Weierstraß-Institut

für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Stability of deep neural networks via discrete rough paths

Christian Bayer[1], Peter K. Friz[2], Nikolas Tapia[2]

submitted: June 23, 2020

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: christian.bayer@wias-berlin.de

[2] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
and
Institut für Mathematik
Technische Universität Berlin
Str. des 17. Juni 136
10623 Berlin
Germany
E-Mail: peter.friz@wias-berlin.de
friz@math.tu-berlin.de
nikolasesteban.tapiamunoz@wias-berlin.de

No. 2732
Berlin 2020

2020 *Mathematics Subject Classification.* 60L10, 60L70, 60L90, 68T07.

Key words and phrases. Deep neural networks, iterated sums signature.

N.T. would like to thank K. Ebrahimi-Fard for many helpful discussions. The authors acknowledge financial support from the MATH+ EF1 Excellence Cluster.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Stability of deep neural networks via discrete rough paths

Christian Bayer, Peter K. Friz, Nikolas Tapia

Abstract

Using rough path techniques, we provide *a priori* estimates for the output of Deep Residual Neural Networks. In particular we derive stability bounds in terms of the total p-variation of trained weights for any $p \geq 1$.

Contents

DOI 10.20347/WIAS.PREPRINT.2732 Berlin 2020

1 Introduction

Since their introduction in 2016 [13], Residual Neural Networks (ResNets) have gained a vast amount of popularity as a preferred network architecture for Machine Learning applications. The general principle is that they allow for deeper networks since they model only the change in the output for each layer. This is achieved by introducing "skip connections" which – at some steps – adjust the output of a layer by adding an earlier layer's output (see Figure 1). The authors argue that this helps precondition the optimization solvers so that increasing the network depth does not result in severe numerical instabilities and performance degradation, as is observed in plain Neural Networks. In particular, this approach allows them to successfully train a Deep Neural Network with hundreds of layers.

In a plain Neural Network, the input vector $\mathbf{w}_{i+1}$ of the $(i+1)$-th hidden layer is given by an application of the weights and the activation function to the input of the previous hidden layer. In symbols

$$\mathbf{w}_{i+1} = \sigma(W_i \mathbf{w}_i)$$

where $\sigma : \mathbb{R}^{d_{i+1}} \to \mathbb{R}^{d_{i+1}}$ and W_i is a $d_{i+1} \times d_i$ matrix. In the ResNet approach, this is modified so that the output to the next hidden layer is given as the sum of the *input* to the previous layer, plus the previous operations; that is,

$$\mathbf{w}_{i+1} = \mathbf{w}_i + \sigma(W_i \mathbf{w}_i). \tag{1}$$

Here, it is assumed that the width of all layers is constant, but the approach can easily be adapted to the more familiar setting of varying widths by applying an appropriate projection to right-hand side of the last equation.

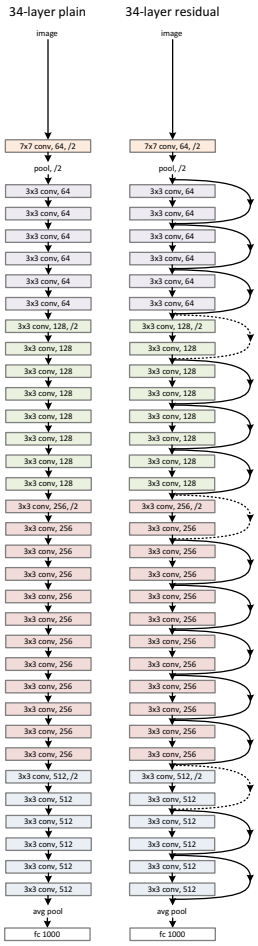


Figure 1: Example ResNet architecture, taken from [13].

Remark 1.1. We simplify notation by leaving out the bias term in the update rule (1). The usual update rule

$$\mathbf{w}_{i+1} = \sigma(W_i \mathbf{w}_i + \mu_i)$$

can be reproduced in the form (1) above by adding a column of consisting of ones to $\mathbf{w}_i$ and an appropriate restriction on W_i to map that column to another column of ones – in the appropriate dimension.

Remark 1.2. In this work, we assume that the architecture follows the update (1) at each layer. In the engineering practice, usually a few layers are skipped over. i.e. the true update may look as follows:

$$\widetilde{\mathbf{w}}_i = \sigma(\widetilde{W}_i \mathbf{w}_i), \quad \mathbf{w}_{i+1} = \mathbf{w}_i + \sigma(W_i \widetilde{\mathbf{w}}_i),$$

skipping over one layer in the process.

It has been pointed out by several authors [5, 10, 11] that the update in eq. (1) can be seen as a step of the Euler scheme for a *controlled ODE* of the form

$$\dot{\mathbf{w}}(t) = \sigma(W(t)\mathbf{w}(t)), \quad \mathbf{w}(0) = \mathbf{w}_0. \tag{2}$$

Therefore, knowledge of stability and convergence of numerical schemes for such systems can be used to derive corresponding results for ResNets, specially since one expects that the behavior of the output layer of the network under consideration will follow closely that of the continuous-time solution of eq. (2) for very deep architectures.

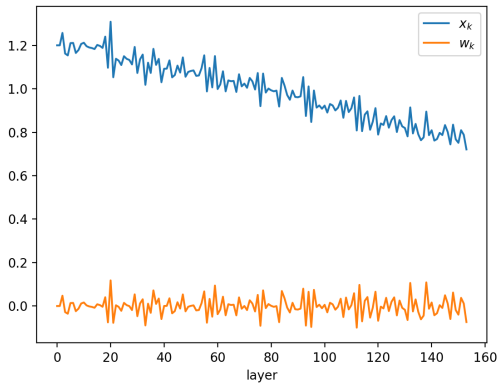
Instead of using continuous techniques, our approach consists of analyzing the evolution of the sequence $(x_0, \dots, x_N)$, where N is the depth of the network, obtained by iteration of eq. (1) directly at the discrete level. This is achieved by carefully estimating p-variation norms of this sequence using analytic techniques borrowed from rough paths theory and the algebraic framework developed in [4]. In particular, we show that for a regular enough activation function and any $p \geq 1$, there is an explicit constant $C_p > 0$ such that

$$|\mathbf{x}_N - \mathbf{x}_0| \leq \inf_{p \in [1,\infty)} \left(C_p^{p-1} \|\sigma\|^p_{C_b^{[p]+1}} |||\mathbb{W}|||^p_{p;[0,N]} \vee \|\sigma\|_{C_b^{[p]+1}} |||\mathbb{W}|||_{p;[0,N]} \right).$$

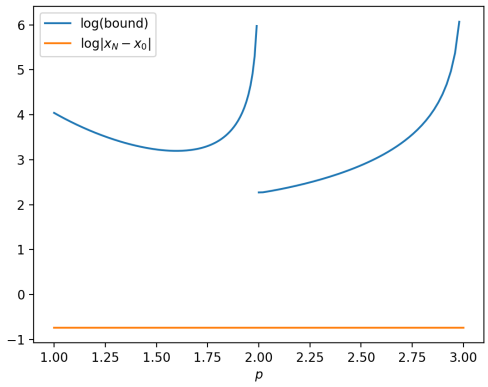
Here, $[p]$ denotes the integer part of p. The symbol $\mathbb{W}$ denotes the discrete signature lift of the weight sequence W and $|||\cdot|||_p$ is an appropriately defined norm on the spaces of lifts (Corollary 4.14). This inequality holds *uniformly* over input data. In practice, the weight matrices are randomly initialized with random i.i.d. values so typically the trained weights are also random. Our estimates hold pathwise, in the sense that the depend only on a single initialization of the weight matrices.

This result is a first step towards the understanding of finer properties of (1), and can be used as a stepping stone in order to prove results about worst-case behavior, i.e. controlling the size of $|\mathbf{w}_N - \tilde{\mathbf{w}}_N|$ in terms of $|\mathbf{w}_0 - \tilde{\mathbf{w}}_0|$.

