 1$. Let $\rho_0 \in H^s$ be uniformly upper- and lower-bounded on $\mathbb{T}^d$ and let $\rho^* \in H^s$ be an objective density satisfying Assumption 6.1. Then, considering the relative entropy loss ℓ, there exists a positive constant $c(\|v\|_{L^2})$ such that for every control parameter $v \in L^2([0,1], H^s)$:

$$\|\nabla L(v)\|_{L^2} \geq c(\|v\|_{L^2}) L(v)^{1+\lambda}.$$

Proof. As in the proof for the L^2 loss, Theorem 5.1 of continuity of the flow, Property 6.3 of continuity of the gradient and continuity of $\partial\ell$ allow to control the following applications:

$$v \in L^2([0,1], H^s) \mapsto (\rho_t)_t \in \mathcal{C}([0,1], H^s),$$
$$v \in L^2([0,1], H^s) \mapsto (I_t)_t \in \mathcal{C}([0,1], H^s),$$
$$v \in L^2([0,1], H^s) \mapsto (\rho_t \nabla I_t)_t \in \mathcal{C}([0,1], H^s),$$

That is, there exists a constant $C(\|v\|_{L^2})$ such that for every $v \in L^2([0,1], H^s)$:

$$\|\rho_t\|_{H^s}, \|I_t\|_{H^s}, \|\rho_t \nabla I_t\| \leq C(\|v\|_{L^2}).$$

Then by Gagliardo-Nirenberg inequalities:

$$\begin{aligned}\|\rho\nabla I\|_{H^{-s}} &\geq \|\rho\nabla I\|_{L^2}^{\frac{2s-1}{s-1}}\|\rho\nabla I\|_{H^{s-1}}^{\frac{-s}{s-1}}\\ &\geq c(\|v\|_{L^2})\|\rho\nabla I\|_{L^2}^{1+\lambda},\end{aligned}$$

with $\lambda = 1 + 1/(s-1)$.

Then, as ρ_0 is uniformly lower-bounded on $\mathbb{T}^d$, one can integrate its lower bound in the constant c:

$$\|\rho\nabla I\|_{H^{-s}} \geq c(\|v\|_{L^2})\|\nabla I\|_{L^2}^{1+\lambda}.$$

Then writing $I_t = I_1 \circ \varphi_1 \circ \varphi_t^{-1}$, and making a change of variables gives:

$$\|\nabla I_t\|_{L^2} = \|d(\varphi_1 \circ \varphi_t^{-1})^T \nabla I_1 \circ \varphi_1 \circ \varphi_t^{-1}\|_{L^2} \geq c(\|v\|_{L^2})\|\nabla I_1\|_{L^2}.$$

Finally, Assumption 6.1 gives the result by using the initial condition for the backward problem which writes:

$$\begin{aligned}\frac{1}{2}\|\nabla I_1\|_{L^2}^2 &= \frac{1}{2}\int |\nabla \log(\rho_1/\rho^*)|^2\, dx\\ &= 2\int \left|\nabla\sqrt{\rho_1/\rho^*}\right|^2 \rho^* dx\\ &\geq \int \rho_1 \log(\rho_1/\rho^*) dx\\ &= \ell(\rho_1).\end{aligned}$$

□

6.5 Linearized dynamic and diffusion phenomenons along the gradient flow

Taking a step back, we would like to offer here an informal discussion, making a link between our problem and the study of a certain class of diffusion PDEs.

The results presented in this section as well as the one presented in previous sections are not really satisfying as they do not allow to show global convergence of the gradient flow. A phenomenon which is well-observed in practice. This difficulty can be partially explained by the fact that two distinct dynamics are interacting:

- ◇ the model flow dynamic : given by the forward equation, with time $t \in [0, 1]$,
- ◇ the gradient flow dynamic : given by the optimization method, with time $\tau \in [0, +\infty)$.

A first step towards a proof of convergence would thus be to understand the impact of small variations of the control parameter v, given by the evolution along the gradient flow, on the deformed object ρ through the forward equation (recall Section 2).

More precisely, consider our density transport model Equation (26). For flow time t and gradient step time τ we will note the variations of the control $\delta v_t = \partial_\tau v_t^\tau = -\nabla L(v^\tau)_t$ and the induced variation of the density $\delta\rho_t = \partial\rho_t^\tau$. Then differentiating Equation (26) w.r.t. τ:

$$\left\{\begin{array}{rcl}\partial_t\delta\rho + \nabla\cdot(\delta\rho v) + \nabla\cdot(\rho\delta v) &=& 0\\ \delta\rho_0 &=& 0.\end{array}\right. \tag{30}$$

Equation (30) is a transport equation with non-zero source term $f_t := -\nabla \cdot (\rho_t \delta v_t)$. Putting aside regularity assumptions, its solution can be expressed for every time $t \in [0,1]$ as:

$$|D\Phi_t| \delta\rho_t \circ \Phi_t = \int_0^t |D\Phi_s| f_s \circ \varphi_s ds, \tag{31}$$

with $\Phi = \Phi^v$ the flow generated by v.

Equation (31) is rather complicated to analyse. However, at initialization $\tau = 0$ we have $v^0 = 0$, thus $\Phi_t = Id$, $\rho_t = \rho$, $I_t = \partial\ell(\rho)$ and the equation simplifies in:

$$\begin{aligned} \delta\rho_1 &= -\nabla \cdot (\rho \nabla L(v)) \\ &= \nabla \cdot (\rho K * (\rho \nabla \partial\ell(\rho))). \end{aligned} \tag{32}$$

A classical technique to analyse the asymptotic behaviour of such equations is to reformulate them in terms of gradient flows for some energy functional in the Wasserstein space of probability measures [3]. Assuming $\rho^* = 1$ on the flat torus and $K = \delta_0$ (that is $V = L^2$ which is not a RKHS):

◇ For the L^2 loss $\ell(\rho) = \frac{1}{2}\|\rho - 1\|^2$, Equation (32) writes:

$$\partial_t \rho = \nabla \cdot (\rho^2 \nabla \rho) = \frac{1}{3}\Delta(\rho^3),$$

◇ For the relative-entropy loss $\ell(\rho) = \mathbf{H}(\rho||1)$, Equation (32) becomes:

$$\partial_t \rho = \nabla \cdot (\rho \nabla \rho) = \frac{1}{2}\Delta(\rho^2),$$

both of which are known to belong to the class of *Porous Medium Equations.* Convergence towards $\rho^* = 1$ as been analyzed via optimal transport in the pioneering work of [45] and numerical approximation properties of the solutions for non-trivial gaussian kernel $K = G_\sigma$ as been studied in [36] (see also [55]). The convergence proof relies on functional inequalities which can be interpreted in terms of convexity of the energy in the Wasserstein space [46, 26, 12], as well as in terms of Polyak-Lojasiewicz type inequalities [9].

