Global convergence of ResNets: From finite to infinite width using linear parameterization

Raphaël Barboni, Gabriel Peyré, François-Xavier Vialard

Global convergence of ResNets: From finite to infinite width using linear parameterization

Raphaël Barboni * Gabriel Peyré ** François-Xavier Vialard °

* ENS - PSL Univ.
** CNRS, ENS - PSL Univ.
° LIGM, Université Gustave Eiffel, CNRS

Abstract

Overparameterization is a key factor in the absence of convexity to explain global convergence of gradient descent (GD) for neural networks. Beside the well studied lazy regime, infinite width (mean field) analysis has been developed for shallow networks, using on convex optimization technics. To bridge the gap between the lazy and mean field regimes, we study Residual Networks (ResNets) in which the residual block has linear parameterization while still being nonlinear. Such ResNets admit both infinite depth and width limits, encoding residual blocks in a Reproducing Kernel Hilbert Space (RKHS). In this limit, we prove a local Polyak-Lojasiewicz inequality. Thus, every critical point is a global minimizer and a local convergence result of GD holds, retrieving the lazy regime. In contrast with other mean-field studies, it applies to both parametric and non-parametric cases under an expressivity condition on the residuals. Our analysis leads to a practical and quantified recipe: starting from a universal RKHS, Random Fourier Features are applied to obtain a finite dimensional parameterization satisfying with high-probability our expressivity condition.

1 Introduction

State of the art supervised learning methods are based on deep neural networks, sometimes heavily overeparameterized, which perfectly fit training data or even noisy data while exhibiting good generalization properties. Such a behaviour appears as a paradox and questions the established theory of "bias-variance trade-off" [10]. That an overparameterized model can fit data perfectly comes as no surprise but this capability does not explain the observed generalization properties. Towards a better understanding of it, one first needs to understand the optimization procedure in the parameter space that selects the interpolation map. This question is tightly linked with the parameterization of the space of maps that are explored and state of the art parameterizations have emerged in the past years. One key architecture that is ubiquitous in deep learning are skip connections, heavily used in *Residual Neural Networks* (ResNets) [25] and it has led to state of the art results in supervised learning. ResNets actually allow to consider a very large number of layers [58].

Continuous models Passing to the limit of infinite depth allows the connection with continuous models (Neural ODE) for which theoretical methods and new algorithms can be designed [12, 55]. Indeed, the similarities between ResNet architectures and discrete numerical schemes motivated the introduction of a continuous neural ODE

$$\dot{z}_t = v(W_t, z_t) \quad \forall t \in [0,1]\,, \tag{1}$$

where $W \in L^2([0,1]\,, \mathbb{R}^m)$ is the parameter of the model and $v : \mathbb{R}^m \times \mathbb{R}^q \to \mathbb{R}^q$ is a *residual transformation* whose output is the *residual term*. These models correspond to limiting models of a discrete ResNet whose depth L tends to infinity [53]. Therefore, their

study brings a theoretical framework for understanding deep ResNet architectures, and more generally very deep NNs [19, 20]. Moreover, their mathematical analysis is facilitated since it allows to leverage a large body of works and tools from analysis and in particular the theory of optimal control [45]. Conversely, methods from numerical analysis can bring inspiration for designing new architectures and new optimization algorithms [38].

RKHS parameterization Most often in the literature studying the training properties of ResNets, the considered residual transformations are *Multi-Layer Perceptrons (MLP)* [17, 2, 24]. Those consist in the composition of several trained linear layers alternatively composed with a non-linear activation function. A 2-layer MLP with width r reads:

$$v : ((W, U), z) \mapsto W\sigma(Uz), \tag{2}$$

where $U \in \mathbb{R}^{r\times q}$ and $W \in \mathbb{R}^{q\times r}$ are the parameters for the "hidden" and the "visible" layer respectively and $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear *activation function* applied component-wise. Popular activation functions are for example the ReLU or the Swish function. Provided with these activations, MLPs enjoy a nice universal approximation property as shown in the seminal work of Barron [7].

In contrast, we consider here a simplified setting where the residual term is linear w.r.t. the parameters while still being nonlinear w.r.t the inputs. Given a feature map $\varphi : \mathbb{R}^q \to \mathbb{R}^r$, we consider as space of residuals the vector space:

$$V \coloneqq \{v : z \mapsto W\varphi(z) | W \in \mathbb{R}^{q\times r}\}, \tag{3}$$

where the matrices $W \in \mathbb{R}^{q\times r}$ are the trained parameters. Compared to Eq. (2), this can be seen as an MLP where the hidden layer is fixed by introducing the feature map $\varphi : z \mapsto \sigma(U^\top z)$ for some feature matrix U. As is standard, the gradient of some loss L w.r.t. W is computed in the sense of the Frobenius metric on the set of matrices:

$$\forall\, W, W' \in \mathbb{R}^{q\times r},\ \langle W, W'\rangle = \mathrm{Tr}(W^\top W'). \tag{4}$$

Such an L^2 penalization induces a metric structure on the set V through the identification $v \leftrightarrow W$ in Eq. (3):

$$\forall\, v, v' \in V,\ \langle v, v'\rangle_V \coloneqq \langle W, W'\rangle. \tag{5}$$

As a finite dimensional space of continuous maps, V has the structure of *Reproducing Kernel Hilbert Space* (RKHS). Moreover, as pointed out in [5], the space V has a natural infinite width limit or mean field limit which is an infinite dimensional RKHS.

In this paper, we are interested in understanding the convergence properties of Gradient Descent (GD) on a ResNet model for which the residual layers are encoded in a – possibly infinite-dimensional – vector-valued RKHS V. For V as in Eq. (3), we stress out that, as the metric on V is induced by the one on $\mathbb{R}^{q\times r}$, GD on V for this metric is strictly equivalent to GD on $\mathbb{R}^{q\times r}$ with the Frobenius metric. Our model is defined as follows:

Definition 1 (RKHS-FlowResNet)**.** *Let V be a RKHS of vector-fields over $\mathbb{R}^q$ and $A \in \mathbb{R}^{q\times d}$, $B \in \mathbb{R}^{d'\times q}$ be matrices. Then for $v \in L^2([0,1], V)$ and a data input $x \in \mathbb{R}^d$, the RKHS-FlowResNet's output is defined as:*

$$F(v, x) \coloneqq Bz_1,$$

where z is the solution to the forward problem

$$\dot{z}_t = v_t(z_t) \quad \textit{and} \quad z_0 = Ax. \tag{6}$$

The variable v will thereafter be called control parameter.

Remark 1. *Note that the matrices A and B are fixed and only the control parameter v is trained. However, we argue that our approach can be simply adapted to the case where B is trained, following for example the proof of [42]. Training A seems more challenging as the model is highly non-linear w.r.t. this parameter.*

Considering such a simplified model comes with shortcomings as well as potential benefits. The main assumption that differs from standard ResNets is linearity in the parameters of the residual blocks. As a comparison, a 2-layer MLP is nonlinear w.r.t. its parameters of the hidden layer. While it is admittedly a simplified setting, the model of Definition 1

still retains the effect of depth and the nonlinearity w.r.t. the input. Indeed, considering V to be a Random Features approximation (c.f. Eq. (29)) of some universal RKHS, the residual blocks are as expressive as a 2-layer MLP as both are dense in the space of continuous functions. Moreover, due to composition of these residual blocks the model's output is still highly non-linear w.r.t. parameters. Therefore, we consider this model as a first step of study towards the general case. In turn, this linearity in parameters naturally leads to an RKHS parameterization which has two important benefits on the theoretical side: **(i)** Flows of vector-fields as implemented by our model in Eq. (6) have already been studied theoretically and for applications in image registration problems [57, 9, 43]. Under some regularity assumptions on the considered RKHS V, one can show that the model's output corresponds to the invertible action of a diffeomorphism by composition on the input [54]. This property was already used in [49] to implement models of *Normalizing Flows* [29] with applications in generative modeling. **(ii)** There is an important literature in Machine Learning about Kernel methods [50]. In practice, various sub-sampling methods exist in order to approximate infinite-dimensional RKHSs with finite-dimensional spaces generated by *Random Fourier Features* (RFF) [46, 47]. Thereby, leveraging results on the approximation bound for RFF [52, 51], we show that the expressiveness properties of universal kernels, such as the Gaussian kernel, can be efficiently recovered using residuals of the form Eq. (3) with a finite number of neurons.

Supervised learning. We consider a map $F : \mathcal{H} \times \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ for some Hilbert space of parameter $\mathcal{H}$ (e.g. the model of Definition 1 with $\mathcal{H} = L^2([0,1], V)$) and a training dataset consisting on a family of inputs $(x^i)_{1\leq i\leq N} \in (\mathbb{R}^d)^N$ and target outputs $(y^i)_{1\leq i\leq N} \in (\mathbb{R}^{d'})^N$. Then for every parameter $v \in \mathcal{H}$, we define the associated *Empirical Risk* as:

$$L(v) \coloneqq \frac{1}{2N} \sum_{1\leq i\leq N} \|F(v, x^i) - y^i\|^2. \tag{7}$$

Remark 2. *For simplicity we consider here the euclidean square distance as a loss on the output space $\mathbb{R}^{d'}$, but our results generalize to any smooth loss satisfying a Polyak-Lojasiewicz inequality (c.f.[11]), e.g. any smooth strongly convex loss.*

Training the model F then amounts to finding a parameter $v^* \in \arg\min_{v\in\mathcal{H}} L(v)$. In order to perform such an *empirical risk minimization (ERM)* task we consider GD on v. For a small step size η, for some initialization $v^0 \in \mathcal{H}$ and for every discrete time step $k \in \mathbb{N}$, the training dynamic reads:

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

Note that we do not consider any additional regularizing term on the loss. In a classical supervised learning one would seek for a parameter minimizing the "regularized" loss $L(v) + \lambda\mathcal{R}(v)$, with $\lambda > 0$ a constant and $\mathcal{R}$ a coercive regularization function. However, we are here interested in the non regularized setting, i.e. $\lambda = 0$ often used in practice. In this case, the generalization property of the computed map is argued to potentially come from the optimization method that adequately shall select a good minimizer of the loss. This implicit regularization depends on the choice of the optimization method [41].

2 Related works and contributions

Recently, several works have addressed the problem of proving convergence of (stochastic) GD in the training of NNs. In [33, 32, 18], the authors focus on the training of "shallow" two layers fully connected NNs and establish convergence of GD in an over-parameterized setting where width of the intermediary layer scales polynomially with the size N of the dataset. More recently, with the same problem setup, [60] shows that the neurons of a teacher network are recovered by a student network optimized with GD as long as the width of the student network is higher than the teacher's one. Formally, their analysis is similar to ours as the result holds if the loss at initialization is already sufficiently low and the proof relies on (local) Polyak-Lojasiewicz inequalities verified by the loss landscape.

Infinite depth. The works of [17, 2, 62, 31, 61, 34, 13, 42] extend those results to arbitrary deep NN in the overparameterized setting. Specifically, the results in [17, 2, 34] apply to deep ResNets. The best result seems to be achieved in [42], with convergence as soon as the last layer has a width $m = \Omega(N^3)$ and at best with linear width. A common feature for those works is to rely on the fact that, for a sufficiently high number of parameters, the model can be well approximated by a linear model corresponding to its first order expansion around the initialization. In [16] this phenomenon, called "lazy regime", is attributed to an inappropriate scaling of the parameters. On the other hand, [35, 34] refer to this phenomenon as "linear" or "kernel regime" and relate it the constancy of the *Neural Tangent Kernel (NTK)* introduced in [26]. However, in all those works the width of intermediary layers has to depend on the depth L of the network. Therefore, these results do not apply to the training of the model in Eq. (1), corresponding to the limit $L \to +\infty$.

Infinite width. The other direction of overparameterization, analyzed in several works [40, 15, 39, 27, 37, 21, 44] is to consider the limit of infinitely wide layers. In such a "mean-field" setting, the model is parameterized by the distribution of the parameters at each layer. In [15, 40, 39, 27] the training dynamic is analyzed as a gradient flow in the Wasserstein space [3], showing that the only stationary distributions are global minimizers of the empirical risk. In [21] a similar result is showed for deep NN with an arbitrary number of infinitely wide layers. In [14, 1], local linear convergence towards the global optimum is shown for two layers NNs in a teacher-student setup with regularized loss. Finally, [37] analyzes the convergence of continuous ResNets with infinitely wide residual layers and shows that every critical point is a global minimizer of the empirical risk. We stress out that these convergence results only apply to infinitely wide NNs. It is not clear if this mean-field limit extends to the parametric setting of MLPs with the Euclidean metric on their parameters. In contrast, a RKHS structure naturally arises when considering a linear parameterization of the residuals. Assumption 1 and Assumption 2 can be satisfied both in a parametric setting with a finite number of features and in a mean-field setting limit where the residuals are generated by a universal kernel.

Contributions. We show convergence results for GD in the training of RKHS-FlowResNets (see Definition 1). These correspond to infinitely deep continuous ResNets with linear parameterization of the residuals. Our first main contribution, in Section 4, shows that under some regularity and expressivity asumptions on the residuals, the associated empirical risk satisfies a (local) Polyak-Lojasiewicz Property 2. A consequence is Theorem 2, which states global convergence of GD towards a global optimum (zero training loss) under the condition that the loss at initialization is already sufficiently low. In the limit where the loss at initialization is arbitrarily small, we recover a linear regime as described in [35, 34]. Our second contribution, in Section 5, shows how this condition for global convergence can be enforced using a first linear layer embedding data into a sufficiently high dimensional space. Thereafter, we show how the assumptions of Theorem 2, can be satisfied for RKHSs generated by a finite number of Random Features, with high probability over the choice of these features. For any dataset $(x^i, y^i)_{1 \leq i \leq N} \in (\mathbb{R}^d \times \mathbb{R}^{d'})^N$, we conclude in Theorem 3 to convergence of GD towards a global minimum of Eq. (7) with high probability when the width of the layers scales polynomially w.r.t. the size of the dataset N and the inverse input data separation δ^{-1}.