To see how our resulting a priori estimate compares to what the smooth theory would imply, we ran a simple numerical experiment, using a pre-trained ResNet 152 of [13] obtained from `https://github.com/BVLC/caffe/wiki/Model-Zoo` (Figure 2b).



(a) Evolution of a single feature through a trained ResNet



(b) Bound in eq. (16) for different values of $p \in [1, 3]$ vs. the difference $|\mathbf{x}_N - \mathbf{x}_0|$

2 Elements of rough analysis

We begin with a brief overview of classical results present in the rough analysis literature. We remark that many of these results are usually stated in terms of continuous-time variables which introduces certain additional difficulties. In our case, no such difficulties arise so the statements and proofs of analogous results become simpler.

2.1 Discrete controls

We recall that in the setting of [15] a *control function* (or simply a *control*) is a function $\omega\colon [0,\infty)\times[0,\infty) \to [0,\infty)$ which is super-additive, in the sense that $\omega(s,u)+\omega(u,t) \le \omega(s,t)$ for all $s<u<t$. In the continuous-time setting, the main motivation for introducing control functions is to measure the size of the increments of a function in a more flexible way than what the natural control $\omega(s,t)=|t-s|$ allows.

Definition 2.1 ([3])**.** *A (discrete) control is a triangular array of non-negative numbers* $(\omega_{k,l} : k<l)$ *such that* $\omega_{k,k}=0$ *and*

$$\omega_{k,l}+\omega_{l,m} \le \omega_{k,m}$$

for all $k<l<m$

Remark 2.2. Observe that for a control ω the maps $l\mapsto\omega_{k,l}$ and $k\mapsto\omega_{k,l}$ are non-decreasing and non-increasing, respectively. Indeed, if $0\le k<l<m\le N$ then

$$\omega_{k,l} \le \omega_{k,l}+\omega_{l,m} \le \omega_{k,m}$$

and

$$\omega_{k,m} \ge \omega_{k,l}+\omega_{l,m} \ge \omega_{l,m}.$$

Now we collect some results on how to produce new controls out of any given control.

Lemma 2.3. *Let* w *be a control and* $\varphi\colon [0,\infty)\to[0,\infty)$ *an increasing convex function such that* $\varphi(0)=0$*. Then* $\tilde{w}_{k,l}:=\varphi(\omega_{k,l})$ *is also a control.*

Proof. Since φ is convex and $\varphi(0)=0$ we have that

$$\varphi(\lambda(x+y)) \le \lambda\varphi(x+y)$$

for any $\lambda\in[0,1]$. Choosing $\lambda=\frac{x}{x+y}$ we obtain

$$\varphi(x) \le \frac{x}{x+y}\varphi(x+y).$$

Similarly, $\varphi(y)\le\frac{y}{x+y}\varphi(x+y)$ so that

$$\varphi(x)+\varphi(y)\le\varphi(x+y),$$

i.e. φ is super-additive.

Therefore, if $0\le k<l<m\le N$,

$$\begin{aligned}\tilde{w}_{k,l}+\tilde{w}_{l,m} &= \varphi(\omega_{k,l})+\varphi(\omega_{l,m})\\ &\le \varphi(\omega_{k,l}+\omega_{l,m})\\ &\le \varphi(\omega_{l,m})=\tilde{\omega}_{l,m}\end{aligned}$$

where the last inequality follows from the monotonicity of φ. □

Remark 2.4. In particular, this implies that if ω is a control, then ω^α is also a control, for any $\alpha>1$.

Lemma 2.5. *Let $\omega, \tilde{\omega}$ be two controls. If $\alpha, \beta > 0$ are such that $\alpha + \beta \geq 1$, then $\hat{\omega}_{k,l} := \omega_{k,l}^{\alpha}\tilde{\omega}_{k,l}^{\beta}$ is also a control.*

Proof. Let $\theta := \alpha + \beta$. By Lemma 2.3, it is enough to show that

$$z_{k,l} := \omega_{k,l}^{\frac{\alpha}{\theta}}\tilde{w}_{k,l}^{\frac{\beta}{\theta}}$$

is a control, since then $\hat{w}_{k,l} = z_{k,l}^{\theta}$ will also be a control. Since $\frac{\alpha}{\theta} + \frac{\beta}{\theta} = 1$, Hölder's inequality implies that

$$\begin{aligned} z_{k,l} + z_{l,m} &\leq (\omega_{k,l} + \omega_{l,m})^{\frac{\alpha}{\theta}}(\tilde{w}_{k,l} + \tilde{w}_{l,m})^{\frac{\beta}{\theta}} \\ &\leq \omega_{k,m}^{\frac{\alpha}{\theta}}\tilde{w}_{k,m}^{\frac{\beta}{\theta}} \end{aligned}$$

and the proof is finished. □

2.2 p-variation

In the following we will deal with *time series*, which are finite sequences of vectors $\mathbf{w} = (\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_N) \in (\mathbb{R}^d)^N$. We will use the convention of indexing time steps with lower indices and components with upper indices, so for example $\mathbf{w}_k^i \in \mathbb{R}$ refers to the i-th component of the k-th entry in the time series $\mathbf{w}$.

We will also need to deal with general *triangular arrays*, which are collections of vectors of the form $(\Xi_{k,l} : 0 \leq k < l \leq N)$. For any time series we define a triangular array $(\mathbf{w}_{k,l})$ by setting $\mathbf{w}_{k,l} := \mathbf{w}_l - \mathbf{w}_k$.

Definition 2.6. *Given $p > 0$, we define the p-variation with respect to a fixed choice of norm $|\cdot|$ on $\mathbb{R}^d$, by*

$$\|\mathbf{w}\|_{p;[k,l]} := \left(\max_{s\in\mathcal{S}_{k,l}} \sum_{j=0}^{\#s} |\mathbf{w}_{s_{j+1}} - \mathbf{w}_{s_j}|^p\right)^{1/p}$$

where the maximum is taken over the set $\mathcal{S}_{k,l}$ of all increasing subsequences

$$s = (s_0 = k, s_1, \ldots, s_m, s_{m+1} = l)$$

of $\{k, k+1, \ldots, l-1, l\}$ and we have set $\#s = m$ for such a sequence. For a triangular array Ξ one can also define its p-variation as

$$\|\Xi\|_{p;[k,l]} := \left(\sup_{s\in\mathcal{S}_{k,l}} \sum_{j=0}^{\#s} |\Xi_{s_j,s_{j+1}}|^p\right)^{1/p}.$$

We observe that in the case where $\Xi_{k,l} = \mathbf{w}_l - \mathbf{w}_k$ both definitions coincide.

Since the trivial sequence $(k, l) \in \mathcal{S}_{k,l}$ we obtain immediately the bound

$$|\Xi_{k,l}| \leq \|\Xi\|_{p;[k,l]} \tag{3}$$

for any $p > 0$. In the particular case where $\Xi_{k,l} = \mathbf{w}_l - \mathbf{w}_k$ we also obtain

$$\|\mathbf{w}\|_{\infty} := \sup_{k=0,\ldots,N} |\mathbf{w}_k| \leq |\mathbf{w}_0| + \|\mathbf{w}\|_{p;[0,N]}.$$

Proposition 2.7. *Let Ξ be a triangular array and $p \geq 0$. Then $\omega_{k,l} := \|\Xi\|^p_{p;[k,l]}$ is a control.*

Proof. Indeed, if $s' \in \mathcal{S}_{k,l}$ and $s'' \in \mathcal{S}_{l,m}$ then $s = (s', s'') \in \mathcal{S}_{k,m}$ and so

$$\sum_{j=0}^{\#s'} |\Xi_{s_j,s_{j+1}}|^p + \sum_{j'=0}^{\#s''} |\Xi_{s'_{j'},s'_{j'+1}}|^p \leq \|\Xi\|^p_{p;k,m}$$

and super-additivity follows from taking the supremum over $\mathcal{S}_{k,l}$ and $\mathcal{S}_{l,m}$. □

Remark 2.8. Since the set $\mathcal{S}_{k,l}$ is finite, the p-variation norm of Ξ is finite for any $p > 0$ and triangular array Ξ. This should be contrasted with the usual setting for rough paths, where one deals with paths in continuous time; in that setting, the p-variation norm can become infinite and this introduces a number of analytical problems which are not present in the present context.