Remark 6.3
Note that recently, several works such as [41, 40, 32] but also [16] or more informally [25] show convergence of the training dynamic of Neural Networks towards the Wasserstein gradient flow of the loss in the mean field limit. However, in contrast with our model of interest, all those works are concerned with shallow networks (with few layers) which they study in the limit of infinite width in order to overcome the lack of convexity.

However, Equation (32) can't fit in this framework as one has to consider a sufficiently smooth kernel K in order to satisfy the admissibility Assumption 5.1. That makes it impossible to interpret the term $\nabla \cdot (\rho K * (\rho \nabla \ell(\rho)))$ as the gradient of an energy functional in the Wasserstein space. Instead, one should directly consider Equation (32) as the gradient of ℓ for the right invariant metric induced by the action of V. We would like to stress out that, in contrast with Wasserstein gradient flows, we lack of theoretical results on the asymptotic behavior of the solutions of such PDEs. Their study could therefore be a first step towards a better understanding of deep ResNets' training dynamic.

7 Scaling to higher dimensions vector field : the problem of RKHS parametrization

In the preceding sections, our model (Definition 5.1) was described by the action of a control parameter v belonging to an abstract RKHS V. In numerical applications, such spaces are described by their kernel function K and with the helps of the representation theorem, allowing to write:

$$v = \sum_{s=1}^{S} K(., z^s)p^s,$$

for some $(z^s)_{1 \leq S \leq S}, (p^s)_{1 \leq s \leq S}$ and some finite S. However, in our case, it is impossible to use such a representation in numerical applications. Indeed, considering the gradient descent $v^{k+1} = -\eta \nabla L(v^k)$ starting at $v^0 = 0$ and using the expression of ∇L in Equation (24) we have:

$$\begin{aligned}
&v^1 = -\nabla L(0) = \sum_{i=1}^{N} K(., x^{i,0})p^{i,0}, \\
&v^2 = v^1 - \nabla L(v^1) = \sum_{i=1}^{N} K(., x^{i,0})p^{i,0} + \sum_{i=1}^{N} K(., x^{i,1})p^{i,1}, \\
&... \\
&v^k = v^{k-1} - \nabla L(v^{k-1}) = \sum_{s=0}^{k-1} \sum_{i=1}^{N} K(., x^{i,s})p^{i,s},
\end{aligned}$$

where for each $k \in \mathbb{N}$, $x^{i,k}$ and $p^{i,k}$ are the forward and backward trajectories of particle i for the control parameter v^k. Therefore, in order to use a representation with the kernel function K, one would need to keep track of an increasing number of trajectories along the gradient descent, which is not numerically feasible.

On the other hand the use of Assumption 5.2 shows that the considered RKHS needs to be of sufficiently high dimension in order to generate sufficiently rich deformations. For that purpose we exhibit here a setting in which RKHS of infinite dimension (such as Sobolev spaces) can be approximated by finite dimensional parametric spaces. We refer to [48] for a more precise presentation.

7.1 General case : defining RKHS through feature maps

We consider in the following a measurable space Ω as well as the functional space $\mathcal{H} := L^2(\Omega, \mu)$ for some positive measure μ on Ω. Then $\mathcal{H}$ is a Hilbert space for the scalar product:

$$\langle f, g \rangle = \int_{\Omega} f \bar{g} d\mu.$$

Linear operators $W : \mathcal{H} \to \mathbb{R}^d$ can therefore be associated with families of vectors $W = (w_i)_{i=1}^{d} \in \mathcal{H}^d$ with for every $f \in \mathcal{H}$:

$$Wf = (\langle w_i, f \rangle)_i,$$

and the associated norm is still an upper-bound for the operator norm:

$$\|W\|_{op} \leq \|W\|_\mu^2 := \sum_i \|w_i\|_\mu^2.$$

Then any application $\varphi : \mathbb{R}^d \to \mathcal{H}$ defines a unique structure of RKHS for which it is the *feature map*. This is the content of the following property:

Property 7.1 (Feature map representation)
Consider an application $\varphi : \mathbb{R}^d \to \mathcal{H}$. Then there exists a unique structure of (vector-valued) RKHS V with kernel K defined as:

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle I_d.$$

V is constituted of those vector-fields v of the form:

$$\forall x \in \mathbb{R}^d, \quad v(x) = W\varphi(x)$$

for some linear operator $W : \mathcal{H} \to \mathbb{R}^d$. φ is called the feature map *associated to V.*

Moreover, the associated scalar product on V is:

$$\forall v, v' \in V, \quad \langle v, v' \rangle_V = \langle W, W' \rangle_\mu = \sum_{i=1}^d \langle w_i, w_i' \rangle.$$

In our model of RKHS-FlowResNet, Property 7.1 allows us to parametrize the RKHS V using only matrices $W \in \mathbb{R}^{d \times q}$. Indeed, considering a finite space $\Omega = \{\omega_1, ...\omega_q\}$ and a positive measure $\mu = \sum_{j=1}^q \mu_j \delta_{\omega_j}$ then linear operators $W : L^2(\Omega, \mu) \to \mathbb{R}^d$ can be associated to $d \times q$ matrices with the scalar product:

$$\begin{aligned}\langle W, W' \rangle_\mu &:= \sum_{i=1}^d \sum_{j=1}^q w_{i,j} w'_{i,j} \mu_j, \\ &= \langle W, W'M \rangle_F\end{aligned}$$

with $\langle ., . \rangle_F$ the canonical Frobenius scalar product and $M = \text{diag}(\mu_1, ..., \mu_q)$.

Therefore, one can still use the canonical scalar product between the control parameters (which for example allows to use automatic differentiation in Section 9), by considering a change of basis in the definition of the vector-fields. The considered space of vector-field is:

$$V := \{v : x \mapsto WN^{-1}\varphi(x) \mid W \in \mathbb{R}^{d \times q}\},$$

with the scalar product:

$$\forall v, v' \in V, \quad \langle v, v' \rangle_V = \langle W, W' \rangle_F = \langle WN^{-1}, W'N^{-1} \rangle_\mu,$$

which defines the same RKHS structure for $N = \sqrt{M}$.