Notations In what follows $\|.\|$ denotes the Euclidean L^2 norm for vectors and the Frobenius norm for matrices. For matrices the spectral norm is denoted $\|.\|_2$, the smallest (resp. greatest) singular value is denoted $\sigma_{\min}$ (resp. $\sigma_{\max}$) and for symmetric matrices the smallest (resp. greatest) eigenvalue is denoted $\lambda_{\min}$ (resp. $\lambda_{\max}$). Given some Hilbert space $\mathcal{H}$, the functional Hilbert space $L^2([0,1], \mathcal{H})$ is denoted $L^2(\mathcal{H})$ or L^2 when there is no ambiguity. The notation $\mathcal{O}$ (resp. Ω) means asymptotically inferior (resp. superior) up to multiplicative factor.

3 Analysis of convergence for overparameterized models

In this section, we review methods for analyzing the convergence of overparameterized machine learning models based on [35, 34]. We refer to the appendix for detailed proofs of the statements.

As presented above, we consider an optimization over the variable v in some Hilbert space $\mathcal{H}$, with fixed input and output data, say $v \mapsto F(v) \coloneqq [F(v, x_i)]_{i=1,\dots,N}$. Therefore, the empirical risk is a function of the parameters $v \in \mathcal{H}$. We say that the model is *overparameterized* whenever the dimension $\dim(\mathcal{H})$ of the parameter space is much larger than the dimension of the output space of $F(v)$, here $d'N$. The RKHS-FlowResNet model defined F in Definition 1 fall into this category as $\mathcal{H}$ is the infinite dimensional functional space $L^2([0,1], V)$.

3.1 A (local) Polyak-Lojasiewicz property

When dealing with overparameterized models, one cannot expect the loss to be convex but one expects the model to perfectly fit the data, that is to reach the global minimum value of 0. In fact, for a sufficient number of parameters, the loss landscape typically possesses a continuum of infinitely many global minima and is non-convex in any neighbourhood of a global minima [35]. One thus rather needs to rely on a set of functional inequalities allowing to control the decrease rate of the loss along GD [36, 11].

Definition 2 ((local) Polyak-Lojasiewicz property)**.** *Let $L : \mathcal{H} \to \mathbb{R}_+$ be a differentiable function. We say that L satisfies a (local) Polyak-Lojasiewicz (PL) property if there exist positive continuous functions $m, M : \mathbb{R}_+ \to \mathbb{R}_+^*$ s.t. for every $v \in \mathcal{H}$*

$$2m(\|v\|)L(v) \leq \|\nabla L(v)\|^2 \leq 2M(\|v\|)L(v). \quad (8)$$

Such functional inequalities have already shown to be relevant for proving convergence guarantees in the training of NNs [22]. A first consequence for a loss L which satisfies the (local) PL property of Definition 2 is that it does not admit any spurious local minima but only global minima. Also, if the training dynamic is bounded, then m and M are uniformly lower- and upper-bounded along the dynamic, implying that L decreases at a linear rate. In most cases, m and M are degenerate when $\|v\| \to +\infty$. When the dynamic is not bounded, L can thus decrease to 0 slower than at a linear rate or even converge towards a strictly positive limit.

3.2 Local convergence result

Because of the degeneracy of m and M, it is in general not possible to conclude to an unconditional convergence of GD towards a global minimizer of the empirical risk. However, PL inequalities are sufficient to prove convergence when the problem is not too hard to solve, that is when the loss at initialization is not too high. Moreover, when using gradient descent stepping, one needs to make a supplementary smoothness assumption on the empirical risk L. This ensures that the loss decreases at each gradient step for a sufficiently small step size.

Definition 3 (Smoothness, Definition 2 of [35])**.** *Let $\beta \geq 0$ be a constant. We say that the function $L : \mathbb{R}^m \to \mathbb{R}$ is β-smooth if for every $v, v' \in \mathbb{R}^m$:*

$$\|L(v') - L(v) - \nabla L(v)(v' - v)\| \leq \frac{\beta}{2}\|v' - v\|^2. \quad (9)$$

The local PL property combined with this smoothness assumption then gives a local convergence result for the convergence of GD towards a global minimizer of the empirical risk.

Theorem 1 (Theorem 6 of [35])**.** *Let $L : \mathbb{R}^m \to \mathbb{R}_+$ be a loss function satisfying a local PL property with local constants m and M. Let $v^0 \in \mathbb{R}^m$ be some parameter initialization such that there exists a radius $R \geq 0$ s.t.*

$$2\sqrt{2}\frac{\sqrt{M(\|v^0\| + R)}}{m(\|v^0\| + R)}\sqrt{L(v^0)} \leq R. \quad (10)$$

Furthermore, assume that L is β-smooth within the ball $B(v^0, R)$. Then for a step size $\eta \leq \beta^{-1}$, GD with initialization v^0 and step size η converges towards a

global minimizer of L with a linear convergence rate, i.e. for every $k \geq 0$:

$$L(v^k) \leq (1 - m(\|v^0\| + R)\eta)^k L(v^0). \tag{11}$$

Moreover, the dynamic is bounded:

$$\|v^k - v^0\| \leq R, \ \forall k \geq 0. \tag{12}$$

4 Properties of RKHS-FlowResNets

In this section we analyze the convergence of GD in the training of the infinitely deep ResNet model of Definition 1. Note that such a model is overparameterized in depth as the parameter space is the infinite dimensional space $L^2([0,1], V)$ and overparameterization can also come from width when the RKHS is high (or even infinite) dimensional. Therefore, our proof of convergence heavily relies on a PL property verified by the empirical risk.

Recall that we consider the training of deep ResNets with a linear parameterization of the residuals. The set of residuals is as in Eq. (3) with the metric of Eq. (5) induced by the Frobenius metric (Eq. (4)). This provides V with a RKHS structure [4], whose associated kernel is given for any $z, z' \in \mathbb{R}^q$ by:

$$K(z, z') \coloneqq \langle \varphi(z), \varphi(z') \rangle \, \mathrm{Id}_q \, .$$

and whose associated feature map is given by φ.

Remark 3. *The definition of $\langle ., . \rangle_V$ in Eq. (5) requires that* $\mathrm{Span}(\{\varphi(z) | z \in \mathbb{R}^q\}) = \mathbb{R}^r$ *in order to associate each $v \in V$ to a unique $W \in \mathbb{R}^{q \times r}$. This is satisfied by all the feature maps φ we consider in the following.*

Given a training dataset composed of input data points $(x^i)_{1 \leq i \leq N} \in (\mathbb{R}^d)^N$ and of target data points $(y^i)_{1 \leq i \leq N} \in (\mathbb{R}^{d'})^N$ we are interested in the task of minimizing the empirical risk of Eq. (7) by GD over v. Analogously to back-propagation in discrete NNs architectures, the gradient of L can be expressed thanks to a backward equation derived by adjoint sensitivity analysis [45].

Property 1. *Let L be the empirical risk in Eq.* (7) *associated to the RKHS-FlowResNet model with a quadratic loss. Let K be the kernel function associated to the RKHS V. Then L is differentiable on $L^2([0,1], V)$, with for every $v \in L^2([0,1], V)$:*

$$\nabla L(v) = \sum_{i=1}^{N} K(., z^i) p^i,$$

where for each index $i \in [\![1, N]\!]$ z^i is the solution of Eq. (6) *with initial condition Ax^i and the* adjoint variable p^i *is the solution to the* backward problem*:*

$$\dot{p}_t^i = -Dv_t(z_t^i)^\top p_t^i \text{ and } p_1^i = -\frac{1}{N} B^\top (B z_1^i - y^i). \tag{13}$$

This explicit formulation of the gradient is directly used to prove the PL property.

4.1 PL property of RKHS-FlowResNets

Following the line of proof sketched in Section 3, we show how to derive PL inequalities of the form Eq. (8) for the empirical loss associated to the RKHS-FlowResNet model. In that purpose we make a few assumptions about the RKHS V. The first one concerns its regularity and allows to control the solutions of Eqs. (6) and (13).

Assumption 1 ((strong) Admissibility)**.** *We say that the RKHS V is* (strongly) admissible *if it is continuously embedded in $W^{2,\infty}(\mathbb{R}^q, \mathbb{R}^q)$. More precisely, there exists a constant $\kappa > 0$ s.t.:*

$$\forall v \in V, \ \|v\| + \|Dv\|_{2,\infty} + \|D^2 v\|_{2,\infty} \leq \kappa \|v\|_V. \tag{14}$$

We note this embedding $V \hookrightarrow W^{2,\infty}(\mathbb{R}^q, \mathbb{R}^q)$.

The embedding $V \hookrightarrow W^{1,\infty}(\mathbb{R}^q, \mathbb{R}^q)$ is a natural assumption in order to ensure the regularity of the flow generated by the control parameter [54, 57] and suffices to prove convergence of a continuous gradient flow on the parameter v. Assumption 1 is a bit stronger because a supplementary smoothness result on the loss landscape is necessary to prove convergence of discrete GD (c.f. Definition 3). In practice, κ

can be computed for smooth kernels thanks to Property 4. For example, the RKHS associated to the Gaussian kernel $k : r \mapsto e^{-r^2/2}$ is (strongly) admissible with $\kappa = 2 + \sqrt{3}$.

The second assumption is related to the expressiveness of V and is a weaker form of the classical universality property of RKHSs.

Assumption 2 (N-universality)**.** *Let K be the kernel function associated to the RKHS V. For a family of points $(z^i)_{1\leq i\leq N} \in (\mathbb{R}^q)^N$, we define the associated kernel matrix as the block matrix:*

$$\mathbb{K}((z^i)_i) \coloneqq (K(z^i, z^j))_{1\leq i,j\leq N}\,.$$

We say that V is N-universal *if for every family of such two-by-two disjoint points $(z^i)_{1\leq i\leq N} \in (\mathbb{R}^q)^N$ the associated kernel matrix $\mathbb{K}$ is positive definite. More precisely we assume:*

$$\Lambda \coloneqq \sup_{(z^i)\in(\mathbb{R}^q)^N} \lambda_{\max}(\mathbb{K}((z^i)_i)) < +\infty \qquad (15)$$

and for every $\delta > 0$:

$$\lambda(\delta^{-1}) \coloneqq \inf_{\substack{(z^i)\in(\mathbb{R}^q)^N \\ \min_{i\neq j}\|z^i - z^j\|\geq\delta}} \lambda_{\min}(\mathbb{K}((z^i)_i)) > 0\,. \quad (16)$$

In particular, satisfying Assumption 2 requires having V of dimension $m \geq N$, but it can be satisfied for finite dimensional RKHSs of dimension $m \leq N^q$, for example by considering a polynomial kernel, or by RKHSs of dimension $m \geq poly(N, q)$ with high probability on the sampling of random features as shown in Section 5. On the other hand, even though the existence of λ follows from simple compactness arguments, it seems to be hardly analytically tractable even for classical kernels such as the Gaussian kernel.

Remark 4. *For a RKHS V as in Eq.* (3)*, the properties of V only depend on φ. An interesting example is when $\varphi : z \mapsto \sigma(Uz)$ with σ an activation function applied component-wise and U a fixed feature matrix. In Section 5 we show that, when considering the complex activation $\sigma : t \mapsto e^{-\imath t}$, both assumptions can be satisfied with high probability. On the other hand, Assumption 1 is not satisfied when considering $\sigma = ReLU$ due to its non-smoothness at* 0*.*

Remark 5. *Note that Λ could also be allowed to depend on some parameters, such as $\max\|z^i\|$. However, as it is a more critical aspect of our analysis, we prefer to highlight the dependency of λ w.r.t. $\min_{i\neq j}\|z^i - z^j\|$. For all the RKHSs studied in what follows, we can always take Λ to be a constant depending on N and q.*

The following PL property is satisfied by the empirical risk L.

Property 2 (RKHS-FlowResNets satisfy PL)**.** *Assume V satisfies Assumption 1 with κ and Assumption 2 with λ and Λ. Let L be the empirical risk in Eq.* (7) *associated to the RKHS-FlowResnet model of Definition 1. Then L satisfies the PL inequalities of Definition 2 with m and M given by:*

$$M(R) = \frac{1}{N}\sigma_{\max}(B^\top)^2\Lambda e^{2\kappa R}, \qquad (17)$$
$$m(R) = \frac{1}{N}\sigma_{\min}(B^\top)^2\lambda\left(\sigma_{\min}(A)^{-1}\delta^{-1}e^{\kappa R}\right)e^{-2\kappa R},$$

where $\delta \coloneqq \min_{i\neq j}\|x^i - x^j\|$ is the data separation.

Sketch of proof. Thanks to Assumption 1, we have for every solution p^i of Eq. (13) and for every $t \in [0, 1]$:

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

Moreover using the initial condition we have:

$$\frac{2\sigma_{\min}(B^\top)^2}{N}L(v) \leq \sum_{i=1}^N \|p_1^i\|^2 \leq \frac{2\sigma_{\max}(B^\top)^2}{N}L(v).$$

With similar arguments one finds that for every $t \in [0, 1]$ and every indices $i, j \in [\![1, N]\!]$:

$$\|z_t^i - z_t^j\| \geq \sigma_{\min}(A)\|x^i - x^j\|e^{-\kappa\|v\|_{L^2}},$$

where z^i is the solution of Eq. (6) with initial condition Ax^i.