Remark 2.9. The p-variation defines a quasi-norm for $0 < p < 1$ (i.e. the triangle inequality fails), and a semi-norm for $p \geq 1$ on time series, since all constant sequences have vanishing p-variation. For $p \geq 1$, it becomes a norm on triangular arrays.

Lemma 2.10. *Let $0 \leq p < q < \infty$. Then $\|\Xi\|_{q;[k,l]} \leq \|\Xi\|_{p;[k,l]}$*

Proof. Observe that, since $\frac{q}{p} > 1$, the inequality

$$\sum_{j=0}^{\#s} |\Xi_{s_j,s_{j+1}}|^q \leq \left(\sum_{j=0}^{\#s} |\Xi_{s_j,s_{j+1}}|^p \right)^{q/p}$$

holds for any $s \in \mathcal{S}_{k,l}$. □

Given a triangular array Ξ, we define another collection $(\delta\Xi_{k,l,m} : 0 \leq k < l < m)$ by

$$\delta\Xi_{k,l,m} := \Xi_{k,m} - \Xi_{k,l} - \Xi_{l,m}.$$

In the special case where $\Xi_{k,l} = \mathbf{w}_l - \mathbf{w}_k$ we see that $\delta\Xi_{k,l,m} = 0$. The operator δ satisfies the following product rule: if $\mathbf{w}$ is a time series and Ξ is a triangular array, consider the triangular array $\mathbf{Z}_{k,l} := \mathbf{w}_k \Xi_{k,l}$. Then

$$\delta\mathbf{Z}_{k,l,m} = \mathbf{w}_k \delta\Xi_{k,l,m} - \mathbf{w}_{k,l}\Xi_{l,m}. \tag{4}$$

Finally we collect here some standard results for further reference.

Lemma 2.11. *Let Ξ be a triangular array and $p \geq 0$. Suppose there is a control w such that*

$$|\Xi_{k,l}| \leq C\omega_{k,l}^{1/p}$$

for all $0 \leq k < l \leq N$ and some constant $C > 0$. Then,

$$\|\Xi\|_{p;[k,l]} \leq C\omega_{k,l}^{1/p}$$

for all $0 \leq k < l \leq N$.

Proof. By hypothesis the inequality

$$|\Xi_{k,l}|^p \leq C^p \omega_{k,l}$$

holds for all $0 \leq k < l \leq N$. By superadditivity of w, if $s \in \mathcal{S}_{k,l}$ then also

$$\sum_{j=0}^{\#s} |\Xi_{s_j,s_{j+1}}|^p \leq C^p \omega_{k,l}.$$

The desired bound follows upon taking the maximum over $s \in \mathcal{S}_{k,l}$. □

Lemma 2.12. *Assume that $p \geq 1$ and*

$$|\Xi_{k,l}| \leq C\omega_{k,l}^{1/p}$$

for all $0 \leq k < l$ such that $\omega_{k,l} \leq 1$. Then

$$\|\Xi\|_{p;[k,l]} \leq C(\omega_{k,l}^{1/p} \vee \omega_{k,l})$$

for all $0 \leq k < l$.

2.3 The Sewing Lemma

At the core of the theory of rough integration lies the Sewing Lemma [7, 8]. Therefore, it is tightly connected with the solution theory of differential equations driven by rough signals. Since our main aim is to perform a precise analysis of the behaviour of discrete equations driven by irregular time-series, it is no doubt that its discrete analogue will play a prominent rôle here as well.

We begin by showing some preliminary results.

Lemma 2.13. *Suppose $s \in \mathcal{S}_{k,l}$ of length $\#s = m$. For any given control w, there exists an integer j^* with $1 \leq j^* \leq m$ such that*

$$\omega_{s_{j^*-1},s_{j^*+1}} \leq \frac{2}{m}\omega_{k,l}.$$

Proof. Suppose, on the contrary, that for any $1 \leq j \leq m$ we have that

$$\omega_{s_{j-1},s_{j+1}} > \frac{2}{m}\omega_{k,l}.$$

Then this would imply that

$$2\omega_{k,l} < \sum_{j=1}^{m} \omega_{s_{j-1},s_{j+1}} \leq 2\omega_{k,l}$$

by super-additivity, which is a contradiction. □

Proposition 2.14 (Discrete sewing)**.** *Let $(\Xi_{k,l})$ be a triangular array, and suppose that there exist two controls w and $\tilde{w}$ such that*

$$|\delta\Xi_{k,l,m}| \leq \omega_{k,l}^{\alpha}\tilde{\omega}_{l,m}^{\beta}$$

for some $\alpha, \beta > 0$ with $\alpha + \beta > 1$. Then

$$\left|\sum_{j=k}^{l-1} \Xi_{j,j+1} - \Xi_{k,l}\right| \leq 2^{(\alpha+\beta)}\zeta(\alpha+\beta)\omega_{k,l}^{\alpha}\tilde{w}_{k,l}^{\beta}$$

where ζ denotes Riemann's zeta function.

Proof. By Remark 2.2 we deduce that $|\delta\Xi_{k,l,m}| \leq \omega_{k,m}^{\alpha}\tilde{\omega}_{k,m}^{\beta}$, and Lemma 2.5 implies that $\hat{\omega} := \omega^{\frac{\alpha}{\theta}}\tilde{\omega}^{\frac{\beta}{\theta}}$ is a control.

Now we apply a Young-style argument to estimate the above difference. First we observe that if $l-k = 1$ then the bound is trivial since the left-hand side vanishes. Therefore we assume that $l-k \geq 2$. By Lemma 2.13 we can find an index $k < j^* < l$ such that

$$\hat{\omega}_{j^*-1,j^*+1} \leq \frac{2}{(l-k-1)}\hat{\omega}_{k,l}.$$

Hence, if we denote by $s := (k, k+1, \dots, j^*-1, j^*+1, \dots, l)$ we have

$$\left|\sum_{j=k}^{l-1}\Xi_{j,j+1} - \sum_{s}\Xi_{s_j,s_{j+1}}\right| = |\delta\Xi_{j^*-1,j^*,j^*+1}| \leq \left(\frac{2}{l-k-1}\right)^{\theta}\hat{\omega}_{k,l}^{\theta}.$$

Then we can apply Lemma 2.13 again to the sequence s to obtain a "coarser" sequence s', containing one less point, and such that

$$\left|\sum_{s}\Xi_{s_j,s_{j+1}} - \sum_{s'}\Xi_{s'_j,s'_{j+1}}\right| \leq \left(\frac{2}{l-k-2}\right)^{\theta}\hat{\omega}_{k,l}^{\theta}.$$

Continuing in this way we obtain a sequence of coarsenings of the full sequence until we get to $s^* = (k,l)$, and by using the triangular inequality we then deduce the estimate

$$\left|\sum_{j=k}^{l-1}\Xi_{j,j+1} - \Xi_{k,l}\right| \leq 2^{\theta}\sum_{r=1}^{l-k-1}\frac{1}{r^{\theta}}\hat{\omega}_{k,l}^{\theta}$$

from where the conclusion follows since $\theta = \alpha + \beta > 1$. □

We will also need the following generalization of the Sewing Lemma, whose proof is straightforward.