7.2 The case of the Fourier Transform

We saw that choosing for Ω a finite space allows us to recover a parametrization of an abstract RKHS that is compatible with automatic differentiation. We will see here that such a parametrization can be used to approximate infinite dimensional RKHS such as the ones defined by Gaussian kernels or Sobolev kernels.

A canonical choice for the feature space Ω would be the euclidean space $\mathbb{R}^d$ provided with a finite positive measure μ. Then choosing as a feature map the frequency projection:

$$\varphi : x \in \mathbb{R}^d \mapsto (\omega \mapsto e^{i\langle x,\omega\rangle}) \in L^2(\Omega, \mu).$$

The vector field associated to a control parameter W then reads in coordinates:

$$v_i(x) = \int_{\mathbb{R}^d} w_i(\omega) e^{i\langle x,\omega\rangle} d\mu(\omega),$$

where one recognizes that $wd\mu$ is the Fourier Transform of v.

The choice of measure μ then defines the RKHS structure. That can be seen by writing the induced RKHS-norm, assuming μ admits a positive density ρ:

$$\begin{aligned} \|v\|_V^2 &= \int_{\mathbb{R}^d} \|w(\omega)\|^2 \rho(\omega) d\omega \\ &= \int_{\mathbb{R}^d} \|\mathcal{F}v(\omega)\|^2 \frac{d\omega}{\rho(\omega)}, \end{aligned}$$

where we used Parseval's formula in the second inequality.

Thus for $\rho(\omega) = (1 + \|\omega\|^2)^{-s}$ we recover the Sobolev space H^s:

$$\|v\|_{H^s}^2 = \int_{\mathbb{R}^d} \|\mathcal{F}v(\omega)\|^2 (1 + \|\omega\|^2)^s d\omega,$$

and for $\rho(\omega) = e^{-\|\omega\|^2/2}$ we recover the RKHS with gaussian kernel:

$$\begin{aligned} K(x, y) &= \langle \varphi(x), \varphi(y)\rangle I_d \\ &= I_d \int_{\mathbb{R}^d} e^{i\langle x-y,\omega\rangle} e^{-\|\omega\|^2/2} d\omega \\ &\propto I_d e^{-\|x-y\|^2/2}. \end{aligned}$$

More generally for any choice of translation invariant kernel (that is $K(x, y) = k(x-y)I_d$), Bochner's theorem (see [49]) indicates that we can chose $\mu = \mathcal{F}k$, then:

$$K(x, y) = \langle \varphi(x), \varphi(y)\rangle I_d = I_d \int_{\mathbb{R}^d} e^{i\langle x-y,\omega\rangle} d\mu(\omega) = k(x - y)I_d.$$

Finally, as shown above, approximating μ by any measure supported on a finite number of atoms will give a way of approximating the corresponding RKHS by a finite dimensional RKHS. For example, assuming μ is of unit total mass and using the approximation $\tilde{\mu} = \frac{1}{q}\sum_{j=1}^{q} \delta_{\omega_j}$ with independent sampling $\omega_j \sim \mu$ gives the following rescaling:

$$v(x) = W\varphi(x), \quad \|v\|_V^2 = \frac{1}{q}\|W\|_F^2$$

with $\varphi(x) = (e^{i\langle x,\omega_j\rangle})_{1\leq j\leq q}$ the vector of *Random Fourier Features.*

8 Towards a global convergence result

The results of the preceding sections only ensure convergence of the gradient flow when the loss at initialization is small enough. They rely on local PL properties combined with an inductive proof, showing that the gradient dynamic converges as long as it remains bounded. However it does not ensure that there exists no spurious attractive local minima located “at infinity” where the bounding constants m and M can become degenerate.

8.1 A case of divergence for the linear model

We give here a counter-example in the linear case where for a particular setting of the problem, the gradient flow starting from identity initialization does not manage to find a global optimum of the loss. This counter-example is inspired from [7].

We begin by giving a particular result in the case where the objective transformation is a linear transformation:

Property 8.1 (Adapted from [7])
Let R be a normal matrix. We recall the setting of Problem ERM in the linear case with quadratic loss:

$$L(W) = \frac{1}{2}\|U_1 - R\|_F^2,$$

associated with the forward ODE Equation (4) with control parameter $W \in L^2([0,1], \mathbb{R}^{d\times d})$.

Then for the gradient flow initialized at $W^0 = 0$, there exists a unitary matrix A and a family of complex numbers $(\lambda_i)_i$ such that for every time $\tau \geq 0$:

$$W^\tau = A\mathrm{diag}(z_i^\tau)A^*,$$

with $(z_i)_i$ solutions of the equation:

$$\begin{cases} \frac{d}{d\tau} z_i^\tau & = & -(e^{z_i^\tau} - \lambda_i)e^{\bar{z}_i^\tau} \\ z_i^0 & = & 0. \end{cases} \tag{33}$$

In particular W_t^τ is independent of t.

Proof. As R is a normal matrix, there exists A a unitary matrix and a family of complex eigenvalues $(\lambda_i)_i$ such that:

$$R = A\mathrm{diag}(\lambda_i)A^*.$$

Then if $\tau \geq 0$ and W is of the form $W^\tau = A\mathrm{diag}(z_i)A^*$, $\nabla L(W)$ is given by:

$$\nabla L(W) = A\mathrm{diag}\left((e^{z_i} - \lambda_i)e^{\bar{z}_i}\right)A^*.$$

Indeed, solving Equation (4) one finds:

$$U_t = A\mathrm{diag}(e^{tz_i})A^*,$$

and then solving the associated backward ODE Equation (9):

$$\begin{aligned} P_t &= A\mathrm{diag}\left((e^{(1-t)\bar{z}_i}\right)A^*P_1, \\ &= -A\mathrm{diag}\left((e^{(1-t)\bar{z}_i}(e^{z_i} - \lambda_i)\right)A^*, \end{aligned}$$

and finally:

$$\nabla L(W) = -PU^* = A\mathrm{diag}\left(e^{\bar{z}_i}(e^{z_i} - \lambda_i)\right) A^*,$$

which gives the desired equation for z_i considering the gradient flow $\frac{d}{d\tau}W^\tau = -\nabla L(W^\tau)$.