Then denoting $\tilde{p}_t$ the vector of the stacked p_t^i and using properties of RKHSs, we have for every $t \in [0, 1]$:

$$\|\nabla L(v)_t\|^2 = \sum_{1\leq i,j\leq N}(p_t^i)^\top K(z_t^i, z_t^j)p_t^j = \langle\tilde{p}_t, \mathbb{K}\tilde{p}_t\rangle,$$

where $\mathbb{K}$ is the kernel matrix associated to the points $(z_t^i)_i$. This last equality gives the result using Assumption 2 and considering the previously derived estimates on p^i and z^i. □

Note that the degeneracy of the bounding functions M, m as $R \to +\infty$ readily appears in Eq. (17). Thus one should not expect these bounds to imply global convergence of GD without making any further assumption. Indeed, cases where GD fails to converge towards a global optimizer of the loss are observed in [8], Section 6, with a setup corresponding to the model of Definition 1 with V as in Eq. (3) and $\varphi = \mathrm{Id}_{\mathbb{R}^q}$.

Also, note that the data separation δ plays an important role in Property 2 as it intervenes in the conditioning of the kernel matrix. In what follows, we always assume the data points to have a data separation lower-bounded by $\delta > 0$.

4.2 Convergence of RKHS-FlowResNets

Thanks to the convergence analysis for overparameterized models detailed in Section 3, our main result follows as a direct consequence of the previous property.

Theorem 2. *Let V satisfy Assumption 1 with constant κ and Assumption 2 with λ, Λ. Let v^0 be some initialization of the control parameter s.t. $\|v^0\|_{L^2} = R_0$ and assume there exists a positive radius $R \geq 0$ with:*

$$\frac{\sqrt{8}\sigma_{\max}(B^\top)\sqrt{N\Lambda L(v^0)}e^{3\kappa(R+R_0)}}{\sigma_{\min}(B^\top)^2\lambda(\sigma_{\min}(A)^{-1}\delta^{-1}e^{\kappa(R+R_0)})} \leq R\,. \quad (18)$$

Then, for a sufficiently small step-size $\eta > 0$, GD with step-size η converges towards a minimizer of the training loss at a linear rate. More precisely, for every $k \geq 0$:

$$L(v^k) \leq (1 - \eta\mu)^k L(v^0), \quad (19)$$

with μ given by:

$$\mu \coloneqq \frac{1}{N}\sigma_{\min}(B^\top)^2\lambda\left(\sigma_{\min}(A)^{-1}\delta^{-1}e^{\kappa(R+R_0)}\right)e^{-2\kappa(R+R_0)}.$$

Moreover, the training dynamic stays bounded in the ball of radius R: $\|v^k - v^0\|_{L^2} \leq R$ for all k.

Sketch of the proof. Following Theorem 2, it only remains to show that the RKHS-FlowResNet model is locally smooth. Consider two control parameters $v, \bar{v} \in L^2([0,1], V)$ and associated solutions $z^i, \bar{z}^i$ and $p^i, \bar{p}^i$ of Eq. (6) and Eq. (13). Then using Assumption 1 one can derive estimates on $z^i - \bar{z}^i$ and $p^i - \bar{p}^i$. The result follows by controlling the quantity $\|\nabla L(v) - \nabla L(\bar{v})\|_{L^2}$ and by establishing a bound of the form: $\|\nabla L(v) - \nabla L(\bar{v})\|_{L^2} \leq C\|v - \bar{v}\|_{L^2}$. □

As Theorem 1, Theorem 2 is a local convergence result in which the condition in Eq. (18) expresses a threshold between two kinds of behaviours: **(i)** if $L(v^0)$ is sufficiently small, the training dynamic converges towards a global minimizer. The limiting behaviour is when the l.h.s. of Eq. (18) tends to 0. Because of a regularizing effect of GD (i.e. that $\|v^k - v^0\|_{L^2} \leq R$), the parameter stays in a ball of arbitrary small radius R all along the training dynamic. In this limit, we recover a "linear" or "kernel" regime where the model is well approximated by its linearization at v^0 [15, 34, 26]. **(ii)** If $L(v^0)$ is too large, the result says nothing about the convergence of the GD. However, it is still observed in practice that the training dynamic often converges towards a global minimizer of the loss [59]. Explaining this phenomenon in a general setting remains a challenging open question.

5 Enforcing global convergence with high dimensional embedding and finite width

As Theorem 2 is a local convergence result, it does not allow to conclude to a general convergence behaviour of GD in the training of RKHS-FlowResNets. In the following, we show how one can enforce the hypothesis of Theorem 2 to be verified and prove two global convergence results. The first one relies on increasing the embedding dimension q in order to satisfy Eq. (18) and applies in the case of infinite width,

i.e. with residual layers in a universal RKHS. The second result recovers global convergence in a finite width regime, relying on a high number r of Random Fourier Features and a high embedding dimension q to satisfy Eq. (18).

For the sake of readability we only consider here the case where V belongs to a restricted class of RKHSs and refer to Appendix C.1 for more general results and complete proofs. For some positive parameter $\nu > 0$ we consider the Matérn kernel k defined in [56]:

$$\forall r \in \mathbb{R}_+,\ k(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{2\pi} r \right)^{\nu} \mathcal{K}_\nu(\frac{\sqrt{2\nu}}{2\pi} r), \quad (20)$$

where Γ is the Gamma function and $\mathcal{K}_\nu$ is the modified Bessel function of the second kind. Equivalently, k can be defined by its frequency distribution over $\mathbb{R}^q$ as:

$$\forall x \in \mathbb{R}^q,\ k(\|x\|) = \int_{\mathbb{R}^q} e^{\imath \langle x, \omega \rangle} \mu_q(\omega) \mathrm{d}\omega, \quad (21)$$
$$\text{with} \quad \mu_q(\omega) = C(q,\nu) \left(1 + \frac{\|\omega\|^2}{2\nu} \right)^{-(\frac{q}{2}+\nu)},$$

where $C(q,\nu)$ is some normalizing constant. For every $q \geq 1$, such a function is known to define a structure of vector-valued RKHS V_q over $\mathbb{R}^q$ [50, 56] corresponding to the Sobolev space $H^{\nu+q/2}(\mathbb{R}^q, \mathbb{R}^q)$. The associated kernel is given for every $z, z' \in \mathbb{R}^q$ by:

$$K_q(z, z') = k(\|z - z'\|) \, \mathrm{Id}_q \, .$$

Note that it is important for this RKHS to depend on the ambient dimension q. In particular the Sobolev space $H^s(\mathbb{R}^q, \mathbb{R}^q)$ is a RKHS if and only if it has regularity $s > q/2$. Assuming $\nu > 2$, μ_q further admits up to 4 finite order moment [28] implying that k is four times differentiable at 0. Then V_q satisfies Assumption 1 with some constant κ depending only on ν and given by Property 4:

$$\kappa = \sqrt{k(0)} + \sqrt{-k''(0)} + \sqrt{k^{(4)}(0)}$$
$$= 1 + \sqrt{\frac{\nu}{\nu - 1}} + \sqrt{\frac{3\nu^2}{(\nu-1)(\nu-2)}}. \quad (22)$$

Also, V_q satisfies Assumption 2 with λ and Λ depending on ν, q and N.

Note that with this choice of scaling for k and μ_q, one recovers the Gaussian kernel $k : r \mapsto e^{-r^2/2}$ in the limit $\nu \to +\infty$ [56]. Thereafter we will consider, $\nu \in (2, +\infty]$, the case $\nu = +\infty$ referring to the Gaussian kernel. We also assume for simplicity that the data distribution is compactly supported, arguing that the proofs can easily be adapted to milder assumptions. In particular there exists some $r_0 \geq 0$ so that every input data x verifies $\|x\| \leq r_0$.

5.1 Global convergence with high-dimensional embedding

We first show how Eq. (18) can be satisfied by considering a sufficiently high embedding dimension q and appropriate embedding matrices A and B. Doing so, the Euclidean distance between the data points, i.e. the model's loss, is preserved whereas the conditioning of the kernel matrix can be controlled.

In what follows, we consider for any $q \geq 1$ the matrices:

$$A_q \coloneqq q^{-1/4} (\mathrm{Id}_d, ..., \mathrm{Id}_d, 0)^\top \in \mathbb{R}^{q \times d},$$
$$B_q \coloneqq q^{1/4} (\mathrm{Id}_{d'}, 0...0) \in \mathbb{R}^{d' \times q},$$

where there are $\lfloor q/d \rfloor$ copies of Id_d in A_q. In particular we have:

$$\sigma_{\min}(A_q) = q^{-1/4} \sqrt{\lfloor q/d \rfloor},$$
$$\sigma_{\min}(B_q^\top) = \sigma_{\max}(B_q^\top) = q^{1/4}$$

and $B_q A_q \in \mathbb{R}^{d' \times d}$ is independent of q.

Proposition 1. *Let $\nu \in (2, +\infty]$. There exists some constant $C \geq 0$ so that for any $N \geq 2$, any $\delta \in (0, 1]$ and any dataset $(x^i, y^i)_{1 \leq i \leq N} \in (\mathbb{R}^{d'} \times \mathbb{R}^d)^N$ with input data separation δ GD converges towards a zero-training-loss optimum in the training of the RKHS-FlowResNet model of Definition 1 with matrices A_q, B_q, RKHS V_q and initialization $v^0 = 0$ as soon as:*

$$q \geq C \big(N^4 + \delta^{-4} \log(N)^4 \big) \, . \quad (23)$$

Sketch of the proof. We give the proof for $\nu < +\infty$.

First of all, note that, as is well-known, the modified Bessel function $\mathcal{K}_\nu$ and thus the Matérn kernel

k as defined in Eq. (20) has exponential decay as r tends to infinity (see for example [6, 30]). Therefore, there exist constants G_ν, H_ν depending only on ν such that for every $r \geq 0$:

$$k(r) \leq G_\nu e^{-H_\nu^{-1} r}. \tag{24}$$

Then, using $d^2 \lfloor q/d \rfloor^2 \geq q(q-2d)$, considering:

$$q \geq 2d + d^2 \frac{H_\nu^4 \log(2G_\nu N)^4}{\delta^4 e^{-4\kappa R}} \tag{25}$$

is enough to ensure that:

$$q^{-1/4}\sqrt{\lfloor q/d \rfloor}\delta e^{-\kappa R} \geq H_\nu \log(2G_\nu N).$$

Then, by the bound in Eq. (24) for any point cloud $(z^i)_{1\leq i\leq N} \in (\mathbb{R}^q)^N$ with data separation $q^{-1/4}\sqrt{\lfloor q/d \rfloor}\delta e^{-\kappa R}$ we have:

$$\forall 1 \leq i < j \leq N,\ |k(\|z^i - z^j\|)| \leq \frac{1}{2N}.$$

Thus, the kernel matrix $\mathbb{K} = (k(\|z^i - z^j\|)\,\mathrm{Id}_q)_{i,j}$ is diagonally dominant with:

$$\lambda_{\min}(\mathbb{K}) \geq 1 - \frac{N-1}{2N} \geq \frac{1}{2},$$

and by definition of λ in Eq. (16):

$$\lambda(\sigma_{\min}(A_q)^{-1}\delta^{-1}e^{\kappa R}) \geq \frac{1}{2}. \tag{26}$$

Moreover, $\Lambda \leq N$ because k is bounded by 1.

Finally:

$$\frac{\sigma_{\max}(B_q^\top)}{\sigma_{\min}(B_q^\top)^2} = q^{-1/4}, \tag{27}$$

and also $\|B_q A_q\|_2$ is independent of q so that $L(0) = \frac{1}{N}\sum_{i=1}^N \|B_q A_q x^i - y^i\|^2 \leq C$, for some constant C independent of q, δ and N because the data distribution has compact support. Putting Eq. (26) and Eq. (37) into Eq. (18):

$$\frac{2\sqrt{2}\sigma_{\max}(B_q^\top)\sqrt{N\Lambda L(0)}e^{3\kappa R}}{\sigma_{\min}(B_q^\top)^2\lambda(\sigma_{\min}(A_q)^{-1}\delta^{-1}e^{-\kappa R})} \leq 4\sqrt{2C}e^{3\kappa R}\frac{N}{q^{1/4}}.$$

Considering $R > 0$ is fixed (c.f. Remark 6), Theorem 2 can be applied as soon as:

$$q \geq 2^{10} C^2 e^{12\kappa R} R^{-4} N^4 \tag{28}$$

and combining this bound with the one in Eq. (25) gives the result. □

Remark 6 (Choice of R)**.** *The proof of Proposition 1 holds for any fixed $R > 0$ whose choice impacts the result through the constant C. There is a trade-off between minimizing $e^{4\kappa R}$ to have a better dependency of q w.r.t. $\delta^{-1}\log(N)$ in Eq. (25) and minimizing $R^{-1}e^{3\kappa R}$ to have a better dependency w.r.t. N in Eq. (28). However, in any case, optimizing w.r.t. R only improves the result up to a constant factor.*

As shown in Appendix C.1, Proposition 1 easily generalizes while considering non-zero initializations v^0, as soon as these initializations scale as $\mathcal{O}(q^{-1/4})$ w.r.t. q. For residuals of the form Eq. (3) with W having i.i.d. Gaussian entries this amounts to add a $\mathcal{O}(q^{-3/4})$ rescaling factor.

5.2 Global convergence with finite width

In the preceding we showed that, in the case of an RKHS defined by a Matérn kernel, convergence of GD can be ensured by increasing the embedding dimension. However, for practical implementations, the form of the residual in Eq. (3) forces us to consider RKHSs defined by feature maps. A way to overcome this difficulty and to benefit from the properties of a wide range of kernel functions is to consider an approximation by *Random Fourier Features (RFF)* [46, 47].