Proposition 2.15 (Generalized discrete sewing)**.** *Suppose that Ξ is a triangular array as before. Suppose that there are controls ω_r and $\tilde{\omega}_r$, and exponents $\alpha_r, \beta_r > 0$ such that $\alpha_r + \beta_r > 1$ for all $r = 1, \dots, n$. If*

$$|\delta\Xi_{k,l,m}| \leq \sum_{r=1}^{n}\omega_{r;k,l}^{\alpha_r}\tilde{\omega}_{r;l,m}^{\beta_r}$$

then

$$\left|\sum_{j=k}^{l-1}\Xi_{j,j+1} - \Xi_{k,l}\right| \leq 2^{\hat{\theta}}\zeta(\hat{\theta})\sum_{r=1}^{n}\omega_{r;k,l}^{\alpha_r}\tilde{\omega}_{r;k,l}^{\beta_r}$$

where $\hat{\theta} := \min_{r=1,\dots,n}\{\alpha_r + \beta_r\}$.

3 The iterated-sums signature

3.1 Quasi-shuffle Hopf algebra

Consider an at most countably infinite set $\mathfrak{A}$, hereafter called the *alphabet*, and whose elements we shall call *letters*. Given $k \geq 1$, a *word of length k over* $\mathfrak{A}$ is a sequence $u = (u_1, \dots, u_k) \in \mathfrak{A}^k$; for

convenience we use the notation $u = u_1 \cdots u_k$ and we denote its length by $\ell(w) := k$. Note that the order of the letters is crucial, so the words a_1a_2 and a_2a_1 are distinct if $a_1 \neq a_2$. There is a single word of length 0, called the *empty word*, and denoted by e. Moreover, we make the convention that $\mathfrak{A}^0 := \{e\}$. The collection of all words over $\mathfrak{A}$ is denoted by

$$\mathfrak{A}^* := \bigcup_{k=0}^{\infty} \mathfrak{A}^k.$$

There is a monoid structure on $\mathfrak{A}^*$ obtained by concatenation of words. Given two words $u = u_1 \cdots u_k \in \mathfrak{A}^k$ and $v = v_1 \cdots v_\ell \in \mathfrak{A}^\ell$, their concatenation is the word

$$uv = u_1 \cdots u_k v_1 \cdots v_\ell \in \mathfrak{A}^{k+\ell}.$$

By definition $ew = we = w$ so that the empty word acts as the neutral element for this composition.

The following construction, which already appeared in [4], is a particular case of the general quasi-shuffle product introduced in [14]. Consider a finite set $A = \{1, \dots, \mathrm{d}\}$ of d distinct symbols. We complete A to a commutative semigroup $\mathfrak{A}$, whose internal law we denote by square brackets; therefore, we obtain a map $[\cdot\cdot] : \mathfrak{A} \times \mathfrak{A} \mapsto \mathfrak{A}$ such that

$$[a_1[a_2a_3]] = [[a_1a_2]a_3], \qquad [a_1a_2] = [a_2a_1]$$

for all $a_1, a_2, a_3 \in \mathfrak{A}$. In view of the first identity, we denote the common result of this operation just by $[a_1a_2a_3]$. Therefore, any element $a \in \mathfrak{A}$ is of the form $a = [\mathrm{i}_1 \cdots \mathrm{i}_k]$ for some $\mathrm{i}_1, \dots, \mathrm{i}_k \in S$. Observe that a word $w \in \mathfrak{A}^*$ has a lenght $\ell(w)$ as before, but now also a *weight* $|w|$ which counts the total number of symbols from S forming it. For example, the word $w = [13][23]$ has length $\ell(w) = 2$ but weight $|w| = 4$.

Now, we let H be the real vector space spanned by $\mathfrak{A}^*$. A generic element of H is a finite linear combination of words from $\mathfrak{A}^*$ with real coefficients.

Example 3.1. If $A = \{1, 2, 3\}$, a generic element of H might look like

$$\sqrt{2}[1] + \frac{3}{2}[112] + \pi^2[1][23].$$

We consider the *quasi-shuffle* product $\star \colon H \times H \to H$ recursively by $w \star e = e \star w = e$ and

$$va \star wb := (v \star wb)a + (va \star w)b + (v \star w)[ab]$$

for $v, w \in \mathfrak{A}^*$ and $a, b \in \mathfrak{A}$. This definition is bilinearly extended to $H \otimes H$. An example:

$$\begin{aligned}[12] \star [3] &= (e \star [3])[12] + ([12] \star e)[3] + (e \star e)[123] \\ &= [3][12] + [12][3] + [123].\end{aligned}$$

We observe in particular that the total weight $|[12]| + |[3]| = 3$ is preserverd but the total length is not, in the sense that all the terms in the right-hand side have the same weight but not the same lenght. In general, it can be proven that if H_n is the real vector space spanned by words of weight exactly n, then the inclusion $H_n \star H_m \subset H_{n+m}$ holds [14]. Since the decomposition

$$H = \bigoplus_{n=0}^{\infty} H_n$$

is clearly true, $(H, \star, e)$ becomes a graded connected algebra, known as the *quasi-shuffle algebra over S*.

The space H can be endowed with another operation, known as the *deconcatenation coproduct* $\Delta\colon H \to H \otimes H$, and defined for $u = u_1 \cdots u_n$ by

$$\Delta(w) := w \otimes e + e \otimes w + \sum_{j=1}^{n-1} u_1 \cdots u_j \otimes u_{j+1} \cdots u_n.$$

This definition is linearly extended to all of H. It can be shown that the compatibility condition $\Delta(v \star w) = \Delta(v) \star \Delta(w)$ is satisfied, thus turning the triple $(H, \star, \Delta)$ into a Hopf algebra.

The space of linear maps $\psi : H \to \mathbb{R}$ can be turned into an algebra by dualizing the coproduct. More precisely, given two such maps φ, ψ, we define

$$(\varphi * \psi)(x) = (\varphi \otimes \psi) \circ \Delta(x), \quad x \in H.$$

For a single word $u = u_1 \cdots u_n$ this yields

$$(\varphi * \psi)(w) = \varphi(w)\psi(e) + \varphi(e)\psi(w) + \sum_{j=1}^{n-1} \varphi(\omega_1 \cdots \omega_j)\psi(\omega_{j+1} \cdots \omega_n). \tag{5}$$

It is a general result that the set G of linear maps such that $\psi(v \star w) = \psi(v)\psi(w)$ forms a group under the convolution product; that is, if $\varphi, \psi \in G$ then $\varphi * \psi \in G$, and for all $\psi \in G$ there exists $\psi^{-1} \in G$ such that $\psi * \psi^{-1} = \psi^{-1} * \psi = \varepsilon$. Here, $\varepsilon \in G$ is the multiplicative linear map such $\varepsilon(e) = 1$ and zero otherwise. The map ε is called the *counit* of H and acts as the neutral element for the group law in G. Elements of G are referred to as *characters over H* (or simply as *characters*) if the underlying Hopf algebra is clear from context.

We now extend a given time series $\mathbf{w}$ by computing additional "features", which we will now describe. These features provide a succint description of the behaviour of $\mathbf{w}$, in the spirit of T. Lyons' iterated-integrals signature. In [4] it is shown that these features capture all time-warping invariants of $\mathbf{w}$. In the remainder of the paper we will see that they are also well suited for giving an analytical description of the behaviour of ResNets, and more generally, of numerical schemes for approximation solutions to ODEs driven by rough signals.