□

A consequence is that it is not possible to obtain the transformation $R = -I_d$ starting from identity. Indeed, setting $\lambda_i = -1$, Equation (33) becomes:

$$\left\{\begin{array}{lcl} \dot{z} & = & -e^{\bar{z}}(e^z + 1) \\ z_0 & = & 0. \end{array}\right.$$

In particular z stays real and setting $x = e^{-z}$:

$$\dot{x} = 1 + 1/x \geq 1,$$

so $x \to +\infty$ and:

$$z = -\log(x) \to -\infty.$$

For the associated deformation we have:

$$U_1 = e^z I_d \to 0.$$

More generally, the dynamic on z corresponds to the gradient flow for the loss $L(z) = \frac{1}{2}|e^z - \lambda|^2$ and a local PL property is verified:

$$|\nabla L(z)|^2 = \left|e^{\bar{z}}(e^z - \lambda)\right|^2 = 2e^{\Re(z)}L(z).$$

The only critical points are $\Re(z) = -\infty$ and $L(z) = 0$ corresponding respectively to $U_1 = 0$ and $U_1 = R$.

More precisely, writing $z = \alpha + i\beta$, $\lambda = re^{i\theta}$ and making the change of variables $\alpha \leftrightarrow e^{-\alpha}$ and $\beta \leftrightarrow \beta - \theta + \pi$, we have the following dynamic:

$$\left\{\begin{array}{lcl} \dot{\alpha} & = & 1/\alpha + r\cos(\beta) \\ \dot{\beta} & = & r\sin\beta/\alpha. \end{array}\right.$$

The only critical points are $\beta = 0, \alpha = 1/r$, corresponding to $L(z) = 0$, and $\beta = \pi, \alpha = +\infty$, corresponding to $L(z) > 0$. The gradient flow converges towards this second critical point for $\beta^0 = 0$, however it converges towards the first one as soon as $\beta^0 = \epsilon$ with $\epsilon > 0$.

8.2 Universality of global convergence

In the preceding counter-example, non-convergence is achieved by taking advantage of the symmetries of the problem. One could for example exhibit similar counter-examples with the RKHS model of Definition 5.1 and symmetric kernel K.

However, we also showed that this divergence behavior is very sensitive to the problem setting. It is in particular not stable in the sense that a small perturbation of the objective (or equivalently of the data input) leads to a significantly different behavior. This evidence supports the fact that global convergence of ResNet is a generic phenomenon: even though one can exhibit counter-examples, one should expect the model to converge for almost every input and almost every initialization.

We summarize this idea in the following conjecture:

Conjecture 1 (Generic convergence)
Consider the framework of the Empirical Risk Minimization problem for the RKHS-FlowResNet model of Definition 5.1. Then there exists a dense subset $\mathcal{D} \subset (\mathbb{R}^d)^N \times (\mathbb{R}^d)^N$ such that for every pair of input-objective point cloud $((x^i)_i, (y^i)_i) \in \mathcal{D}$ the gradient flow with identity initialization $v^0 = 0$ converges towards a 0-training loss global optimum, i.e.:

$$v^\tau \xrightarrow[\tau \to +\infty]{} v^*, \quad \textit{with} \quad L(v^*) = 0.$$

9 Numerical experiments

We show here some numerical examples in order to illustrate the results that were presented in this report. In order to realise those experiments, the models of Definitions 3.1, 4.1 and 5.1 were implemented in `Pytorch` and optimization of the parameters was performed thanks to automatic differentiation.

All experiments were made using the input dimension $d = 2$ in order to plot the trajectories of the points along the forward problems. We considered optimization with Gradient Descent or with Stochastic Gradient Descent.

Remark 9.1 ("Batch" loss and "global" loss)
In the case of optimization with SGD and in contrast with GD, the computed losses are averaged on the number of data points considered. This is done in order to compare the "batch" loss, computed at each iteration so as to perform back-propagation, and the "global" loss, computed regularly only in order to estimate the efficiency of the training process.

9.1 Linear models

In the experiment of Figure 1, we aim at observing the way the convergence condition of Equation (6) can be enforced by increasing the dimension of the matrices A and B in the linear model of Definition 3.1. We therefore set the inputs to be randomly distributed, $X = (x^i_j)^{1 \leq i \leq N}_{1 \leq j \leq d}$ with $x^i_j \sim \mathcal{N}(0,1)$ and the objective outputs to be a linear transformation of the inputs, $Y = R_\theta X$, with R_θ being the rotation matrix of angle θ, and $\theta = \pi - \epsilon$ with $\epsilon \ll 1$. Note that we are therefore in the setting of Section 8, and that such a problem is "hard" for the linear model without considering an embedding in a space of higher dimension $q > d$.

Two things can be observed in Figure 1:

- ◇ The problem is significantly easier when raising the embedding dimension up to $q = 2d = 4$ (right). Indeed, starting from initialization, the training loss immediately decreases at a linear rate and only few iterations of gradient descent are needed in order to achieve convergence towards a zero training loss optimum. This can be explained by the fact that raising the dimension allows trajectories that are not allowed otherwise. Thereby generated flows (which are projected in the original input space) look like straight lines whereas they look much more like rotations for $q = d = 2$.

- ◇ In the case where $q = d = 2$ (left), starting from initialization, the training loss only poorly decreases but still, at the end, reaches a global minimum. Indeed we observed that the point clouds are matched by the flow of the corresponding forward ODE. This kind of behavior cannot be explained by our local convergence result. In fact it is, to the best of our knowledge, still an open problem to understand this kind of behavior in the training dynamic of deep ResNets.