More precisely, given $q \geq 1$, recall the definition of the Matérn kernel k in terms of its frequency distribution μ_q over $\mathbb{R}^q$ in Eq. (21) and for any sampling $\omega^1, ..., \omega^r \overset{iid}{\sim} \mu_q$ of size r, consider the feature map:

$$\varphi : z \in \mathbb{R}^q \mapsto \frac{1}{\sqrt{r}}(e^{\imath\langle z, \omega^j\rangle})_{1\leq j\leq r} \in \mathbb{C}^r. \tag{29}$$

Remark 7 (Sampling)**.** *Note that μ_q identifies as the density of a q-variate t-distribution with shape*

parameter 2ν [28]. Sampling over μ_q can be achieved using that for $Y \sim \mathcal{N}(0, \mathrm{Id}_q)$ and for u distributed according to $\chi^2_{2\nu}$, the chi-squared distribution with 2ν degrees of freedom, $Y/\sqrt{u/2\nu}$ is distributed according to μ_q.

In other words, considering the complex activation $\sigma : t \mapsto e^{\imath t}$ applied component-wise and $U \coloneqq (\omega^1 | \dots | \omega^r) \in \mathbb{R}^{q\times r}$ the feature matrix, we have $\varphi(z) = \frac{1}{\sqrt{r}}\sigma(U^\top z)$. Recall, that such a feature map defines a structure of RKHS:

$$\hat{V}_q \coloneqq \{W\varphi(.) \mid W \in \mathbb{R}^{q\times r}\},$$

Such a $\hat{V}_q$ can be viewed as a finite-dimensional approximation of the universal RKHS V_q as it is associated to the kernel function $\hat{K}_q(z,z') \coloneqq \hat{k}(z,z')\,\mathrm{Id}_q$, with:

$$\hat{k}(z,z') \coloneqq \langle\varphi(z),\varphi(z')\rangle = \frac{1}{r}\sum_{j=1}^{r} e^{\imath\langle z-z',\omega^j\rangle} \xrightarrow{r\to+\infty} k(\|z-z'\|),$$

where the convergence holds almost surely by the law of large numbers.

Given any $q \geq 1$, we show that, with high probability over the choice of features, $\hat{V}_q$ recovers the properties of admissibility and universality of V_q as soon as r is sufficiently high w.r.t. q and N:

Proposition 2. *Consider any $q, N \geq 2$ and any $\epsilon, \tau, R > 0$.*

***(i)** Assume $\nu > 4$. For $r \geq \Omega(\tau q^8)$, with probability greater than $1-\tau^{-1}$, $\hat{V}_q$ satisfies Assumption 1 with some $\hat{\kappa} \leq \kappa + 1$.*

***(ii)** For $r \geq \Omega(\epsilon^{-2}N^2(q\log(\|A\|_2 r_0 + R) + \tau))$, with probability greater than $1-e^{-\tau}$, for any control parameter $v \in L^2([0,1], \hat{V}_q)$ s.t. $\|v\|_{L^2} \leq R$ and any time $t \in [0,1]$:*

$$\lambda_{\min}(\hat{\mathbb{K}}((z_t^i)_i)) \geq \lambda_{\min}(\mathbb{K}((z_t^i)_i)) - \epsilon,$$

where the $(z^i)_i$ are the solutions to Eq. (6) and $\hat{\mathbb{K}}$, $\mathbb{K}$ are the kernel matrices associated to $\hat{k}$ and k respectively.

Sketch of Proof. **Proof of (i).** First of all, note that for $\nu > 4$, μ_q admits up to 8^{th}-order finite moments and these can be bounded uniformly in q [28].

Let φ be the feature map of Eq. (29). Then for every $z \in \mathbb{R}^q$, $\|\varphi(z)\| \leq 1$ so that for every $v \in \hat{V}_q$, $\|v\|_\infty \leq \|W\|\|\varphi\|_\infty \leq \|v\|_V$.

For the differential Dv we have for every $z \in \mathbb{R}^q$:

$$D\varphi(z) = \frac{1}{\sqrt{r}}\left(\omega_i^j e^{-\imath\langle z,\omega^j\rangle}\right)_{\substack{1\leq i\leq q\\ 1\leq j\leq r}} \in \mathbb{R}^{r\times q}.$$

Summing on the index j gives by the law of large numbers that $D\varphi(z)^* D\varphi(z) = \frac{1}{r}\sum_{j=1}^r \omega^j(\omega^j)^\top$ converges in probability to $-k''(0)\,\mathrm{Id}_q$ as $r \to +\infty$. The convergence rate depends on q and can be controlled by the Bienaymé-Chebyshev inequality, using that μ_q has finite 4^{th} order moments. Finally, for r sufficiently high w.r.t. q, τ and α, one has $\|Dv\|_{2,\infty} = \|WD\varphi\|_{2,\infty} \leq (-k''(0)+\alpha)\|v\|_{\hat{V}_q}$, with probability greater than $1-\tau^{-1}$.

The same idea applies to bound $\|D^2 v\|_{2,\infty}$, using that μ_q has finite 8^{th}-order moments. The result follows using that κ is given by Eq. (22).

Proof of (ii). For $t \in [0,1]$, we consider $(z_t^i)_i$ the solutions of of Eq. (6) for some control parameter $v \in L^2([0,1], \hat{V}_q)$ and we introduce the kernel matrices:

$$\hat{\mathbb{K}}_t = (\hat{K}_q(z_t^i, z_t^j))_{1\leq i,j\leq N},\ \mathbb{K}_t = (K_q(z_t^i, z_t^j))_{1\leq i,j\leq N}.$$

Using the first point, we know that if $\|v\|_{L^2} \leq R$, then $\|z_t^i\| \leq \|Ax^i\| + (\kappa+1)R$. Then, using Theorem 1 in [51], we have for every indices i,j and every $t \in [0,1]$:

$$\mathbb{P}\Big(|\hat{k}(z_t^i, z_t^j) - k(\|z_t^i - z_t^j\|)| \geq \frac{h(q,R)+\sqrt{2\tau}}{\sqrt{r}}\Big) \leq e^{-\tau},$$

with $h(q,R) = \mathcal{O}(\sqrt{q\log(\|A\|_2 r_0 + R)})$. Therefore, choosing $r \geq \Omega\left(\epsilon^{-2}N^2(q\log(\|A\|_2 r_0 + R) + \tau)\right)$, we have with probability greater than $1-e^{-\tau}$, $\lambda_{\min}(\hat{\mathbb{K}}_t) \geq \lambda_{\min}(\mathbb{K}_t) - \epsilon$, for any $t \in [0,1]$. □

Finally, combining the results of Proposition 1 and Proposition 2 we obtain a global convergence theorem for ResNets of finite width.

Theorem 3 (Global convergence)**.** *Assume $\nu > 4$. There exists some constant $C \geq 0$ such that, for any $N \geq 2$, any $\delta \in (0, 1]$, any dataset $(x^i, y^i) \in (\mathbb{R}^d \times \mathbb{R}^{d'})^N$ with input data separation δ and any $\tau > 0$, GD initialized at $v^0 = 0$ converges with probability at least $1 - \tau^{-1}$ towards a zero training loss optimum in the training of the RKHS-FlowResNet model of Definition 1 with a feature map φ such as in Eq.* (29) *as soon as:*

$$q \geq C\big(N^4 + \delta^{-4}\log(N)^4\big),\ r \geq C\tau q^8. \tag{30}$$

Proof. Consider $R = 1$. Thanks to Proposition 1, we can have q large enough so that in Eq. (18):

$$\frac{8\sqrt{2}\sigma_{\max}(B^\top)\sqrt{N\Lambda L(v^0)}e^{3(\kappa+1)}}{\sigma_{\min}(B^\top)^2\lambda(\sigma_{\min}(A)^{-1}\delta^{-1}e^{(\kappa+1)})} \leq 1,$$

for some matrices $B \in \mathbb{R}^{d'\times q}, A \in \mathbb{R}^{q\times d}$ and with κ, λ and Λ associated to k. For the Matérn kernel k in Eq. (20), this is achieved as soon as $q \geq \Omega(N^4 + \delta^{-4}\log(N)^4)$. Also, by the proof of Proposition 1 we can have A such that:

$$\lambda(\sigma_{\min}(A)^{-1}\delta^{-1}e^{(\kappa+1)}) \geq \frac{1}{2}$$

Then, taking $\epsilon = \frac{1}{4}$ in Proposition 2, the condition in Eq. (18) is satisfied by $\hat{V}_q$ with probability greater than $1 - \tau^{-1}$ as soon as $r \geq \Omega(\tau q^8 + \tau q N^2 \log(1 + \|A\|_2 r_0))$. Recalling the proof of Proposition 1 we have $\|A\|_2 = q^{-1/4}\sqrt{\lfloor q/d \rfloor}$ and the dominant term is $\Omega(\tau q^8)$. □

Theorem 3 concludes that it suffices to have a networks width $r \geq \Omega(\tau(N^{32} + \delta^{-32}\log(N)^{32}))$ (i.e. $r \geq poly(\tau, N, \delta^{-1})$) in order to have convergence of GD towards a global optimum with probability greater than $1 - \tau^{-1}$. Note that in order to obtain this result, the choice of the matrices A_q, B_q in Proposition 1 is not restrictive and the dependency of q and r w.r.t. N and δ might be improved for other well-chosen embedding matrices.

We show an illustration in Fig. 1. We chose to perform experiences on small synthetic datasets and with a Matérn kernel of finite parameter ν instead of Gaussian kernels. This last choice is motivated by the fact that we observed that Gaussian kernels led to poorer performances. We observe an important acceleration of GD when q and r increase simultaneously, starting from $q = d$. On the other hand, there seems to be only a low impact of increasing q as soon as $q \geq Nd$. When q is fixed, a similar behaviour is observed when increasing r. GD performs poorly for $r \leq q$ and equally well for values of r above a certain threshold.

6 Conclusion

In this paper, we have identified a relevant infinite width limit (RKHS-FlowResNet) for a simplified model of ResNet. We showed that GD converges linearly when training this model and that a network's width polynomial w.r.t. to the size of the dataset is sufficient to maintain this property.

A natural extension of our result is to study the convergence of GD when also training the hidden layers of the residuals. A first step towards this general case consists in studying the corresponding mean field model where the residuals are parameterized by density distributions over the neurons [15, 40, 39, 27, 37, 21] for each residual blocks. Interestingly, such a parametrization of the residual blocks is still linear in this measure and thus fits into our framework of linear in parameters. However, it would require a finer mathematical analysis to obtain similar results.

Acknowledgements

The work of Gabriel Peyré was supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19- P3IA-0001 (PRAIRIE 3IA Institute) and by the European Research Council (ERC project NORIA).

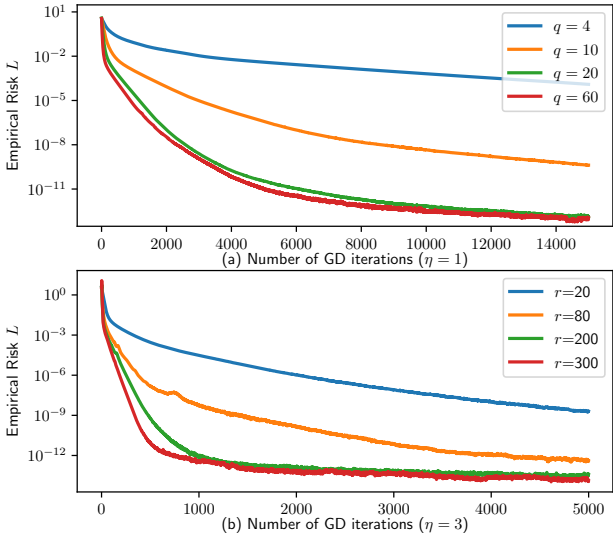

Figure 1: Evolution of the empirical risk along GD. Each plot is an average over 12 batches. Each batch consists of 10 points in dimension $d = d' = 2$ with $x \sim \mathcal{N}(0, \mathrm{Id}_2)$ and $y = -x + 0.2\epsilon$, $\epsilon \sim \mathcal{N}(0, \mathrm{Id}_2)$. In (a) the dataset is embedded in a varying dimension q and we set $r = 2q$. In (b) the dataset is embedded a fixed dimension $q = 30$ and a varying r is used. V_q is the Sobolev space $H^{(q+5)/2}$ approximated by RFFs and we used embedding matrices $A_q = (\mathrm{Id}_d, 0, ..., 0)^\top \in \mathbb{R}^{q \times d}$ and $B_q = (\mathrm{Id}_{d'}, ..., \mathrm{Id}_{d'}) \in \mathbb{R}^{d' \times q}$.

References

[1] S. Akiyama and T. Suzuki, *On learnability via gradient method for two-layer relu neural networks in teacher-student setting*, arXiv e-prints, (2021), pp. arXiv–2106.

[2] 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.

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

[4] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American mathematical society, 68 (1950), pp. 337–404.

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

[6] Á. Baricz, *Bounds for modified bessel functions of the first and second kinds*, Proceedings of the Edinburgh Mathematical Society, 53 (2010), pp. 575–599.

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

[8] 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.

[9] 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.

[10] M. Belkin, D. Hsu, S. Ma, and S. Mandal, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proceedings of the National Academy of Sciences, 116 (2019), pp. 15849–15854.

[11] 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.

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

[13] Z. Chen, Y. Cao, D. Zou, and Q. Gu, *How much over-parameterization is sufficient to learn deep relu networks?*, in International Conference on Learning Representations, 2020.

[14] L. Chizat, *Sparse optimization on measures with over-parameterized gradient descent*, Mathematical Programming, (2021), pp. 1–46.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] C. Fang, J. Lee, P. Yang, and T. Zhang, *Modeling from features: a mean-field framework for over-parameterized deep neural networks*, in Conference on Learning Theory, PMLR, 2021, pp. 1887–1936.

[22] S. Frei and Q. Gu, *Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent*, in Thirty-Fifth Conference on Neural Information Processing Systems, 2021.

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

[24] M. Hardt and T. Ma, *Identity Matters in Deep Learning*, arXiv:1611.04231 [cs, stat], (2018). arXiv: 1611.04231.

[25] 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.

[26] 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.

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

[28] B. M. G. Kibria and A. Joarder, *A short review of multivariate t-distribution*, Journal of Statistical Research ISSN, 40 (2006), pp. 256–422.