Recall from the previous section that H is defined to be the real-linear span of the set of words $\mathfrak{A}^*$ over the free commutative semigroup $\mathfrak{A}$ generated by $A = \{1, \dots, d\}$. Given two indices $0 \le k < l \le N$ we define a linear map $\mathbb{W}_{k,l} : H \to \mathbb{R}$ in three steps:

1 Define the *extended increments* $\mathbf{w}_{k,l}^{[i_1 \cdots i_n]}$ of $\mathbf{w}$ by simply multiplying out the individual increments in the corresponding directions, that is, for each n-tuple $(i_1, \dots, i_n) \in A^n$,

$$\mathbf{w}_{k,l}^{[i_1 \cdots i_n]} := \mathbf{w}_{k,l}^{i_1} \cdots \mathbf{w}_{k,l}^{i_n}.$$

2 Each word $w = \omega_1 \cdots \omega_n \in \mathfrak{A}^*$ is mapped to an iterated sum,

$$\mathbb{W}_{k,l}^{\omega} := \sum_{k \le j_1 < \cdots < j_n < l} \mathbf{w}_{j_1, j_1+1}^{\omega_1} \cdots \mathbf{w}_{j_n, j_n+1}^{\omega_n}.$$

By definition, the empty word is mapped always to 1.

3 Finally, the map $w \mapsto \mathbb{W}^{\omega}_{k,l}$ is linearly extended to H.

Definition 3.2. *The collection* $\mathbb{W} := (\mathbb{W}_{k,l} : 0 \le k < l \le N)$ *is the* iterated-sums signature *of the time series* $\mathbf{w}$.

Example 3.3. Before moving forward, we present some examples.

$$\mathbb{W}^{[i_1][i_2]}_{k,l} = \sum_{j=k}^{l-1} (\mathbf{w}^{i_1}_j - \mathbf{w}^{i_1}_k)(\mathbf{w}^{i_2}_{j+1} - \mathbf{w}^{i_2}_j)$$
$$\mathbb{W}^{[i_1 i_2]}_{k,l} = \sum_{j=k}^{l-1} (\mathbf{w}^{i_1}_{j+1} - \mathbf{w}^{i_1}_j)(\mathbf{w}^{i_2}_{j+1} - \mathbf{w}^{i_2}_j).$$

We remark from this example that the following relation

$$\begin{aligned}
\mathbb{W}^{[i_1]}_{k,l}\mathbb{W}^{[i_2]}_{k,l} &= \mathbf{w}^{i_1}_{k,l}\mathbf{w}^{i_2}_{k,l} \\
&= \sum_{j_1=k}^{l-1} (\mathbf{w}^{i_1}_{j_1+1} - \mathbf{w}^{i_1}_{j_1}) \sum_{j_2=k}^{l-1} (\mathbf{w}^{i_2}_{j_2+1} - \mathbf{w}^{i_2}_{j_2}) \\
&= \sum_{k \le j_1 < j_2 < l} \mathbf{w}^{i_1}_{j_1,j_1+1}\mathbf{w}^{i_2}_{j_2,j_2+1} + \sum_{k \le j_2 < j_1 < l} \mathbf{w}^{i_1}_{j_1,j_1+1}\mathbf{w}^{i_2}_{j_2,j_2+1} + \sum_{j=k}^{l-1} \mathbf{w}^{i_1}_{j,j+1}\mathbf{w}^{i_2}_{j,j+1} \\
&= \mathbb{W}^{[i_1][i_2]}_{k,l} + \mathbb{W}^{[i_2][i_1]}_{k,l} + \mathbb{W}^{[i_1 i_2]}_{k,l} \\
&= \mathbb{W}^{[i_1]\star[i_2]}_{k,l}.
\end{aligned}$$

From this we can observe two things: the entries in the iterated-sums signature are **not** linearly independent from each other, and the quasi-shuffle product defined in Section 3.1 completely describes the combinatorial properties of the relations between them. The map $\mathbb{W}$ enjoys several other combinatorial properties which are nicely described by the quasi-shuffle Hopf algebra, which we now recall from [4].

Theorem 3.4. *The iterated-sums signature satisfies:*

1 *the quasi-shuffle identities: for any* $v, w \in H$,

$$\mathbb{W}^{v}_{k,l}\mathbb{W}^{\omega}_{k,l} = \mathbb{W}^{v\star w}_{k,l}.$$

This means that $\mathbb{W}_{k,l} \in G$ *for all* $k < l$.

2 *Chen's identity: for any* $k < l < m$ *we have* $\mathbb{W}_{k,l} * \mathbb{W}_{l,m} = \mathbb{W}_{k,m}$.

3 *Recursive computation: for any* $u = u_1 \cdots u_n$ *we have*

$$\mathbb{W}^{u}_{k,l} = \sum_{j=k}^{l-1} \mathbb{W}^{u_1 \cdots u_{n-1}}_{k,j} \mathbf{w}^{u_n}_{j,j+1}.$$

Remark 3.5. From Chen's identity it can be deduced that $\mathbb{W}_{k,l} = \mathbb{W}^{-1}_{0,k} * \mathbb{W}_{0,l}$. Therefore, the iterated-sums signature is characterized by the sequence $k \mapsto \mathbb{W}_{0,k}$. This identity can also be exploited during numerical computations to reduce the amount of work required to compute $\mathbb{W}_{k,l}$ for $0 < k < l < N$.

Indeed, one only really ever needs to compute iterated sums while calculating the value of $\mathbb{W}^u_{0,k}$ for all words u up to a certain weight, and all $0 \le k \le N$. Once these values are stored, the inverse $\mathbb{W}^{-1}_{k,l}$ can be computed by using the *antipode*[1], i.e. $(\mathbb{W}^{-1})^u_{k,l} = \mathbb{W}^{\alpha(u)}_{k,l}$ which only needs to compute linear combinations of the already stored quantities; these can also be stored. Finally, the value of $\mathbb{W}^u_{k,l}$ can easily be recovered from Chen's identity, since by eq. (5), $\mathbb{W}_{k,l} = \mathbb{W}^{-1}_{0,k} * \mathbb{W}_{0,l}$ is also given as a linear combination of already stored quantities.

The number of features contained in the iterated-sums signature, that is, the dimension of the linear span $H_n := \langle u \in \mathfrak{A}^* : |u| = n \rangle$ or, equivalently, the number of words $u \in \mathfrak{A}^*$ with $|u| = n$ is known [4, Remark 2.3]. The first few are

$$\dim H_1 = d, \quad \dim H_2 = \frac{d(3d+1)}{2}, \quad \dim H_3 = \frac{d(13d^2+9d+2)}{6}, \dots$$

We choose norms $|\cdot|_n$ on each of the finite-dimensional spaces H_n, subject to some compatibility conditions, and define the vector $\mathbb{W}^{(n)} = (\mathbb{W}^u : |u| = n)$. Finaly, we set

$$|||\mathbb{W}|||_{p;[k,l]} := \max_{n=1,\dots,[p]} \|\mathbb{W}^{(n)}\|^{1/n}_{p/n;[k,l]}$$

where the p-variation of $\mathbb{W}^{(n)}$ is computed with respect to the chosen norm on H_n. The trivial inequality

$$\|\mathbb{W}^u\|_{p/n;[k,l]} \le |||\mathbb{W}|||^n_{p;[k,l]}$$

can easily be seen to hold for all words $u \in H_n$.

4 Controlled difference equations

In this section we consider equations of the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \sum_{i=1}^{d} f_i(\mathbf{x}_k)(\mathbf{w}^i_{k+1} - \mathbf{w}^i_k), \quad \mathbf{x}_0 = \xi \in \mathbb{R}^m \tag{6}$$

for some vector fields $f_1, \dots, f_d$ on $\mathbb{R}^m$, and where k ranges between 0 and some fixed time horizon $N \in \mathbb{N}$. Our main aim is to obtain some control over the size of the end-point value $\mathbf{x}_N$ of the solution.

In view of the previous sections, and in particular of the bound in eq. (3), we will try to obtain good estimates for the p-variation norm $\|\mathbf{x}\|_{p;[0,N]}$. Of course, such estimates will require some assumptions on the vector fields. It turns out that we will not only be able to control the "large scale" behavior of $\mathbf{x}$, but we will also obtain a cascade of estimates of some remainder terms, reminiscent of a Taylor expansion.