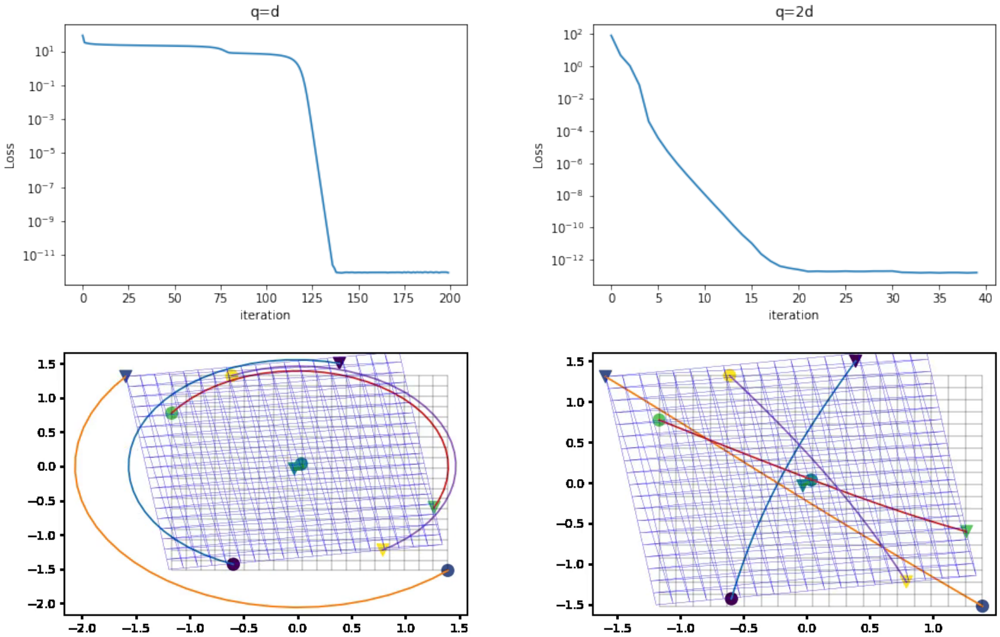

Figure 1: Evolution of the loss during training with Gradient Descent (top) and generated flows (bottom) for linear models: circles represent the input data, triangles of the same color the associated objectives, and plain lines are the trajectories of the forward ODE at the end of training. The dark grid is deformed onto the blue grid.
Objective data: $Y = R_\theta X$, with $\theta = \pi - \epsilon$, $N = 20$ data points with embedding dimension $q = d = 2$ (left) or $q = 2d = 4$ (right).

Note that we obtained the same kind of plots considering an optimization of the parameters with Stochastic Gradient Descent in Figure 2.

9.2 RKHS models

In the experiments of Figure 3, we aim at observing the advantages brought by a non-linear model such as the one of Definition 4.1 in comparison with the linear model of Definition 3.1. Motivated by the approximation properties presented in Section 7, we consider for a non-linear feature map φ the map of *Random Fourier Features* with frequencies sampled from the probability density $\rho(\omega) \propto (3 + \|\omega\|^2)^{-5/2}$. Therefore, the considered space of residual terms is an approximation of the Sobolev space $V = H^{5/2}(\mathbb{R}^2, \mathbb{R}^2)$.

Remark 9.2 (Sampling from a multivariate t-distribution)
In dimension $d = 2$, the probability distribution with density $\rho(\omega) \propto (3 + \|\omega\|^2)^{-5/2}$ is the multivariate t-distribution with 3 degrees of freedom, noted $\mathbf{t}_3(0, I_2)$. Considering the random variable $z = y/\sqrt{u/3}$ with $y \sim \mathcal{N}(0, I_2)$ and $u \sim \chi^2_3$, the multivariate chi-squared distribution with 3 degrees of freedom, then $z - \mathbb{E}[z] \sim \mathbf{t}_3(0, I_2)$.

As shown in Section 8, we observe in Figure 3 that the gradient descent does not converge in the training of the Linear-FlowResNet model. Because of the symmetries of the problem, the parameters diverge to infinity and asymptotically the model implements a degenerate deformation where the whole plane shrinks at the origin.

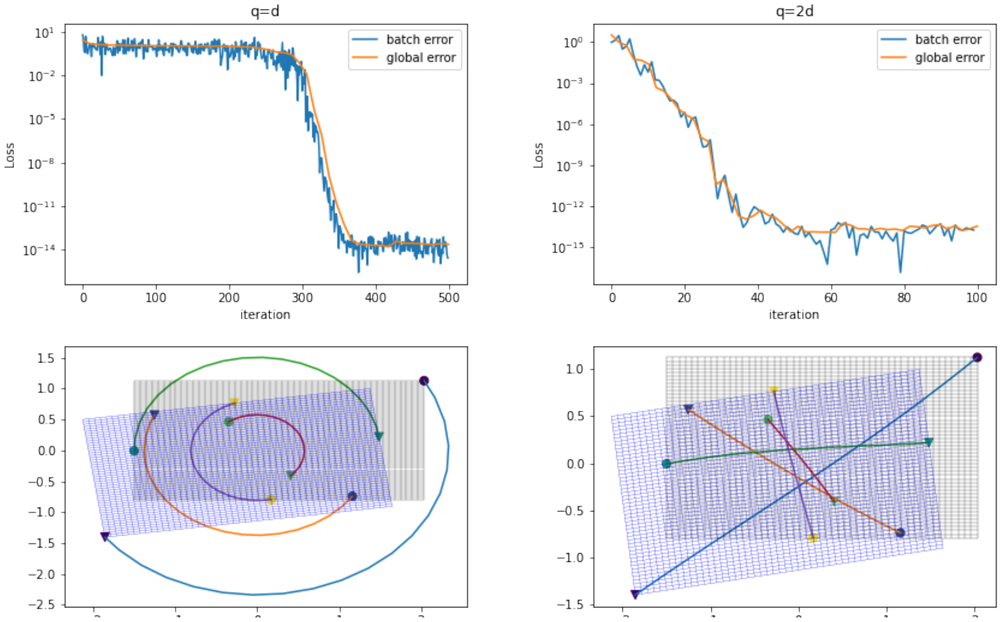

Figure 2: Evolution of the loss during training with SGD (top) and generated flows (bottom) for linear models: circles represent the input data, triangles of the same color the associated objectives, and plain lines are the trajectories of the forward ODE at the end of training. The dark grid is deformed onto the blue grid.
Objective data: $Y = R_\theta X$, with $\theta = \pi - \epsilon$, $N = 200$ data points with embedding dimension $q = d = 2$ (left) or $q = 2d = 4$ (right).

On the other hand, because the RKHS-FlowResNet model parametrized on the Sobolev space $V = H^{5/2}$ is much more expressive, we see that gradient descent achieves a better training loss while training this model. Furthermore, the decrease rate of the training loss increases with the number of features. It is normal as the model is as expressive as the set of residual terms is “rich”, that is as the number of features is high. However, even though the model interpolates the training dataset, the implemented deformation of the space is not smooth, which could be a desirable property, and does not seem to get smoother as the number of features increases.

10 Conclusion

We presented several Machine Learning models for which we studied the convergence properties under gradient flow or gradient descent in a supervised learning setting. All of the models are continuous and represent a rescaled limit of discrete ResNet architectures. We studied a model of linear deformation in Section 3 and general non-linear models in Sections 4 and 5, motivating the use of an RKHS parametrization. In each case, relying on functional inequalities of Polyak-Lojasiewicz type, we showed that, under some assumptions, the loss landscape admits no spurious local minima and a sufficiently small loss at initialization allows to recover a convergence towards a global optimum at a linear rate. Thereby we exhibited a numerical threshold in the transition towards a “linear regime”, which can be reached for example by embedding the data in sufficiently high dimension.