[29] 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.

[30] A. Laforgia and P. Natalini, *Some inequalities for modified bessel functions*, Journal of Inequalities and Applications, 2010 (2010), pp. 1–10.

[31] 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.

[32] Y. Li and Y. Liang, *Learning overparameterized neural networks via stochastic gradient descent on structured data*, Advances in neural information processing systems, (2018).

[33] Y. Li and Y. Yuan, *Convergence analysis of two-layer neural networks with relu activation*, Advances in Neural Information Processing Systems, 30 (2017), pp. 597–607.

[34] 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).

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

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

[37] Y. Lu, C. Ma, Y. Lu, J. Lu, and L. Ying, *A mean field analysis of deep resnet and beyond: Towards provably optimization via over-parameterization from depth*, in International

Conference on Machine Learning, PMLR, 2020, pp. 6426–6436.

[38] Y. Lu, A. Zhong, Q. Li, and B. Dong, *Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations*, in International Conference on Machine Learning, PMLR, 2018, pp. 3276–3285.

[39] 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.

[40] 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.

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

[42] 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.

[43] M. Niethammer, Y. Huang, and F.-X. Vialard, *Geodesic regression for image time-series*, in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part II, G. Fichtinger, A. Martel, and T. Peters, eds., Berlin, Heidelberg, 2011, Springer Berlin Heidelberg, pp. 655–662.

[44] H. T. Pham and P.-M. Nguyen, *Global convergence of three-layer neural networks in the mean field regime*, in International Conference on Learning Representations, 2020.

[45] L. S. Pontryagin, *Mathematical theory of optimal processes*, CRC press, 1987.

[46] 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.

[47] ——, *Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning*, Advances in Neural Information Processing Systems, 21 (2008), pp. 1313–1320.

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

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

[50] B. Schölkopf, A. J. Smola, F. Bach, et al., *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT Press, 2002.

[51] B. Sriperumbudur and Z. Szabo, *Optimal rates for random fourier features*, Advances in Neural Information Processing Systems, 28 (2015), pp. 1144–1152.

[52] D. J. Sutherland and J. Schneider, *On the error of random fourier features*, in Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, 2015, pp. 862–871.

[53] M. Thorpe and Y. van Gennip, *Deep limits of residual neural networks*, arXiv preprint arXiv:1810.11741, (2018).

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

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

[56] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, vol. 2, MIT press Cambridge, MA, 2006.

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

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

[59] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning (still) requires rethinking generalization*, Communications of the ACM, 64 (2021), pp. 107–115.

[60] M. Zhou, R. Ge, and C. Jin, *A local convergence theory for mildly over-parameterized two-layer neural network*, in COLT, 2021.

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

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

A Proofs of Section 3

We give a proof of Theorem 1. This essentially follows the proof given in [35].

Proof of Theorem 1. Assume the loss L satisfies Definition 2 with M and m and that Eq. (10) is satisfied at initialization $v^0 \in \mathbb{R}^m$. The proof proceeds by induction over the gradient step k

Assume the convergence rate Eq. (11) and the regularization bound Eq. (12) is satisfied for every $l \leq k$. Then at step $k+1$:

$$\begin{aligned}\|v^{k+1} - v^0\| &= \|\eta \sum_{l=0}^{k} \nabla L(v^l)\| \\ &\leq \eta \sum_{l=0}^{k} \|\nabla L(v^l)\| \\ &\leq \eta \sum_{l=0}^{k} \sqrt{2M(\|v^l\|)L(v^l)}.\end{aligned}$$

Using the induction hypothesis and setting $\mu = m(\|v^0\| + R)$ we have:

$$\begin{aligned}\|v^{k+1} - v^0\| &\leq \eta\sqrt{2M(\|v^0\| + R)L(v^0)} \sum_{l=0}^{k} (1-\eta\mu)^{-l/2} \\ &\leq \eta\sqrt{2M(\|v^0\| + R)L(v^0)}(1 - \sqrt{1-\eta\mu})^{-1} \\ &\leq \frac{2}{\mu}\sqrt{2M(\|v^0\| + R)L(v^0)} \\ &\leq R,\end{aligned}$$

where the last inequality is Eq. (10). We thus recovered the regularization bound Eq. (12) at step $k+1$.

Moreover, because v^{k+1} is located in $B(v^0, R)$ we have thanks to the smoothness assumption:

$$\begin{aligned}L(v^{k+1}) &\leq L(v^k) - \eta\|\nabla L(v^k)\|^2 + \eta^2\frac{\beta}{2}\|\nabla L(v)\|^2 \\ &\leq L(v^k) - \frac{\eta}{2}\|\nabla L(v^k)\|^2,\end{aligned}$$

because $\eta \leq \beta^{-1}$. Thus using the lower bound in the PL inequality Eq. (8):

$$L(v^{k+1}) \leq L(v^k)(1 - m(\|v^0\| + R)\eta),$$

which gives the convergence rate of Eq. (11) at step $k+1$ by induction on k.

□

B Proofs of Section 4

B.1 About the definition of RKHS-FlowResNets

Before deriving proofs for the properties of our RKHS-FlowResNet model, it is interesting to study carefully the well-posedness of Definition 1. Indeed, because the control parameter v is only integrable in time and not continuous, the Cauchy-Lipschitz theorem does not ensure that there exist solutions to Eq. (6). Instead we rely on a weaker notion of solution and use a result from Carathéodory (Section I.5 in [23]).

Proposition 3. *Let V be some RKHS satisfying Assumption 1 and $v \in L^2([0,1], V)$ be some control parameter. Then for every $x \in \mathbb{R}^d$ there exists a unique solution z of Eq. (6) in the weak sense of absolutely continuous functions. More precisely there exists a unique $z \in H^1([0,1], \mathbb{R}^q)$ such that for every $t \in [0,1]$:*

$$z_t = Ax + \int_0^t v_s(z_s)ds\,. \tag{31}$$

Proof. The map $(t,z) \in [0,1]\times\mathbb{R}^q \mapsto v_t(z)$ is measurable and by Assumption 1 we have for every $t \in [0,1]$ and every $z \in \mathbb{R}^q$:

$$\|v_t(z)\| \leq \kappa\|v_t\|_V\,,$$

whose upper-bound is integrable w.r.t. $t \in [0,1]$. Then, applying Theorem 5.1 of [23] gives a unique absolutely continuous solution z of Eq. (31). Applying Assumption 1 once again, we have that $\dot{z}$ is square integrable and thus z is in H^1. □

In the paper, every equality implying derivatives has to be understood in the sense of weak derivatives of H^1 functions. In particular, this notion allows to perform integration by parts, which is used in the following proof of Property 1.

Proof Property 1. Consider the optimization problem of minimizing the empirical risk of Eq. (7) with F the RKHS-FlowResNet model of Definition 1 and a dataset $(x^i, y^i)_{1\leq i\leq N} \in (\mathbb{R}^d \times \mathbb{R}^{d'})^N$. Introducing for every index $i \in [\![1, N]\!]$ the variables $z^i \in H^1([0,1], \mathbb{R}^q)$ solutions of Eq. (6), this can be viewed as an optimisation problem over $((z^i)_i, v)$ under the constraint that Eq. (6) is satisfied:

$$\min_{\substack{(z^i)_i \in H^1(\mathbb{R}^q)^N \\ v\in L^2(V)}} \frac{1}{2N}\sum_{i=1}^N \|Bz_1^i - y^i\|^2$$
$$\text{with } \forall i \in [\![1,N]\!], \left\{ \begin{array}{rcl} \dot{z}_t^i &=& v_t(z_t^i)\ \forall t \in [0,1] \\ z_0^i &=& Ax^i. \end{array}\right.$$

Introducing the adjoint variables $(p^i)_i \in H^1(\mathbb{R}^q)^N$, the Lagrangian of the optimization problem is defined as:

$$\begin{aligned}\mathcal{L}((z^i),(P^i),v) :=& \sum_{i=1}^N \big(\frac{1}{2N}\|Bz_1^i - y^i\| \\ &+ \int_0^1 \langle p_t^i, \dot{z}_t^i - v_t(z_t^i)\rangle \mathrm{d}t\big) \\ =& \sum_{i=1}^N \big(\frac{1}{2N}\|Bz_1^i - y^i\| + \left[\langle p_t^i, z_t^i\rangle\right]_0^1 \\ &- \int_0^1 \langle \dot{p}_t^i, z_t^i\rangle \mathrm{d}t - \int_0^t \langle p_t^i, v_t(z_t^i)\rangle \mathrm{d}t\big),\end{aligned}$$

where the second equality is established by integration by parts. Therefore, the condition for optimality over z^i is equivalent to Eq. (13). For every index i:

$$\nabla_{z^i}\mathcal{L} = 0 \Leftrightarrow \left\{ \begin{array}{rcl} \dot{p}_t^i &=& -Dv_t(z_t^i)p_t^i \\ p_1^i &=& -\frac{1}{N}B^\top(Bz_1^i - y^i), \end{array}\right.$$

which has to be understand in the sense of weak solutions in H^1.

The gradient of L is obtained by differentiating over the v variable. Denoting δ_z^p the linear form $v \mapsto \langle v(z), p\rangle$, we have:

$$\begin{aligned}\nabla L(v) &= \nabla_v \mathcal{L}((z^i),(p^i),v) \\ &= -\sum_{i=1}^N K * \delta_{z^i}^{p^i} \\ &= -\sum_{i=1}^N K(.,z^i)p^i,\end{aligned}$$

with K the kernel function of the RKHS V and $K*: V^* \to V$ the associated isometry[1].

□

B.2 Proof of Property 2

We prove here that for any given dataset $(x^i, y^i)_{1\leq i\leq N} \in (\mathbb{R}^d \times \mathbb{R}^{d'})^N$, the empirical risk L associated to the RKHS-FlowResNet model satisfies a (local) Polyak-Lojasiewicz property. As stated in Property 2. The proof uses Assumption 1 to derive estimates on the solutions of Eq. (6) and Eq. (13), which we give in the following lemma:

Lemma 1. *Let V satisfy Assumption 1 with constant κ and let $v \in L^2([0,1], V)$ be some control parameter.*

(i) *Let $(z^i)_{1\leq i\leq N}$ be the solutions of Eq. (6) for some data inputs $(x^i)_{1\leq i\leq N} \in (\mathbb{R}^d)^N$. Then for every indices $i, j \in [\![1, N]\!]$ and every time $t \in [0,1]$:*

$$\|z^i - z^j\| \geq \sigma_{\min}(A)e^{-\kappa\|v\|_{L^2}}\|x^i - x^j\|. \qquad (32)$$

(ii) *Let $(p^i)_{1\leq i\leq N}$ be the solutions of Eq. (13) associated to $(z^i)_{1\leq i\leq N}$ with objective outputs $(y^i)_{1\leq i\leq N} \in (\mathbb{R}^{d'})^N$. Then for every $i \in [\![1, N]\!]$ and every time $t \in [0,1]$:*

$$\|p_t^i\| \geq \frac{\sigma_{\min}(B^\top)}{N}e^{-\kappa\|v\|_L^2}\|Bz_1^i - y^i\|,$$
$$\|p_t^i\| \leq \frac{\sigma_{\max}(B^\top)}{N}e^{\kappa\|v\|_L^2}\|Bz_1^i - y^i\|.$$

Proof of Lemma 1. **Proof of (i)** Let $i, j \in [\![1, N]\!]$. Assume by contradiction that for some time $t \in [0,1]$ we have:

$$\|z_t^i - z_t^j\| < e^{-\kappa\|v\|_{L^2}}\|z_0^i - z_0^j\|.$$

[1]The notation $K*$ reminds of convolution which is the case when the kernel is translation invariant.

Then because z^i and z^j are absolutely continuous, $\|z^i - z^j\|^2$ is absolutely continuous and for any time $s \in [0,1]$:

$$\begin{aligned}\|z_s^i - z_s^j\|^2 &= \|z_t^i - z_t^j\|^2 + 2\int_t^s \langle v_r(z_r^i) - v_r(z_r^j), z_r^i - z_r^j\rangle \mathrm{d}r \\ &\leq \|z_t^i - z_t^j\|^2 + 2\int_t^s \kappa \|v_r\|_V \|z_r^i - z_r^j\|^2 \mathrm{d}r,\end{aligned}$$

where the inequality follows from $\|Dv_r\|_{2,\infty} \leq \kappa \|v_r\|_V$. Applying Grönwall's lemma, we have:

$$\|z_s^i - z_s^j\|^2 \leq \|z_t^i - z_t^j\|^2 e^{2\kappa \|v\|_{L^2}},$$

and by setting $s = 0$:

$$\|z_0^i - z_0^j\|^2 \leq \|z_t^i - z_t^j\|^2 e^{2\kappa \|v\|_{L^2}} < \|z_0^i - z_0^j\|,$$

which is a contradiction. Therefore for any time $t \in [0,1]$:

$$\|z_t^i - z_t^j\| \geq e^{-\kappa \|v\|_{L^2}} \|z_0^i - z_0^j\|,$$

and the result follows by considering the initial condition $z_0^i = Ax^i$.

Proof of (ii) Let $i \in [\![1, N]\!]$ be any index and let p^i be the solution of Eq. (13) with initial condition $p_1^i = -\frac{1}{N} B^\top (Bz_1^i - y^i)$. Then because p^i is absolutely continuous, $\|p^i\|$ is absolutely continuous and for any time $t \leq s \in [0,1]$:

$$\|p_t^i\|^2 = \|p_1^i\|^2 - 2\int_1^t \langle Dv_s(z_s^i) p_s^i, p_s^i \rangle \mathrm{d}s,$$

so that using Assumption 1 we have:

$$\|p_s^i\|^2 \leq \|p_t^i\|^2 + 2\int_t^s \kappa \|v_r\|_V \|p_r^i\|^2 \mathrm{d}r\,.$$

Using Grönwall's lemma in the first inequality and setting $s = 0$ we have:

$$\|p_1^i\|^2 \leq \|p_t^i\|^2 e^{2\kappa \|v\|_{L^2}},$$

and proceeding by contradiction (such as in (i)) we have:

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

The result follows by considering the initial condition on p_1^i. □

Provided those estimates on z^i and p^i, it remains to use Assumption 2 in order to conclude.