The techniques needed to obtain those bounds will depend crucially on $p \in [1, \infty)$. At first, we distinguish two basic regimes: $p \in [1, 2)$ and $p \in [2, \infty)$. By analogy with the rough paths literature, we call the former the *Young regime*, and the latter the *rough regime* – even though there is strictly no notion of roughness in our setting. The rough regime can be further subdivided into the cases where

[1] This is a standard result in the theory of Hopf algebras. In the quasi-shuffle setting, $\alpha : H \to H$ admits an explicit expression in terms of compositions [14, 4]

$p \in [n, n+1)$, which we call the *level n rough regime*. The terminology will make itself clear later down the road.

A central tool for constructing solutions to ODEs driven by rough paths are the so-called *controlled paths*, introduced by Gubinelli [8]. See also [12]. In a nutshell, the notion of "controlledness" contains all the necessary analytical estimates needed for the definition of a rough integral which then is used to give sense to solutions of Rough Differential Equations. In the present setting no such definition is needed since there are no divergences appearing from considering eq. (6). Nonetheless, we can still derive similar bounds. Note however that in our case the estimates are *proven* rather than *assumed*.

Given a vector field $f : \mathbb{R}^m \to \mathbb{R}^m$ of class C_b^n, i.e. it and all its derivatives up to order n are bounded, we define

$$\|f\|_{C_b^n} := \max_{k=1,\dots,n} \|D^k f\|_\infty.$$

If $F = (f_1, \dots, f_d)$ is a collection of vector fields on $\mathbb{R}^n$ of class C_b^n (or, equivalently, a map in $C_b^n(\mathbb{R}^n, \mathbb{R}^{dn})$), we define

$$\|F\|_{C_b^n} := \max_{i=1,\dots,d} \|f_i\|_{C_b^n}.$$

4.1 The Young regime

In this regime, we can easily obtain good bounds with minimal assumptions on the f_i. These bounds have already been shown by Davie [3], but it will be an enlightening exercise to go through the proof in full details, since it will lay the foundations for our approach in the rough regime. Also, our methods are slightly different and already in this case they highlight the importance of the rôle played by the Sewing Lemma (Propositions 2.14 and 2.15).

Before beginning we define the remainder

$$R_{k,l} := \mathbf{x}_{k,l} - \sum_{i=1}^{d} f_i(\mathbf{x}_k)\mathbf{w}^i_{k,l} \tag{7}$$

so that

$$\mathbf{x}_{k,l} = \sum_{i=1}^{d} f_i(\mathbf{x}_k)\mathbf{w}^i_{k,l} + R_{k,l}.$$

Theorem 4.1. *Let $1 \le p < 2$, and suppose that $F = (f_1, \dots, f_d)$ is a collection of vector fields in $\mathbb{R}^n$, of class C_b^1. The bound*

$$\|\mathbf{x}\|_{p;[k,l]} \le \left(2^p C_p^{p-1} \|F\|^p_{C_b^1} \|\mathbf{w}\|^p_{p;[k,l]} \vee 2\|F\|_{C_b^1} \|\mathbf{w}\|_{p;[k,l]}\right)$$

holds, with

$$C_p := 2^{2/p}\zeta(2/p).$$

Proof. Consider the triangular array $\Xi_{k,l} := \sum_i f_i(\mathbf{x}_k)\mathbf{w}^i_{k,l}$. By eq. (4) we immediately see that

$$\delta\Xi_{k,l,m} = -\sum_i (f_i(\mathbf{x}_l) - f_i(\mathbf{x}_k))\mathbf{w}^i_{l,m},$$

so that the usual Lipschitz bound implies

$$|\delta\Xi_{k,l,m}| \le \|F\|_{C_b^1}\|\mathbf{x}\|_{p;[k,l]}\|\mathbf{w}\|_{p;[k,l]},$$

and the hypothesis of Proposition 2.14 is satisfied since $2/p > 1$. Thus, we obtain

$$\left|\sum_{j=k}^{l-1}\Xi_{j,j+1} - \Xi_{k,l}\right| \le C_p\|F\|_{C_b^1}\|\mathbf{x}\|_{p;[k,l]}\|\mathbf{w}\|_{p;[k,l]}.$$

with $C_p := 2^{2/p}\zeta(2/p)$. Now, we observe that by eq. (6),

$$\sum_{j=k}^{l-1}\Xi_{j,j+1} = \mathbf{x}_{k,l}$$

thus obtaining

$$|R_{k,l}| \le C_p\|F\|_{C_b^1}\|\mathbf{x}\|_{p;[k,l]}\|\mathbf{w}\|_{p;[k,l]}. \tag{8}$$

By Lemma 2.11, the same bound holds if we replace $|R_{k,l}|$ on the left-hand side by $\|R\|_{p/2;[k,l]}$.

Using the relation between the remainder R and the increments of $\mathbf{x}$ we get

$$|\mathbf{x}_{k,l}| \le C_p\|F\|_{C_b^1}\|\mathbf{x}\|_{p;[k,l]}\|\mathbf{w}\|_{p;[k,l]} + \|F\|_{C_b^1}\|\mathbf{w}\|_{p;[k,l]}$$

for all $0 \le l < k \le N$. We deduce that

$$\|\mathbf{x}\|^p_{p;[k,l]} \le 2^{p-1}C_p\|F\|^p_{C_b^1}\|\mathbf{x}\|^p_{p;[k,l]}\|\mathbf{w}\|^p_{p;[k,l]} + 2^{p-1}\|F\|^p_{C_b^1}\|\mathbf{w}\|^p_{p;[k,l]}.$$

If we now consider a pair $k < l$ such that $\bar\omega_{k,l}^{1/p} := 2C_p\|F\|_{C_b^1}\|\mathbf{w}\|_{p;[k,l]} \le 1$, we obtain

$$\|\mathbf{x}\|^p_{p;[k,l]} \le 2^p\|F\|^p_{C^1}\|\mathbf{w}\|^p_{p;[k,l]} = C_p^{-p}\bar\omega_{k,l}$$

for all such (k,l). In particular

$$|\mathbf{x}_{k,l}| \le 2\|F\|_{C_b^1}C_p^{-1}\bar\omega_{k,l}^{1/p}.$$

From Lemma 2.12 we then get

$$\begin{aligned}\|\mathbf{x}\|_{p;[k,l]} &\le C_p^{-1}\left(\bar\omega_{k,l} \vee \bar\omega_{k,l}^{1/p}\right)\\ &= C_p^{-1}\left(2^p\|F\|^p_{C_b^1}C_p^p\|\mathbf{w}\|^p_{p;[k,l]} \vee 2\|F\|_{C_b^1}C_p\|\mathbf{w}\|^p_{p;[k,l]}\right)\end{aligned}$$

from where the result follows. □

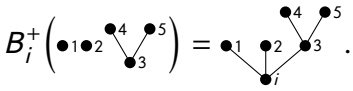
4.2 The rough regime

Before continuing we review the combinatorial setting for describing the composition of vector fields. We fix a collection $f_1,\dots,f_d$ of vector fields on $\mathbb{R}^n$. It turns out that a convenient framework for describing the kind of expansions we are looking for, is that of pre-Lie algebras.

Let $\mathcal{T}$ denote the set of non-planar rooted trees labeled by an at most countable index set I. By a slight abuse of notation, we denote by the same symbol the linear span of this set, i.e. the vector space formed by linear combinations of decorated trees. By *forest* we mean a disjoint union of trees, and we denote by $\mathcal{F}$ the set of all forests. There is a unique *empty forests* which we denote by $\emptyset$.

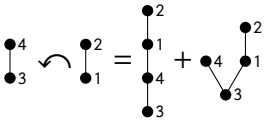
Given a label in I, we define an operator $B_i^+ : \mathcal{F} \to \mathcal{T}$ such that the image $B_i^+(\tau_1\cdots\tau_n)$ is obtained by grafting each of the trees $\tau_1,\dots,\tau_n$ onto a new root labeled by i.