A second aspect of our work is the connection we made between the problem of con-

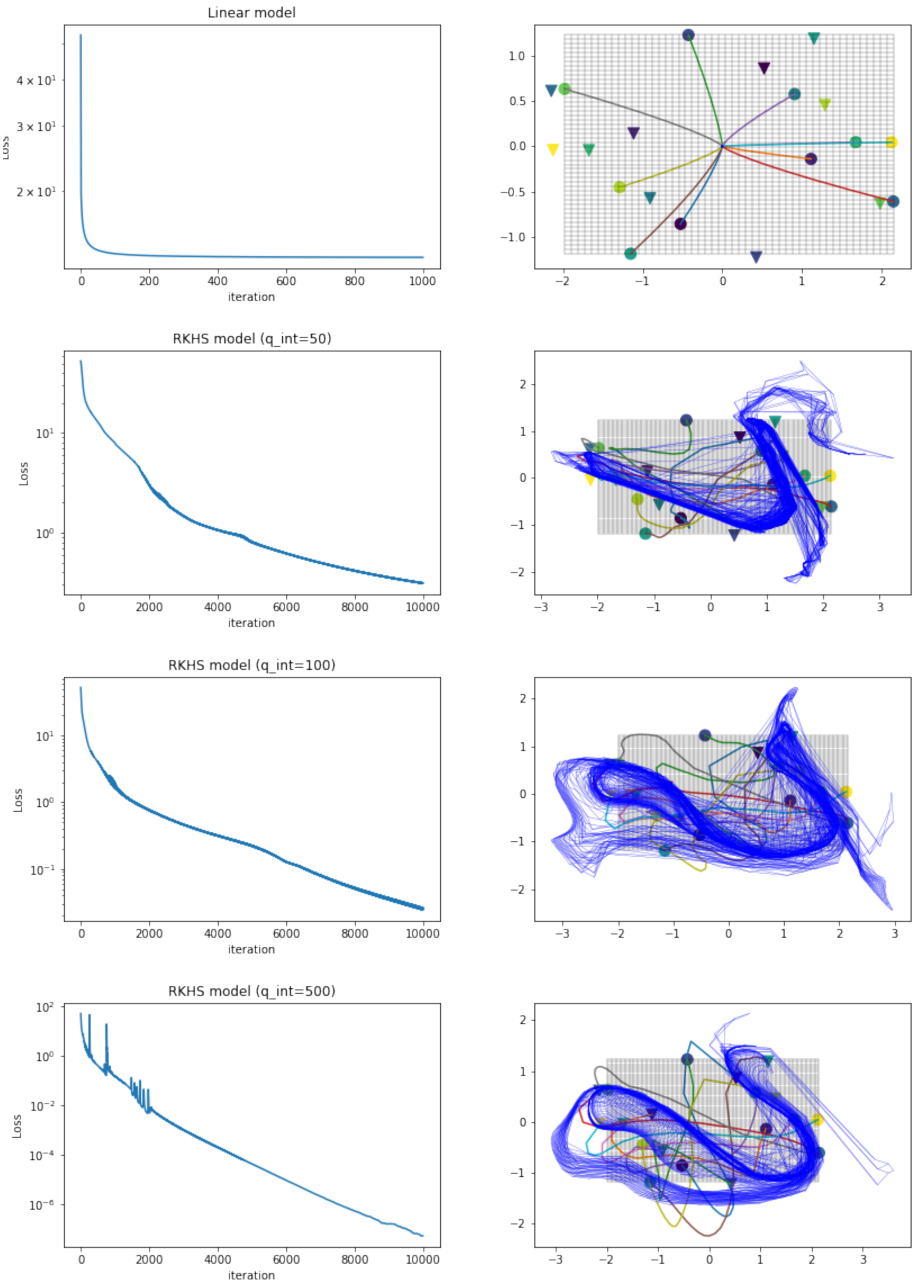

Figure 3: Evolution of the loss during training with Gradient Descent (left) and generated flows (right) for the linear model (top) and for RKHS models with increasing dimension `q_int` of the feature space.
The dark grid are deformed onto the blue grids by the model at the end of training.
Objective data: $Y = -X$, $N = 10$ data points.

vergence of these residual models and tools from mathematical image registration by using an RKHS parametrization of the vector-fields. This allowed us to consider in Section 6 Normalizing Flow models implementing diffeomorphic deformations of the data, with applications in generative modeling. We proved that there are no spurious local minima in the loss landscape of our model, but, due to the infinite dimension of the density matching problem, the functional inequalities of Kurdyka-Lojasiewicz type we obtained are too weak to conclude to a convergence result.

Open questions and further work Those results are by many aspects unsatisfying as they do not explain in full generality how first order methods such as gradient descent manage to optimize complex overparametrized models so as to reach a global minimum of the training loss. A phenomenon which is yet well observed in practice. We point out two lines of work which we believe are relevant in order to fulfill this program:

◇ Leveraging implicit bias : Our results only express a “local” convergence behavior and rely on conditions at initialization. This does not explain the general behavior of training dynamics of overparametrized models which, for many instances, include long phases of slow improvement with important variations of the parameters before entering in a phase of linear convergence (see Section 9) in the neighborhood of the optimum. One could explain this phenomenon by studying in depth the *Implicit Biases* induced by gradient descent in the optimization of overparametrized models [43, 28, 27, 51, 42]. The general idea is that the functional inequalities in Properties 3.1 and 5.2 describe global properties of the loss landscape, whereas one should try to derive stronger properties that are only verified along the gradient descent. Ultimately, an objective would be to prove Conjecture 1.

◇ Parametrize residual layers in the Barron space : RKHS parametrization of the displacement vector field was firstly motivated by generalizing the non-linear model of Definition 4.1. However, such a model is not really representative of the architectures that are used in practice. In particular, the considered residual layer does not have the property to be a universal approximator in contrast for example with Multilayer Perceptrons (MLP) with one hidden layer and a ReLU or Sigmoid activation function, as introduced in the seminal work of Barron [6]. Therefore, a natural extension of our work would be to consider models parametrized by deformation vector fields in a “Barron space”. In practice it amounts to training a supplementary linear hidden layer for each residual layer, thus further complicating the analysis as the residual terms will then no longer be linear w.r.t. the control parameter. Such a parametrization has already been considered but in the mean field limit of infinite width [16, 22] and [5] gives a comparison of the two (RKHS vs. Barron space) approaches.