Proof of Property 2. Let $v \in L^2([0,1], V)$ and consider the form of the gradient of L given by Property 1 with $(z^i)_{1 \leq i \leq N}$ the solutions of Eq. (6) and $(p^i)_{1 \leq i \leq N}$ the solutions of Eq. (13). Let $t \in [0,1]$, then by definition of the norm in RKHSs:

$$\|\nabla L(v)_t\|_V^2 = \sum_{1 \leq i,j \leq N} (p_t^i)^\top K(z_t^i, z_t^j) p_t^j\,,$$

where we recall that K is the kernel associated to V. Noting $p \coloneqq (p_t^i) \in \mathbb{R}^{Nq}$, the vector of the stacked $(p_t^i)_{1 \leq i \leq N}$, and $\mathbb{K}$ the kernel matrix associated to the family of points $(z_t^i)_i$, we have:

$$\|\nabla L(v)_t\|_V^2 = \langle p, \mathbb{K} p \rangle\,.$$

Then by Assumption 2, there exists a non-increasing function λ and a constant Λ such that:

$$\begin{aligned}&\|\nabla L(v)_t\|_V^2 \leq \Lambda \|p\|^2, \\ &\|\nabla L(v)_t\|_V^2 \geq \lambda(\max_{1 \leq i,j \leq N} \|z_t^i - z_t^j\|^{-1}) \|p\|^2.\end{aligned}$$

Using (i) in Lemma 1 we have:

$$\lambda(\max_{1 \leq i,j \leq N} \|z_t^i - z_t^j\|^{-1}) \geq \lambda(\sigma_{\min}(A)^{-1} \delta^{-1} e^{\kappa \|v\|_{L^2}}),$$

where $\delta \coloneqq \min_{1 \leq i,j \leq N} \|x^i - x^j\|$ is the data separation. Finally the result follows by using (ii). More precisely:

$$\begin{aligned}\|p\|^2 &= \sum_{i=1}^N \|p_t^i\|^2 \\ &\leq \frac{\sigma_{\max}(B^\top)^2}{N^2} e^{2\kappa \|v\|_{L^2}} \sum_{i=1}^N \|Bz_1^i - y^i\|^2 \\ &= 2\frac{\sigma_{\max}(B^\top)^2}{N} e^{2\kappa \|v\|_{L^2}} L(v),\end{aligned}$$

and in the same manner:

$$\|p\|^2 \geq 2\frac{\sigma_{\min}(B^\top)^2}{N} e^{-2\kappa \|v\|_{L^2}} L(v).$$

□

B.3 Proof of Theorem 2

Theorem 2 is a direct consequence of Property 2. In order to apply Theorem 1, it suffices to show that L satisfies some smoothness assumption:

Property 3 (Smoothness of L)**.** *Let V be some RKHS satisfying Assumption 1. Let L be the empirical risk defined on $L^2([0,1],V)$ and associated to the RKHS-FlowResNet model. Then there exists a continuous function $\mathbf{C}:\mathbb{R}_+\to\mathbb{R}_+^*$ such that for every $R\geq 0$ and every $v,\bar{v}\in L^2([0,1],V)$ with $\|v\|_{L^2},\|\bar{v}\|_{L^2}\leq R$:*

$$\|\nabla L(v)-\nabla L(\bar{v})\|_{L^2}\leq \mathbf{C}(R)\|v-\bar{v}\|_{L^2}.$$

We note κ the constant associated to Assumption 1. The proof of Property 3 relies on the following lemma:

Lemma 2. *Let $v,\bar{v}\in L^2([0,1],V)$ be some control parameters and $R\geq 0$ be some radius such that $\|v\|_{L^2},\|\bar{v}\|_{L^2}\leq R$. Let $(x,y)\in\mathbb{R}^d\times\mathbb{R}^{d'}$ be some pair of data input / objective output.*

(i) *Let $z,\bar{z}$ be solutions of Eq.* (6) *with parameter v and $\bar{v}$ respectively and with the same initial condition Ax, then for any $t\in[0,1]$:*

$$\|z_t-\bar{z}_t\|\leq \kappa e^{\kappa R}\|v-\bar{v}\|_{L^2}.$$

(ii) *Let $p,\bar{p}$ be solutions of Eq.* (13) *with parameter v and $\bar{v}$ respectively and with initial condition $\frac{1}{N}B^\top(Bz_1-y)$ and $\frac{1}{N}B^\top(B\bar{z}_1-y)$, then for any $t\in[0,1]$:*

$$\|p_t-\bar{p}_t\|\leq \frac{\kappa e^{2\kappa R}\|B\|_2}{N}\|v-\bar{v}\|_{L^2}\left[\|B\|_2+\|B(\bar{z}_1-y)\|(1+Re^{\kappa R})\right].$$

Proof of Lemma 2. **Proof of (i)** For every time $t\in[0,1]$ we have:

$$\begin{aligned}z_t-\bar{z}_t&=\int_0^t\big(v_s(z_s)-\bar{v}_s(\bar{z}_s)\big)\mathrm{d}s\\&=\int_0^t\big(v_s(z_s)-v_s(\bar{z}_s)+v_s(\bar{z}_s)-\bar{v}_s(\bar{z}_s)\big)\mathrm{d}s,\end{aligned}$$

and by triangle inequality:

$$\begin{aligned}\|z_t-\bar{z}_t\|&\leq\int_0^t\big(\|v_s(z_s)-v_s(\bar{z}_s)\|+\|v_s(\bar{z}_s)-\bar{v}_s(\bar{z}_s)\|\big)\mathrm{d}s\\&\leq\int_0^t\kappa\|v_s\|_V\|z_s-\bar{z}_s\|\mathrm{d}s+\int_0^t\kappa\|v_s-\bar{v}_s\|_V\mathrm{d}s,\end{aligned}$$

where we used Assumption 1 in the second inequality. Therefore, by Grönwall's lemma:

$$\begin{aligned}\|z_t-\bar{z}_t\|&\leq\kappa e^{\kappa\|v\|_{L^2}}\int_0^t\|v_s-\bar{v}_s\|_V\mathrm{d}s\\&\leq\kappa e^{\kappa R}\|v-\bar{v}\|_{L^2}.\end{aligned}$$

Proof of (ii) For any $t\in[0,1]$ we have:

$$\begin{aligned}p_t-\bar{p}_t=&(p_1-\bar{p}_1)-\int_1^t\big(Dv_s(z_s)^\top p_s-D\bar{v}_s(\bar{z}_s)^\top\bar{p}_s\big)\mathrm{d}s\\=&(p_1-\bar{p}_1)-\int_1^t Dv_s(z_s)^\top(p_s-\bar{p}_s)\mathrm{d}s\\&-\int_1^t\big(Dv_s(z_s)-Dv_s(\bar{z}_s)\big)^\top\bar{p}_s\mathrm{d}s\\&-\int_1^t\big(Dv_s(\bar{z}_s)-D\bar{v}_s(\bar{z}_s)\big)^\top\bar{p}_s\mathrm{d}s,\end{aligned}$$

and using the triangle inequality and Assumption 1:

$$\begin{aligned}\|p_t-\bar{p}_t\|\leq&\|p_1-\bar{p}_1\|+\int_t^1\kappa\|v_s\|_V\|p_s-\bar{p}_s\|\mathrm{d}s\\&+\int_t^1\kappa\|v_s\|_V\|z_s-\bar{z}_s\|\|\bar{p}_s\|\mathrm{d}s\\&+\int_t^1\kappa\|v_s-\bar{v}_s\|_V\|\bar{p}_s\|\mathrm{d}s.\end{aligned}$$

Then, using Grönwall's lemma backward in time gives:

$$\begin{aligned}\|p_t-\bar{p}_t\|\leq&\|p_1-\bar{p}_1\|e^{\kappa\|v\|_{L^2}}\\&+\kappa e^{\kappa\|v\|_{L^2}}\int_t^1\|v_s-\bar{v}_s\|_V\|\bar{p}_s\|\mathrm{d}s\\&+\kappa e^{\kappa\|v\|_{L^2}}\int_t^1\|v_s\|_V\|z_s-\bar{z}_s\|\|\bar{p}_s\|\mathrm{d}s.\end{aligned}$$

On one hand, because of (i) we have for every $s\in[0,1]$:

$$\|z_s-\bar{z}_s\|\leq\kappa e^{\kappa R}\|v-\bar{v}\|_{L^2},$$

and also:

$$\|p_1 - \bar{p}_1\| = \frac{1}{N}\|B^\top B(z_1 - \bar{z}_1)\| \leq \frac{\|B\|_2^2}{N}\kappa e^{\kappa R}\|v - \bar{v}\|_{L^2}.$$

On the other hand, recalling (ii) of Lemma 1, for every $s \in [0,1]$:

$$\|\bar{p}_s\| \leq \frac{\sigma_{\max}(B^\top)}{N}e^{\kappa R}\|Bz_1 - y\|.$$

The result follows by putting these estimates in the preceding inequality:

$$\begin{aligned}\|p_t - \bar{p}_t\| \leq & \frac{\|B\|_2^2}{N}\kappa e^{2\kappa R}\|v - \bar{v}\|_{L^2} \\ & + \frac{\sigma_{\max}(B^\top)}{N}\kappa e^{2\kappa R}\|B(\bar{z}_1 - y)\|\|v - \bar{v}\|_{L^2} \\ & + R\frac{\sigma_{\max}(B^\top)}{N}\kappa^2 e^{3\kappa R}\|B(\bar{z}_1 - y)\|\|v - \bar{v}\|_{L^2}.\end{aligned}$$

□

Proof of Property 3. Let $v, \bar{v} \in L^2([0,1], V)$ with $\|v\|_{L^2}, \|\bar{v}\|_{L^2} \leq R$. Then taking the same notation as in Lemma 2, we have for any $t \in [0,1]$:

$$\begin{aligned}\nabla L(v)_t - \nabla L(\bar{v})_t &= \sum_{i=1}^N K(., z_t^i)p_t^i - \sum_{i=1}^N K(.\bar{z}_t^i)\bar{p}_t^i \\ &= \sum_{i=1}^N K(., z_t^i)(p_t^i - \bar{p}_t^i) \\ &\quad + \sum_{i=1}^N (K(., z_t^i) - K(.\bar{z}_t^i))\bar{p}_t^i,\end{aligned}$$

and we can write $\|\nabla L(v)_t - \nabla L(\bar{v})_t\|_V \leq T_1 + T_2$ with:

$$T_1 = \|\sum_{i=1}^N K(., z_t^i)(p_t^i - \bar{p}_t^i)\|_V,$$

$$T_2 = \|\sum_{i=1}^N (K(., z_t^i) - K(.\bar{z}_t^i))\bar{p}_t^i\|_V.$$

First we consider deriving an upper bound on T_1. Note that by the definition of the norm in RKHSs and by Assumption 2 we have:

$$\begin{aligned}T_1^2 &= \sum_{1\leq i,j\leq N}(p_t^i - \bar{p}_t^i)^\top K(z_t^i, z_t^j)(p_t^j - \bar{p}_t^j) \\ &\leq \Lambda \sum_{i=1}^N \|p_t^i - \bar{p}_t^i\|^2.\end{aligned}$$

Therefore, using (ii) from Lemma 2 to bound $\|p_t^i - \bar{p}_t^i\|$ for every index i we get:

$$T_1^2 \leq \Lambda \mathbf{C}_1^2 \|v - \bar{v}\|_{L^2}^2,$$

with:

$$\begin{aligned}\mathbf{C}_1^2 &= \sum_{i=1}^N \frac{\kappa^2 e^{4\kappa R}\|B\|_2^2}{N^2}\left[\|B\|_2 + \|B(\bar{z}_1^i - y)\|(1 + Re^{\kappa R})\right]^2 \\ &\leq \sum_{i=1}^N \frac{2\kappa^2 e^{4\kappa R}\|B\|_2^2}{N^2}\left[\|B\|_2^2 + \|B(\bar{z}_1^i - y)\|^2(1 + Re^{\kappa R})^2\right] \\ &\leq \frac{2\kappa^2 e^{4\kappa R}\|B\|_2^4}{N} + \frac{4\kappa^2 e^{4\kappa R}\|B\|_2^2}{N}(1 + Re^{\kappa R})^2 L(\bar{v}),\end{aligned}$$

where we recognised $L(\bar{v})$ in the third line. By continuity of L we can define for every $R \geq 0$:

$$L^*(R) \coloneqq \sup_{\|v\|_{L^2}\leq R} L(v).$$

And therefore:

$$\begin{aligned}\mathbf{C}_1^2 &\leq \frac{2\kappa^2 e^{4\kappa R}\|B\|_2^4}{N} + \frac{4\kappa^2 e^{4\kappa R}\|B\|_2^2}{N}(1 + Re^{\kappa R})^2 L^*(R) \\ &=: \mathbf{C}_3(R)^2.\end{aligned}$$

We then consider deriving an upper-bound on T_2. By triangle inequality:

$$T_2 \leq \sum_{i=1}^N \|(K(., z_t^i) - K(., \bar{z}_t^i))\bar{p^i}_t\|_V.$$

Consider any $\alpha \in V$, then for any index $i \in [\![1, N]\!]$, by the reproducing property:

$$\begin{aligned}\langle (K(., z_t^i) - K(., \bar{z}_t^i))\bar{p}_t^i, \alpha\rangle_V &= \langle \alpha(z_t^i) - \alpha(\bar{z}_t^i), \bar{p}_t^i\rangle \\ &\leq \kappa\|\alpha\|_V\|z_t^i - \bar{z}_t^i\|\|\bar{p}_t^i\|,\end{aligned}$$

where we used the Cauchy-Schwarz inequality and Assumption 1 applied to α. Therefore, by duality:

$$\|(K(.,z_t^i) - K(.,\bar{z}_t^i))\bar{p^i}_t\|_V \leq \kappa \|z_t^i - \bar{z}_t^i\| \|\bar{p}_t^i\|.$$

Using the estimates of Lemma 1 and Lemma 2 we get:

$$\begin{aligned}\|(K(.,z_t^i) - K(.,\bar{z}_t^i))\bar{p^i}_t\|_V \\ \leq \frac{\kappa^2 e^{2\kappa R}\|B\|_2}{N}\|B\bar{z}_1^i - y^i\|\|v-\bar{v}\|_{L^2}.\end{aligned}$$

And finally, using Cauchy-Schwarz inequality and recognizing $L(\bar{v})$ we have:

$$\begin{aligned}T_2^2 &\leq N\sum_{i=1}^N \|(K(.,z_t^i) - K(.,\bar{z}_t^i))\bar{p^i}_t\|_V^2 \\ &\leq \mathbf{C}_2^2\|v-\bar{v}\|_{L^2}^2,\end{aligned}$$

with:

$$\begin{aligned}\mathbf{C}_2^2 &= 2\kappa^4 e^{4\kappa R}\|B\|_2^2 L(\bar{v}) \\ &\leq 2\kappa^4 e^{4\kappa R}\|B\|_2^2 L^*(R) =: \mathbf{C}_4(R)^2.\end{aligned}$$

Therefore we obtain the result by setting:

$$\mathbf{C}(R) = \left[\Lambda \mathbf{C}_3(R)^2 + \mathbf{C}_4(R)^2\right]^{1/2}.$$

□

Provided with Property 3, we can finish the proof of Theorem 2.