Example 4.2.



Given two trees $\tau, \sigma \in \mathcal{T}$, $\tau \curvearrowleft \sigma$ denotes the linear combination of trees obtained by grafting the root of σ onto every vertex of τ.

Example 4.3.



We note that the operator B_i^+ defined above corresponds to the special case where τ is a tree containing a single node.

It can be shown that the grafting operation satisfies the (right) pre-Lie relation

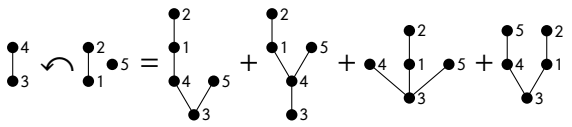
$$(\tau \curvearrowleft \sigma) \curvearrowleft \gamma - \tau \curvearrowleft (\sigma \curvearrowleft \gamma) = (\tau \curvearrowleft \gamma) \curvearrowleft \sigma - \tau \curvearrowleft (\gamma \curvearrowleft \sigma),$$

i.e. the associator $a_{\curvearrowleft}(\tau, \sigma, \gamma) := (\tau \curvearrowleft \sigma) \curvearrowleft \gamma - \tau \curvearrowleft (\sigma \curvearrowleft \gamma)$ is symmetric in σ, γ. In fact, the pair $\mathcal{T}, \curvearrowleft$ is the *free pre-Lie* algebra on $\#I$ generators [2].

The linear span of forests, also denoted by $\mathcal{F}$, can be thought of as the free polynomial algebra over $\mathcal{T}$; that is, a forests can be uniquely identified with a commuting polynomial with variables indexed by $\mathcal{T}$. In order to keep the notation simple, we identified these variables with their indices. In this sense, single forests correspond to monomials of trees. In particular, we use the standard product notation $\sigma = \sigma_1 \cdots \sigma_k \in \mathcal{F}$

The grafting operator can be upgraded to an operator $\curvearrowleft : \mathcal{T} \times \mathcal{F} \to \mathcal{F}$ given by grafting every forests on the right onto every vertex of the tree on the left. As a special case, we also define $\tau \curvearrowleft \emptyset := \tau$ for every $\tau \in \mathcal{T}$.

Example 4.4. Continuing with the above example,



The extended grafting is such that $\mathcal{T}$ becomes what is known as a *symmetric brace algebra* [17]. In particular, the following formula holds.

Proposition 4.5. *Let $\tau_1, \dots, \tau_n \in \mathcal{T}$, $\sigma = \sigma_1 \cdots \sigma_k \in \mathcal{F}$ and $i \in I$. Then*

$$B_i^+(\tau_1 \cdots \tau_n) \curvearrowleft \sigma = \sum_{I_1, \dots, I_{n+1}} B_i^+((\tau_1 \curvearrowleft \sigma_{J_1}) \cdots (\tau_n \curvearrowleft \sigma_{J_n}) \sigma_{J_{n+1}}).$$

where the sum is over all ways of decomposing the set $\{1, \dots, k\}$ into $n+1$ disjoint subset (some may be empty), and $\sigma_\emptyset = \emptyset$, $\sigma_J = \sigma_{j_1} \cdots \sigma_{j_r}$ for $\{j_1, \dots, j_r\} \subset \{1, \dots, k\}$.

Vector fields also satisfy the pre-Lie relation under composition as differential operators. Recall that any smooth vector field f on $\mathbb{R}^n$ can be identified with a first-order differential operator acting on C^1 functions by

$$(f \triangleright g)(x) := Dg(x) f(x).$$

Then, if f, g, h are vector fields of class C^1,

$$\begin{aligned}(f \rhd g) \rhd h - f \rhd (g \rhd h) &= D[Df(x)g(x)]h(x) - Df(x)Dg(x)h \\ &= D^2 f(x)(g(x), h(x))\end{aligned}$$

which is clearly symmetric in g, h. This justifies the following definition.

Definition 4.6. *Let $f_i : i \in I$ be vector fields in $\mathbb{R}^n$. We recursively define the* elementary vector fields *f_τ for $\tau \in \mathcal{T}$ recursively by $f_{\bullet_i} = f_i$ and*

$$f_{B_i^+(\tau_1 \cdots \tau_n)}(\mathbf{x}) = D^n f_i(\mathbf{x})(f_{\tau_1}(\mathbf{x}), \dots, f_{\tau_n}(\mathbf{x})).$$

This definition is extended to linearly to $\mathcal{F}$ in such a way that $f_{\tau_1 \cdots \tau_n} \equiv 0$ if $n \geq 1$.[2] *We also set $f_\emptyset = \mathrm{id}$.*

This definition is consistent because of the universality property of $(\mathcal{T}, \curvearrowright)$. As a direct consequence of Proposition 4.5 we obtain

Lemma 4.7. *Let $\tau_1, \dots, \tau_n \in \mathcal{T}$, $\sigma = \sigma_1 \cdots \sigma_k \in \mathcal{F}$ and $i \in I$. Then*

$$f_{B_i^+(\tau_1 \cdots \tau_n) \curvearrowright \sigma}(\mathbf{x}) = \sum_{r=n}^{n+k} \frac{1}{(r-n)!} \sum_{\rho, \gamma} D^r f_i(\mathbf{x})(f_{\tau_1 \curvearrowright \rho_1}, \dots, f_{\tau_1 \curvearrowright \rho_n}, f_{\gamma_1}, \dots, f_{\gamma_{r-n}}).$$

where the sum is over all forests $\rho_1, \dots, \rho_n, \gamma_1, \dots, \gamma_{r-n} \in \mathcal{F}$ such that $\sigma = \rho_1 \cdots \rho_n \cdot \gamma_1 \cdots \gamma_{r-n}$.

Proof. By Proposition 4.5, the formula

$$f_{B_i^+(\tau_1 \dots, \tau_n) \curvearrowright \sigma}(\mathbf{x}) = \sum_{I_1, \dots, I_{n+1}} f_{B_i^+((\tau_1 \curvearrowright \sigma_{J_1}) \cdots (\tau_n \curvearrowright \sigma_{J_n}) \sigma_{J_{n+1}})}(\mathbf{x})$$

holds. We split this sum according to the size of $J_{n+1} \subset \{1, \dots, k\}$ to obtain, in accordance with Definition 4.6,

$$f_{B_i^+(\tau_1 \dots, \tau_n) \curvearrowright \sigma}(\mathbf{x}) = \sum_{r=0}^{k} \sum_{I_1, \dots, I_n} \frac{1}{r!} D^{n+r} f_i(f_{\tau_1 \curvearrowright \sigma_{J_1}}, \dots, f_{\tau_n \curvearrowright \sigma_{J_n}}, f_{\sigma_1'}, \cdots, f_{\sigma_r'})$$

where $\sigma_{J_{n+1}} = \sigma_1' \cdots \sigma_r'$ and the combinatorial factor appears because of the symmetry of this expression with respect to $\sigma_{J_{n+1}}$. The result follows from this formula by substituting indices in the first summation and identifying $\rho_k = \sigma_{J_k}, \gamma_k = \sigma_k'$. □

For notational simplicity, instead of eq. (6), we address the slightly more general problem

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k + \sum_{s=1}^{[p]} \sum_{i_1, \dots, i_s = 1}^{d} f_{[\mathrm{i}_1 \cdots \mathrm{i}_s]}(\mathbf{x}_k)(\mathbf{w}_{k+1}^{i_1} - \mathbf{w}_k^{i_1}) \cdots (\mathbf{w}_{k+1}^{i_s} - \mathbf{w}_k^{i_s}) \\ &= \mathbf{x}_k + \sum_{s=1}^{[p]} f_{[\mathrm{i}_1 \cdots \mathrm{i}_s]}(\mathbf{x}_k) \mathbb{W}_{k,k+1}^{[\mathrm{i}_1 \cdots \mathrm{i}_s]}\end{aligned} \tag{9}$$

This includes eq. (6) as the special case where we set $f_{[\mathrm{i}_1 \cdots \mathrm{i}_s]} \equiv 0$ for $s > 1$. Equation (9) is a discrete analogue of a truncated version of Marcus' canonical extension [16].