References

[1] Z. Allen-Zhu, Y. Li, and Z. Song, *A convergence theory for deep learning via over-parameterization*, in International Conference on Machine Learning, PMLR, 2019, pp. 242–252.

[2] L. Ambrosio, *Transport equation and cauchy problem for non-smooth vector fields*, in Calculus of variations and nonlinear partial differential equations, Springer, 2008, pp. 1–41.

[3] L. Ambrosio, N. Gigli, and G. Savare, *Gradient flows: in metric spaces and in the space of probability measures*, Lectures in mathematics ETH Zürich, (2008).

[4] C. Ané, S. Blachère, D. Chafaï, P. Fougères, I. Gentil, F. Malrieu, C. Roberto, and G. Scheffer, *Sur les inégalités de Sobolev logarithmiques*, no. 10 in Panoramas et synthèses, Société Mathématique de France ; Diffusion, AMS, Paris : Providence, RI, 2000.

[5] F. Bach, *Breaking the curse of dimensionality with convex neural networks*, The Journal of Machine Learning Research, 18 (2017), pp. 629–681.

[6] A. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Transactions on Information Theory, 39 (1993), pp. 930–945.

[7] P. Bartlett, D. Helmbold, and P. Long, *Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks*, in International Conference on Machine Learning, PMLR, 2018, pp. 521–530.

[8] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, *Computing large deformation metric mappings via geodesic flows of diffeomorphisms*, International journal of computer vision, 61 (2005), pp. 139–157.

[9] A. Blanchet and J. Bolte, *A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions*, Journal of Functional Analysis, 275 (2018), pp. 1650–1673.

[10] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, *Characterizations of Łojasiewicz inequalities: Subgradient flows, talweg, convexity*, Transactions of the American Mathematical Society, 362 (2009), pp. 3319–3363.

[11] M. Bruveris and F.-X. Vialard, *On completeness of groups of diffeomorphisms*, Journal of the European Mathematical Society, 19 (2017), pp. 1507–1544.

[12] J. A. Carrillo, R. J. McCann, and C. Villani, *Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates*, Revista Matematica Iberoamericana, 19 (2003), pp. 971–1018.

[13] B. Chang, L. Meng, E. Haber, L. Ruthotto, D. Begert, and E. Holtham, *Reversible architectures for arbitrarily deep residual neural networks*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[14] B. Chang, L. Meng, E. Haber, F. Tung, and D. Begert, *Multi-level residual networks from dynamical systems view*, in International Conference on Learning Representations, 2018.

[15] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, *Neural ordinary differential equations*, Advances in Neural Information Processing Systems, (2018).

[16] L. Chizat and F. Bach, *On the global convergence of gradient descent for overparameterized models using optimal transport*, Advances in Neural Information Processing Systems, 31 (2018), pp. 3036–3046.

[17] L. Chizat, E. Oyallon, and F. Bach, *On lazy training in differentiable programming*, in NeurIPS 2019-33rd Conference on Neural Information Processing Systems, 2019, pp. 2937–2947.

[18] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, *Gradient descent finds global minima of deep neural networks*, in International Conference on Machine Learning, PMLR, 2019, pp. 1675–1685.

[19] S. S. Du, X. Zhai, B. Poczos, and A. Singh, *Gradient descent provably optimizes over-parameterized neural networks*, in International Conference on Learning Representations, 2018.

[20] E. Dupont, A. Doucet, and Y. W. Teh, *Augmented neural odes*, in Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 3140–3150.

[21] W. E, J. Han, and Q. Li, *A mean-field optimal control formulation of deep learning*, Research in the Mathematical Sciences, 6 (2019), p. 10.

[22] W. E, C. Ma, and L. Wu, *The Barron Space and the Flow-Induced Function Spaces for Neural Network Models*, Constructive Approximation, (2021).

[23] I. Ekeland, *The hopf-rinow theorem in infinite dimension*, Journal of Differential Geometry, 13 (1978), pp. 287–301.

[24] W. Fedus, B. Zoph, and N. Shazeer, *Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity*, arXiv preprint arXiv:2101.03961, (2021).

[25] X. FERNÁNDEZ-REAL and A. FIGALLI, *The continuous formulation of shallow neural networks as wasserstein-type gradient flows.*

[26] W. Gangbo and R. J. McCann, *The geometry of optimal transportation*, Acta Mathematica, 177 (1996), pp. 113–161.

[27] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, *Characterizing implicit bias in terms of optimization geometry*, in International Conference on Machine Learning, PMLR, 2018, pp. 1832–1841.

[28] ——, *Implicit bias of gradient descent on linear convolutional networks*, In Advances in Neural Information Processing Systems, (2018).

[29] J. K. Hale, *Ordinary differential equations*, Dover Publications, Mineola, N.Y, dover ed ed., 2009. OCLC: ocn294885198.

[30] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 2016, IEEE, pp. 770–778.

[31] A. Jacot, F. Gabriel, and C. Hongler, *Neural tangent kernel: convergence and generalization in neural networks (invited paper)*, in Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Italy, June 2021, ACM, pp. 6–6.

[32] A. Javanmard, M. Mondelli, and A. Montanari, *Analysis of a two-layer neural network via displacement convexity*, The Annals of Statistics, 48 (2020), pp. 3619–3642.

[33] I. Kobyzev, S. Prince, and M. Brubaker, *Normalizing Flows: An Introduction and Review of Current Methods*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2020), pp. 1–1.

[34] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, in Annales de l'institut Fourier, vol. 48, 1998, pp. 769–783.

[35] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, *Wide neural networks of any depth evolve as linear models under gradient descent*, Advances in neural information processing systems, 32 (2019), pp. 8572–8583.

[36] P.-L. Lions and S. Mas-Gallic, *Une méthode particulaire déterministe pour des équations diffusives non linéaires*, Comptes Rendus de l'Académie des Sciences-Series I-Mathematics, 332 (2001), pp. 369–376.

[37] C. Liu, L. Zhu, and M. Belkin, *On the linearity of large non-linear models: when and why the tangent kernel is constant*, Advances in Neural Information Processing Systems, 33 (2020).

[38] ——, *Loss landscapes and optimization in over-parameterized non-linear systems and neural networks*, arXiv:2003.00307 [cs, math, stat], (2021). arXiv: 2003.00307.

[39] S. Lojasiewicz, *Sur les trajectoires du gradient d'une fonction analytique*, Seminari di geometria, 1983 (1982), pp. 115–117.

[40] S. Mei, T. Misiakiewicz, and A. Montanari, *Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit*, in Conference on Learning Theory, PMLR, 2019, pp. 2388–2464.