Proof of Theorem 2. By Property 2, L satisfies the PL inqualities of Definition 2 and the proof is a direct corollary of Theorem 1. It only remains to show that the smoothness condition of Definition 3 is verified.

Let $v, \bar{v} \in L^2([0,1], V)$ such that $\|v\|_{L^2}, \|\bar{v}\|_{L^2} \leq R$ for some radius $R \geq 0$. Then we have:

$$\begin{aligned}L(\bar{v}) =& L(v) + \int_0^1 \nabla L(v + t(\bar{v}-v)).(\bar{v}-v)\mathrm{d}t \\ =& L(v) + \nabla L(v).(\bar{v}-v) \\ &+ \int_0^1 \left[\nabla L(v+t(\bar{v}-v)) - \nabla L(v)\right]\cdot(\bar{v}-v)\mathrm{d}t.\end{aligned}$$

Using Property 3, there exists some $\mathbf{C}(R)$ such that:

$$\|\nabla L(v + t(\bar{v}-v)) - \nabla L(v)\|_{L^2} \leq t\mathbf{C}(R)\|\bar{v}-v\|_{L^2}.$$

This gives the inequality:

$$L(\bar{v}) \leq L(v) + \nabla L(v)\cdot(\bar{v}-v) + \frac{\mathbf{C}(R)}{2}\|\bar{v}-v\|_{L^2}^2,$$

which is the desired result. □

C Proofs of Section 5

The results in Section 5 show how the condition for convergence in Eq. (18) can be enforced by considering suitable RKHSs of vector-fields in sufficiently high dimension q. We give in Appendix C.3 examples of suitable kernels.

In the following, we assume that for every $q \geq 1$ we are provided with a function $k_q : \mathbb{R}_+ \to \mathbb{R}$ such that the induced symmetric rotationally-invariant kernel K_q defined by:

$$\forall z, z' \in \mathbb{R}^q,\ K_q(z,z') = k_q(\|z-z'\|)\,\mathrm{Id}_q, \qquad (33)$$

is a positive-definite kernel over $\mathbb{R}^q$. Without loss of generality, one can assume k_q to be normalized, that is $k_q(0) = 1$. We note V_q the vector-valued RKHS associated to K_q. The properties of V_q are then entirely determined by k_q. In particular, smoothness of the kernel at 0 implies regularity of the vector-fields in V_q:

Property 4 (Regularity of V_q)**.** *Let $k_q : \mathbb{R}_+ \to \mathbb{R}$ be some function defining a positive symmetric kernel K_q. If k_q is 4 times differentiable at 0, with $k_q'(0) = k_q^{(3)}(0) = 0$. Then V_q satisfies Assumption 1 with constant $\kappa = \sqrt{k_q(0)} + \sqrt{-k_q''(0)} + \sqrt{k_q^{(4)}(0)}$.*

As a consequence, if the derivatives of k_q can be bounded uniformly over q then V_q satisfies Assumption 1 with some constant κ independent of q. This, is the case for the Matérn kernel k defined in Eq. (20).

Proof. The proof proceeds by duality arguments. For $q \geq 1$, consider some $v \in V_q$. Then for any $z \in \mathbb{R}^q$

and any $\alpha \in V_q$, by the reproducing properties of RKHSs:

$$\begin{aligned}\langle v(z), \alpha\rangle &= \langle v, K_q(., z)\alpha\rangle_{V_q} \\ &\leq \|v\|_{V_q}\|K_q(., z)\alpha\|_{V_q} \\ &= \|v\|_{V_q}\big(\langle \alpha, K_q(z, z)\alpha\rangle\big)^{1/2} \\ &\leq \sqrt{k_q(0)}\|v\|_{V_q}\|\alpha\|.\end{aligned}$$

Therefore, by duality $\|v(z)\| \leq \sqrt{k_q(0)}\|v\|_{V_q}$ and then by taking the supremum over $z \in \mathbb{R}^q$:

$$\|v\|_\infty \leq k_q(0)\|v\|_{V_q}.$$

Then for any $z \in \mathbb{R}^q$ any $\alpha, \beta \in \mathbb{R}^q$ and any $h \in \mathbb{R}_+$:

$$\begin{aligned}&\langle v(z + h\alpha) - v(z), \beta\rangle \\ &= \langle v, (K_q(., z + h\alpha) - K_q(., z))\beta\rangle \\ &\leq \|v\|_{V_q}\|(K_q(., z + h\alpha) - K_q(., z))\beta\|_{V_q}.\end{aligned}$$

In the r.h.s we have using the Taylor's expansion of k_q at 0.:

$$\begin{aligned}&\|(K_q(., z + h\alpha) - K_q(., z))\beta\|^2_{V_q} \\ &= \begin{pmatrix} \beta \\ -\beta \end{pmatrix}^\top \begin{pmatrix} k_q(0)Id_q & k_q(h\|\alpha\|)Id_q \\ k_q(h\|\alpha\|)Id_q & k(0)Id_q \end{pmatrix} \begin{pmatrix} \beta \\ -\beta \end{pmatrix} \\ &= 2\|\beta\|^2(k_q(0) - k_q(h\|\alpha\|)) \\ &= -\|\beta\|^2h^2\|\alpha\|^2k_q''(0) + o(h^2).\end{aligned}$$

Taking the limit $h \to 0$:

$$\begin{aligned}\langle Dv(z)\alpha, \beta\rangle &= \lim_{h\to 0} h^{-1}\langle v(z + h\alpha) - v(z), \beta\rangle \\ &\leq \sqrt{-k_q''(0)}\|v\|_{V_q}\|\alpha\|\|\beta\|,\end{aligned}$$

and therefore $\|Dv(z)\|_2 \leq \sqrt{-k_q''(0)}\|v\|_{V_q}$.

Finally, let us bound $\|D^2v\|_{2,\infty}$. For any $z \in \mathbb{R}^q$ any $\alpha, \beta, \gamma \in \mathbb{R}^q$ and any $h, l \geq 0$ we have in the same manner:

$$\begin{aligned}&\langle v(z + h\beta + l\alpha) - v(z + h\beta) - v(z + l\alpha) + v(z), \gamma\rangle \\ &\leq \|v\|_{V_q}\|\beta\|\|\alpha\|\|\gamma\|hl\sqrt{k_q^{(4)}(0)} + o(hl)\end{aligned}$$

where the second line is obtained by Taylor expansion of k_q at 0. Therefore, passing to the limit $h, l \to 0$:

$$\begin{aligned}&\langle D^2v(z)(\alpha, \beta), \gamma\rangle \\ &= \lim_{h,l\to 0} h^{-1}l^{-1}\langle v(z + h\beta + l\alpha) \\ &\qquad - v(z + h\beta) - v(z + l\alpha) + v(z), \gamma\rangle \\ &\leq \sqrt{k_q^{(4)}(0)}\|v\|_{V_q}\|\beta\|\|\alpha\|\|\gamma\|,\end{aligned}$$

and therefore $\|D^2v(z)\|_2 \leq \sqrt{k_q^{(4)}(0)}\|v\|_{V_q}$.

Setting $\kappa = \sqrt{k_q(0)} + \sqrt{-k_q''(0)} + \sqrt{k_q^{(4)}(0)}$ we obtain the result. Moreover, choosing appropriate v in the above proof, inequalities become sharp and one observes that the constant κ is optimal.

□

C.1 Enforcing convergence with high dimensional embedding and universal kernels

Here we investigate the dependency of Eq. (18) w.r.t. q, δ and N for the class of RKHS V_q and thereby recover the proof of Proposition 1.

We make the following assumption concerning the decay of k_q at infinity:

Assumption 3 (Decay of k_q)**.** *For every $q \geq 1$, $k_q(x)$ tends to 0 when x tends to infinity and we note $\beta_{q,N} > 0$ s.t.:*

$$\forall x \geq \beta_{q,N},\ |k_q(x)| \leq \frac{1}{2N}.$$

Moreover for fixed N we assume that

$$\beta_{q,N} = o_{q\to+\infty}(q^{1/4}).$$

Recall that for any $q \geq 1$ we consider the matrices:

$$\begin{aligned}A_q &\coloneqq q^{-1/4}(\mathrm{Id}_d, ..., \mathrm{Id}_d, 0)^\top \in \mathbb{R}^{q\times d}, \\ B_q &\coloneqq q^{1/4}(\mathrm{Id}_{d'}, 0...0) \in \mathbb{R}^{d'\times q},\end{aligned}$$

where there are $\lfloor q/d \rfloor$ copies of Id_d in A_q. In particular we have:

$$\begin{aligned}\sigma_{\min}(A_q) &= q^{-1/4}\sqrt{\lfloor q/d \rfloor} \simeq q^{1/4}, \\ \sigma_{\min}(B_q^\top) &= \sigma_{\max}(B_q^\top) = q^{1/4}\end{aligned}$$

and $B_qA_q \in \mathbb{R}^{d'\times d}$ is independent of q. We also consider for every $q \geq 1$ some control parameter initialization $V_q^0 \in L^2(V_q)$ such that $\|v_q^0\|_{L^2} \leq R_0 q^{-1/4}$ and recall that we assume the data distribution to be compactly supported.

Proposition 4. *Assume Assumption 3 is satisfied and V_q satisfies Assumption 1 for every $q \geq 1$ with constant κ independent of q. Then there exists some constant $C \geq 0$ so that for any $N \geq 2$, any $\delta \in (0,1]$ and any dataset $(x^i, y^i)_{1\leq i\leq N} \in (\mathbb{R}^{d'}\times\mathbb{R}^d)^N$ with input data separation δ GD converges towards a zero-training-loss optimum in the training of the RLHS-FlowResNet model of Definition 1 with matrices A_q, B_q, RKHS V_q and initialization v_q^0 as soon as:*

$$q \geq CN^4, \text{ and } q \geq C\delta^{-4}\beta_{q,N}^4. \tag{34}$$

Note that the second condition in Eq. (34) can always be ensured for large enough q thanks to Assumption 3. In the case of the Matérn kernel k defined in Eq. (20), such an assumption is verified because it has exponential decay and it is independent of q. Hence, Proposition 1 is a direct consequence of Proposition 4.

Proof of Proposition 4. Let $q \geq 1$. Using the fact that $d^2\lfloor q/d\rfloor^2 \geq q(q-2d)$, considering:

$$q \geq 2d + d^2\frac{\beta_{q,N}^4}{\delta^4 e^{-4\kappa(R+R_0)}} \tag{35}$$

is enough to ensure that:

$$q^{-1/4}\sqrt{\lfloor q/d\rfloor}\delta e^{-\kappa(R+R_0)} \geq \beta_{q,N}.$$

Then, by Assumption 3 for any point cloud $(z^i)_{1\leq i\leq N} \in (\mathbb{R}^q)^N$ with data separation $q^{-1/4}\sqrt{\lfloor q/d\rfloor}\delta e^{-\kappa(R+R_0)}$ we have:

$$\forall 1 \leq i < j \leq N,\ |k_q(\|z^i - z^j\|)| \leq \frac{1}{2N}.$$

Thus, the kernel matrix $\mathbb{K} = (k_q(\|z^i - z^j\|)\,\mathrm{Id}_q)_{i,j}$ is diagonally dominant with:

$$\lambda_{\min}(\mathbb{K}) \geq 1 - \frac{N-1}{2N} \geq \frac{1}{2},$$

and by definition of λ in Eq. (18):

$$\lambda(\sigma_{\min}(A_q)^{-1}\delta^{-1}e^{\kappa(R+R_0)}) \geq \frac{1}{2}. \tag{36}$$

Moreover, $\Lambda \leq N$ because k_q is bounded by 1.