[41] S. Mei, A. Montanari, and P.-M. Nguyen, *A mean field view of the landscape of two-layer neural networks*, Proceedings of the National Academy of Sciences, 115 (2018), pp. E7665–E7671.

[42] E. Moroshko, B. E. Woodworth, S. Gunasekar, J. D. Lee, N. Srebro, and D. Soudry, *Implicit bias in deep linear classification: Initialization scale vs training accuracy*, Advances in Neural Information Processing Systems, 33 (2020).

[43] B. Neyshabur, *Implicit regularization in deep learning*, arXiv preprint arXiv:1709.01953, (2017).

[44] Q. Nguyen, *On the Proof of Global Convergence of Gradient Descent for Deep ReLU Networks with Linear Widths*, arXiv:2101.09612 [cs, stat], (2021). arXiv: 2101.09612.

[45] F. OTTO, *The geometry of dissipative evolution equations: the porous medium equation*, Comm. Partial Differential Equations, 26 (2001), pp. 101–174.

[46] F. Otto and C. Villani, *Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality*, Journal of Functional Analysis, 173 (2000), pp. 361–400.

[47] H. Owhadi, *Do ideas have shape? Plato's theory of forms as the continuous limit of artificial neural networks*, arXiv:2008.03920 [cs, stat], (2020). arXiv: 2008.03920.

[48] A. Rahimi and B. Recht, *Random features for large-scale kernel machines*, in Proceedings of the 20th International Conference on Neural Information Processing Systems, 2007, pp. 1177–1184.

[49] W. Rudin, *Fourier analysis on groups*, Courier Dover Publications, 2017.

[50] H. Salman, P. Yadollahpour, T. Fletcher, and K. Batmanghelich, *Deep diffeomorphic normalizing flows*, arXiv e-prints, (2018), pp. arXiv–1810.

[51] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, *The implicit bias of gradient descent on separable data*, The Journal of Machine Learning Research, 19 (2018), pp. 2822–2878.

[52] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[53] A. Trouvé, *Diffeomorphisms groups and pattern matching in image analysis*, International journal of computer vision, 28 (1998), pp. 213–221.

[54] A. Trouvé and L. Younes, *Local geometry of deformable templates*, SIAM journal on mathematical analysis, 37 (2005), pp. 17–59.

[55] J. L. Vazquez, *The Porous Medium Equation: Mathematical Theory*, Oxford University Press on Demand, 2007.

[56] F.-X. Vialard, R. Kwitt, S. Wei, and M. Niethammer, *A shooting formulation of deep learning*, Advances in Neural Information Processing Systems, 33 (2020).

[57] E. Weinan, *A proposal on machine learning via dynamical systems*, Communications in Mathematics and Statistics, 5 (2017), pp. 1–11.

[58] L. Wu, Q. Wang, and C. Ma, *Global convergence of gradient descent for deep linear residual networks*, Advances in Neural Information Processing Systems, 32 (2019), pp. 13389–13398.

[59] L. Younes, *Shapes and Diffeomorphisms*, vol. 171 of Applied Mathematical Sciences, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[60] S. Zagoruyko and N. Komodakis, *Wide residual networks*, in British Machine Vision Conference 2016, British Machine Vision Association, 2016.

[61] H. Zhang, X. Gao, J. Unterman, and T. Arodz, *Approximation capabilities of neural odes and invertible residual networks*, in International Conference on Machine Learning, PMLR, 2020, pp. 11086–11095.

[62] D. Zou, Y. Cao, D. Zhou, and Q. Gu, *Gradient descent optimizes over-parameterized deep ReLU networks*, Machine Learning, 109 (2020), pp. 467–492.

[63] D. Zou, P. M. Long, and Q. Gu, *On the global convergence of training deep linear resnets*, in International Conference on Learning Representations, 2019.

Appendices

A Caratheodory existence and uniqueness theory for ODEs

We are interested in the study of solution of the initial value problem:

$$\left\{ \begin{array}{rcl} x(0) & = & x_0 \\ \dot{x}(t) & = & F(\theta(t), x(t)), \end{array} \right. \tag{34}$$

where $x_0 \in \mathbb{R}^d$, $F : \mathbb{R}^q \times \mathbb{R}^d \to \mathbb{R}^d$ is a continuous function and θ is a *control parameter* living in some control space of function $\mathcal{B} \subset \mathcal{F}([0,1], \mathbb{R}^q)$. Cauchy-Lipschitz theorem gives existence and uniqueness of a solution for problem Equation (34) but makes strong regularity assumption on control parameter θ such as local lipschitzianity whereas we would like $\mathcal{B}$ to be a space of less regular functions such as $L^r([0,1], \mathbb{R}^q)$ for $1 \leq r < +\infty$.

In this context ones needs to consider solutions in a weaker sense. Indeed, observe that for a continuous control parameter θ, Equation (34) is equivalent to:

$$x(t) = x_0 + \int_0^t F(\theta(s), x(s)) ds, \tag{35}$$

where only integrability of θ is needed if one only needs equality to be true in the sense of absolutely continuous functions. Existence and uniqueness of a solution is given by the following theorem (see [29]):

Theorem A.1 (Caratheodory)
Assume that $f : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ is such that f is measurable in t for each fixed x and continuous in x for each fixed t. Further assume that for each compact set $U \subset \mathbb{R}^d$ there exist a measurable function k_U such that:

$$\|f(t,x) - f(t,y)\| \leq k_U(t)\|x - y\|, \ \forall t \in [0,1], \ \forall x, y \in U.$$

Then, for every $x_0 \in \mathbb{R}^d$, there exist an unique absolutely continuous curve x satisfying:

$$x(t) = x_0 + \int_0^t f(s, x(s)) ds, \ \forall t \in [0,1].$$

The following corollary applies to our case:

Corollary A.1
Assume that for each compact set $U \subset \mathbb{R}^d$ there exist a continuous function k_U such that:

$$\|F(\theta,x) - F(\theta,y)\| \leq k_U(\|\theta\|)\|x - y\|, \ \forall \theta \in \mathbb{R}^q, \ \forall x, y \in U.$$

Then, for every integrable control parameter θ there exist an unique absolutely continuous solution of equation Equation (35).

The assumption is for example satisfied for a single layer perceptron with a non-linear activation function, i.e. when $F(\theta, x) = \theta\varphi(x)$ with $\theta \in \mathbb{R}^{d \times q}$ and $\varphi : \mathbb{R}^d \to \mathbb{R}^q$ is continuous.