Let $x \in B(0, r_0)$ and assume z is a solution of Eq. (6) for the control parameter v_q^0 and with initial condition A_qx. We have at time $t = 1$:

$$z_1 = A_qx + \int_0^1 (v_q^0)_t(z_t)\mathrm{d}t,$$

so that by triangle inequality and Assumption 1:

$$\|z_1 - A_qx\| \leq \kappa\|v_q^0\|_{L^2},$$

and then because $\|v_q^0\| \leq R_0q^{-1/4}$ and the dataset is compactly supported:

$$\begin{aligned}\|F(v_q^0, x)\| &= \|B_qz_1\| \\ &\leq \|B_qA_qx\| + \|B_q(z_1 - A_qx)\| \\ &\leq \|B_qA_q\|_2r_0 + \kappa R_0,\end{aligned}$$

with B_qA_q independent of q. Thus $L(v_q^0) \leq C$ for some constant C independent of q, N and δ.

Finally:

$$\frac{\sigma_{\max}(B_q^\top)}{\sigma_{\min}(B_q^\top)^2} = q^{-1/4}, \tag{37}$$

and putting Eq. (26) and Eq. (37) into the l.h.s. Eq. (18) gives:

$$\frac{2\sqrt{2}\sigma_{\max}(B_q^\top)\sqrt{N\Lambda L(0)}e^{3\kappa(R+R_0)}}{\sigma_{\min}(B_q^\top)^2\lambda(\sigma_{\min}(A_q)^{-1}\delta^{-1}e^{-\kappa(R+R_0)})} \leq 4\sqrt{2C}e^{3\kappa(R+R_0)}\frac{N}{q^{1/4}}.$$

Considering $R > 0$ is fixed (c.f. Remark 6), Theorem 2 can be applied as soon as:

$$q \geq 2^{10}C^2e^{12\kappa(R+R_0)}R^{-4}N^4 \tag{38}$$

and combining this bound with the one in Eq. (35) gives the result. □

C.2 Enforcing convergence with high dimensional embedding en finite dimensional kernels

We recover here the result of Proposition 2 for the more general kernel k_q. In particular notice that, as an application of Bochner's theorem [48], for every $q \geq 1$ there exists some probability measure μ_q over $\mathbb{R}^q$ such that:

$$\forall z \in \mathbb{R}^q,\ k_q(\|z\|) = \int_{\mathbb{R}^q} e^{\imath\langle z,\omega\rangle} \mathrm{d}\mu_q(\omega). \tag{39}$$

Then, such as in Eq. (29) for the Matérn kernel, for any independent sampling $\omega^j \sim \mu_q$ of size r one can consider the feature map:

$$\varphi : z \mapsto \left(e^{\imath\langle z,\omega^j\rangle}\right)_{1\leq j\leq r} \in \mathbb{C}^r. \tag{40}$$

Such a feature map induces a structure of RKHS $\hat{V}_q$ which is the set of residuals of Eq. (3) with activation φ. The associated kernel is $\hat{K}_q : (z, z') \mapsto \hat{k}_q(z, z')\,\mathrm{Id}_q$ with:

$$\begin{aligned}\forall z, z' \in \mathbb{R}^q,\ \hat{k}_q(z, z') &\coloneqq \langle\varphi(z), \varphi(z')\rangle \\ &\xrightarrow{r\to+\infty} k_q(\|z - z'\|),\end{aligned}$$

almost surely, by the law of large numbers.

We make the following assumption on μ_q:

Assumption 4 (Moments of μ_q)**.** *The measure μ_q admits finite moments up to order 8:*

$$\mathbb{E}_{\mu_q}\left[\prod_{j=1}^{8} |\omega_{i_j}|\right] < \infty,\ \forall i_1, ..., i_8 \in [\![1, q]\!].$$

Moreover, we assume those moments are independent of q.

Note that Assumption 4 implies regularity on the function k_q. Indeed by Fourier inversion theorem we have for every $r \in \mathbb{R}_+$ and every $\theta \in \mathbb{S}^{d-1}$:

$$k_q(r) = \mathbb{E}_{\mu_q}\left[e^{\imath r\langle\theta,\omega\rangle}\right].$$

By theorems of derivation under the integral k_q is 8^{th}-time differentiable on $\mathbb{R}_+$ and for $0 \leq l \leq 8$:

$$k_q^{(l)}(r) = \mathbb{E}_{\mu_q}\left[(\imath\langle\theta,\omega\rangle)^l e^{\imath r\langle\theta,\omega\rangle}\right].$$

In particular, k_q is four time differentiable at 0 and:

$$\begin{aligned}k'(0) &= \mathbb{E}_{\mu_q}\left[\imath\langle\theta,\omega\rangle\right] \\ k^{(3)}(0) &= \mathbb{E}_{\mu_q}\left[-\imath\langle\theta,\omega\rangle^3\right]\end{aligned}$$

Therefore, $k_q'(0)$ and $k_q^{(3)}(0)$ are in $\imath\mathbb{R} \cap \mathbb{R} = \{0\}$ and Property 4 holds. Moreover, as the moments are independent of q, the associated κ is also independent of q.

Proposition 5. *Consider any $q, N \geq 1$ and any $\epsilon, \tau, R > 0$.*

(i) *Assume Assumption 4 is satisfied. For $r \geq \Omega(\tau q^8)$, with probability greater than $1 - \tau^{-1}$, $\hat{V}_q$ satisfies Assumption 1 with some $\hat{\kappa} \leq \kappa + 1$.*

(ii) *For $r \geq \Omega(\epsilon^{-2}N^2(q\log(\|A\|_2 r_0 + R) + \tau))$, with probability greater than $1 - e^{-\tau}$, for any control parameter $v \in L^2([0,1], \hat{V}_q)$ s.t. $\|v\|_{L^2} \leq R$ and any time $t \in [0,1]$:*

$$\lambda_{\min}(\hat{\mathbb{K}}((z_t^i)_i)) \geq \lambda_{\min}(\mathbb{K}((z_t^i)_i)) - \epsilon,$$

where the $(z^i)_i$ are the solutions to Eq. (6) and $\hat{\mathbb{K}}$, $\mathbb{K}$ are the kernel matrices associated to $\hat{k}$ and k respectively.

As Assumption 4 is satisfied for the Matérn kernel k defined in Eq. (20) as soon as $\nu > 4$, Proposition 2 is a direct consequence of Proposition 5.

Proof of Proposition 5. As the proof of (ii) already holds in full generality it only remains to show that (i) is true for general functions k_q satisfying our assumptions.

Proof of (i) We already saw that thanks to the assumption on the moments of μ_q, the RKHS V_q associated to k_q satisfies Assumption 1 with constant κ.

Then we want to prove that for sufficiently high r, the RKHS $\hat{V}_q$ generated by the feature map φ in Eq. (29), satisfies Assumption 1.

Let $v \in \hat{V}_q$ be of the form:

$$v : z \mapsto W\varphi(z)$$

for some $W \in \mathbb{R}^{q\times r}$. For $z \in \mathbb{R}^q$, $\|\varphi(z)\| = 1$ and thus:

$$\|v(z)\| = \|W\varphi(z)\| \leq \|W\| = \|v\|_{\hat{V}_q},$$

so that $\|v\|_\infty \leq \|v\|_{\hat{V}_q}$.

Then $Dv(z) = WD\varphi(z)$ and by the law of large number we have for any $\theta \in \mathbb{S}^{q-1}$:

$$\begin{aligned}\|D\varphi(z)\theta\|^2 &= \frac{1}{r}\sum_{j=1}^{r}\sum_{1\leq k,l\leq q}\omega_k^j\omega_l^j\theta_k\theta_l \\ &= \frac{1}{r}\sum_{j=1}^{r}\langle\omega^j,\theta\rangle^2 \\ &\xrightarrow{r\to+\infty} \mathbb{E}_{\mu_q}\left[\langle\omega,\theta\rangle^2\right] = -k_q''(0).\end{aligned}$$

Because μ_q admits finite fourth order moments, the rate of convergence can be controlled using Chebyshev's inequality. For every indices $k, l \in [\![1, q]\!]$:

$$\mathbb{P}\Big(\left|\frac{1}{r}\sum_{j=1}^{r}\omega_k^j\omega_l^j - \mathbb{E}_{\mu_q}\left[\omega_k\omega_l\right]\right| \geq \alpha/q\Big) \leq \frac{q^2\mathbb{E}_{\mu_q}\left[\omega_k^2\omega_l^2\right]}{\alpha^2 r}.$$

For $r \geq \Omega(\frac{q^4\tau}{\alpha^2})$ we have with probability greater than $1-\tau^{-1}$ that the above inequality is satisfied for every indices k, l. Thus for every $z \in \mathbb{R}^q$ and every $\theta \in \mathbb{S}^{q-1}$:

$$\begin{aligned}&\left|\|D\varphi(z)\theta\|^2 + k_q''(0)\right| \\ &\leq \sum_{1\leq k,l\leq q}|\theta_k\theta_l|\Big|\frac{1}{r}\sum_{j=1}^{r}\omega_k^j\omega_l^j - \mathbb{E}_{\mu_q}\left[\omega_k\omega_l\right]\Big| \\ &\leq \sum_{1\leq k,l\leq q}|\theta_k\theta_l|\frac{\alpha}{q} \leq \alpha,\end{aligned}$$

using Chauchy-Schwarz inequality in the last line. We can thus conclude:

$$\|D\varphi\|_{2,\infty}^2 \leq -k_q''(0) + \alpha.$$

The same arguments holds for $D^2v(z) = WD^2\varphi(z)$. For any $\theta \in \mathbb{S}^{q-1}$ we have:

$$D^2\varphi(z)(\theta,\theta) = \left(\frac{1}{\sqrt{r}}\sum_{1\leq k,l\leq q} -e^{\imath\langle z,\omega^j\rangle}\omega_k^j\omega_l^j\theta_k\theta_l\right)_{1\leq j\leq r}.$$

Passing to the squared norm we get:

$$\begin{aligned}&\|D^2\varphi(z)(\theta,\theta)\|^2 \\ &\quad= \frac{1}{r}\sum_{j=1}^{r}\sum_{1\leq k,l,s,t\leq q}\omega_k^j\omega_l^j\omega_s^j\omega_t^j\theta_k\theta_l\theta_s\theta_t \\ &\quad\xrightarrow{r\to+\infty}\sum_{1\leq k,l,s,t\leq q}\mathbb{E}_{\mu_q}\left[\omega_k\omega_l\omega_s\omega_t\right]\theta_k\theta_l\theta_s\theta_t \\ &\quad= \mathbb{E}_{\mu_q}\left[\langle\omega,\theta\rangle^4\right] = k_q^{(4)}(0).\end{aligned}$$

Then because μ_q admits 8^{th} order moments, we can control the convergence in probability by Chebyshev's inequality. For $r \geq \Omega(\frac{q^8\tau}{\alpha^2})$ we have with probability greater than $1-\tau^{-1}$:

$$\|D^2\varphi\|_{2,\infty}^2 \leq k_q^{(4)}(0) + \alpha.$$

Finally $\hat{V}_q$ satisfies Assumption 1 with:

$$\hat{\kappa} \leq (k_q(0))^{1/2} + (-k_q''(0))^{1/2} + (k_q^{(4)}(0))^{1/2} + 1$$

for α sufficiently low. □

Note that the assumption of finite 8^{th} moments is only needed to have a control of the convergence rate of $\hat{k}_q$ towards k_q in probability. Following the proof and by the law of large numbers, assuming finite 4^{th}-order moments is sufficient to have convergence almost surely. Also, we used the Chebyshev's inequality in order to control the convergence rate. Making stronger assumptions on the decay of μ_q, such as subgaussianity for example, could have led to faster convergence by using sharper concentration inequalities.

C.3 Example of appropriate kernels

We show here that the Matérn kernel of parameter $\nu \in (8, +\infty]$ satisfies Assumption 3 and Assumption 4.

Gaussian kernel The Gaussian kernel defined by for some parameter $\sigma > 0$ by $k_q(r) = e^{-\frac{\sigma^2 r^2}{2}}$. In this case the frequency distribution μ_q is the multivariate normal of variance σ and has a density which is given for every $\omega \in \mathbb{R}^q$ by:

$$\mu_q(\omega) = \frac{1}{(2\pi\sigma^2)^{q/2}} e^{-\frac{\|\omega\|^2}{2\sigma^2}},$$

This distribution admits finite moments of every order which are independent of q. Also, k_q is four times differentiable at 0 and by Property 4 the associated V_q is (strongly) admissible with $\kappa = 2 + \sqrt{3}$

Moreover Assumption 3 as one has $|k_q(x)| \leq 1/2N$ if:

$$x \geq \beta_{q,N} = \frac{2}{\sigma^2}\sqrt{\log(2N)}.$$

Matérn kernel Sobolev spaces $H^s(\mathbb{R}^q, \mathbb{R}^q)$ which are RKHSs as soon as $s > q/2$. Given some $\nu > 0$, the kernel k_q associated to $H^{(q/2+\nu)}(\mathbb{R}^q, \mathbb{R}^q)$ is independent of q and is defined in Eq. (20). It is associated with the multivariate t-distribution:

$$\mu_q(\omega) = C(q,\nu)(1 + \frac{\|\omega\|^2}{2\nu})^{-(\nu+q/2)},$$

for some normalising constant $C(q,\nu)$. Therefore, μ_q admits l^{th} order moments as soon as $\nu \geq l/2$, and those moments are bounded independently of q (see [28] for the computation of moments). In particular, for $\nu > 2$, k_q is four times differentiable at 0 with $k''(0) = \nu/(\nu-1)$ and $k^{(4)}(0) = 3\nu^2/(\nu-1)(\nu-2)$. Thus by Property 4, V_q is (strongly) admissible with:

$$\kappa = 1 + \sqrt{\frac{\nu}{(\nu-1)}} + \sqrt{\frac{3\nu^2}{(\nu-1)(\nu-2)}}.$$

Because k_q has exponential decay (see for example [30]), there exist constants H_ν, G_ν such that:

$$|k_q(r)| \leq G_\nu e^{-H_\nu^{-1} r}$$

and Assumption 3 is satisfied with

$$\beta_{q,N} = H_\nu \log(2G_\nu N).$$