[2] Maps with this property are sometimes called *infinitesimal characters* in the literature.

Remark 4.8. Formally speaking, the formalism we have introduced also serves to handle the case where the right-hand side of eq. (9) is replaced by an infinite series containing all possible contractions. To avoid any analytical complication that might arise from such considerations we refrain from doing so, but we note that at least the combinatorial formulas carry over without any difficulties.

We particularize the sets $\mathcal{T}$ and $\mathcal{F}$ of decorated trees and forests to decorations from the extended alphabet $\mathfrak{A}$. Concretely, this means that any contraction of the form $a = [\mathrm{i}_1 \cdots \mathrm{i}_k] \in \mathfrak{A}$ can appear attached to a node.

We also consider the *contracting arborification* $\mathfrak{a}_c \colon \mathcal{F} \to H$ [1, 6]. It is recursively defined by

$$\mathfrak{a}_c(B_i^+(\tau_1 \cdots \tau_k)) = (\mathfrak{a}_c(\tau_1) \star \cdots \star \mathfrak{a}_c(\tau_k))[\mathrm{i}].$$

This formula defines the image of $\mathfrak{a}_c$ over $\mathcal{T}$. Since $\mathcal{F}$ is the free polynomial algebra over trees, this definition admits a unique extension to $\mathcal{F}$, which is given by

$$\mathfrak{a}_c(\tau_1 \cdots \tau_n) = \mathfrak{a}_c(\tau_1) \star \cdots \star \mathfrak{a}_c(\tau_n).$$

Example 4.9.

$$\mathfrak{a}_c(\bullet_1\bullet_2) = [1][2] + [2][1] + [12],$$
$$\mathfrak{a}_c\left(\begin{smallmatrix}\bullet^1 & & \bullet^2\\ & \bullet_3 & \end{smallmatrix}\right) = (\mathfrak{a}_c(\bullet_1) \star \mathfrak{a}_c(\bullet_2))[3] = [1][2][3] + [2][1][3] + [12][3].$$

Finally, we introduce a *grading* on forests which is compatible with the weight of words in H. For $\tau = \bullet_a, a \in \mathfrak{A}$ we set $|\tau| := |a|$. Inductively, we define $|B_a^+(\tau_1 \cdots \tau_n)| := |\tau_1 \cdots \tau_n| + |a|$ and $|\tau_1 \cdots \tau_n| := |\tau_1| + \cdots + |\tau_n|$.

Concretely, the grading on $\mathcal{F}$ counts the total weight in $\mathfrak{A}$ of the decorations appearing in the forest. Note that this differs with the usual grading introduced on $\mathcal{F}$ by the number of nodes. As a simple example, the tree $\bullet_{[23]}$ has degree two instead of just one. By analogy with the definition for words, for every integer $n \geq 0$ we let $\mathcal{F}_n$ be the set of forests with weight exactly equal to n, and $\mathcal{F}_{(n)}$ is the union of the $\mathcal{F}_k$ for $k \leq n$. We also define $F^0_{(n)} := \mathcal{F}_{(n)} \setminus \{\varnothing\}$.

We use the contracting arborification map to transform eq. (9) into its arborified version. First, we define a transformed path $\overline{\mathbb{W}}^\tau_{k,l} := \mathbb{W}^{\mathfrak{a}_c(\tau)}_{k,l}$, for all forests $\tau \in \mathcal{F}_d$.

Example 4.10. For all $a, b, c \in \mathfrak{A}$ we have

$$\overline{\mathbb{W}}^{\bullet^a}_{k,l} = \mathbb{W}^a_{k,l}, \quad \overline{\mathbb{W}}^{\begin{smallmatrix}\bullet^b & \bullet^c\\ \bullet_a & \end{smallmatrix}}_{k,l} = \mathbb{W}^{[b][c][a]}_{k,l} + \mathbb{W}^{[c][b][a]}_{k,l} + \mathbb{W}^{[bc][a]}_{k,l}.$$

By its very definition, the aborified map $\overline{\mathbb{W}}$ satisfies the product rule

$$\overline{\mathbb{W}}^\tau_{k,l}\overline{\mathbb{W}}^\sigma_{k,l} = \overline{\mathbb{W}}^{\tau\sigma}_{k,l}$$

for all forests $\tau, \sigma \in \mathcal{F}_d$. Moreover, it satisfies Chen's relation

$$\delta\overline{\mathbb{W}}^\tau_{k,l,m} = \overline{\mathbb{W}}^\tau_{k,m} - \overline{\mathbb{W}}^\tau_{k,l} - \overline{\mathbb{W}}^\tau_{l,m} = \sum_{(\tau)} \overline{\mathbb{W}}^{\tau_1}_{k,l}\overline{\mathbb{W}}^{\tau_2}_{l,m}.$$

for all $\tau \in \mathcal{T}$. Here, the terms in the sum are obtained from τ by performing *admissible cuts*, and we use the convention where τ_2 is the connected component containing the root. See e.g. [9, 12] for further details. In particular, this sum can be rewritten as

$$\sum_{\rho \in \mathcal{T}, \sigma \in \mathcal{F}} c(\sigma, \rho; \tau)\overline{\mathbb{W}}^\sigma_{k,l}\overline{\mathbb{W}}^\rho_{l,m} \tag{10}$$

where the coefficient $c(\sigma, \rho; \tau)$ is the coefficient of τ in the linear combination $\rho \curvearrowleft \sigma$.

We are now ready to prove a formal Taylor-like expansion for the large scale increments of the solution $\mathbf{x}$ of eq. (9). Given $p \geq 2$, we $[p] \in \mathbb{N}$ denote its integer part, i.e. the unique integer such that $[p] \leq p \leq [p] + 1$. With this notation, the controlled difference equation becomes

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \sum_{a \in \mathfrak{A}: |a| \leq [p]} f_{\bullet_a}(\mathbf{x}_k) \overline{\mathbb{W}}^{\bullet_a}_{k,k+1}.$$

Theorem 4.11. *For any $p \geq 1$, the inequality*

$$\|\mathbf{x}\|_{p;[k,l]} \leq \left(C_p^{p-1} \|F\|^p_{C_b^{[p]+1}} |\!|\!| \mathbb{W} |\!|\!|^p_{p;[k,l]} \vee \|F\|_{C_b^{[p]+1}} |\!|\!| \mathbb{W} |\!|\!|_{p;[k,l]} \right)$$

holds, with

$$C_p := 2^{([p]+1)/p} \zeta\left(\frac{[p]+1}{p} \right).$$

We prove Theorem 4.11 in various stages, but first we observe that since eq. (9) is bilinear in f and $\mathbb{W}$ we can –and do– assume that $\|F\|_{C_b^{[p]+1}} \leq 1$. The first step is to consider the "germ"

$$\overline{\Xi}_{k,l} := \sum_{\mathcal{F}^0_{([p])}} \frac{1}{\Sigma(\tau)} f_\tau(\mathbf{x}_k) \overline{\mathbb{W}}^\tau_{k,l} \tag{11}$$

where we recall that $\Sigma(\tau)$ is the internal symmetry factor of the tree τ [1].

For a given tree $\tau \in \mathcal{T}$ we define the remainder term

$$\overline{R}^\tau_{k,l} := f_\tau(\mathbf{x}_l) - \sum_{|\sigma| \leq [p] - |\tau|} \frac{1}{\Sigma(\sigma)} f_{\tau \curvearrowright \sigma}(\mathbf{x}_k) \overline{\mathbb{W}}^\sigma_{k,l}, \tag{12}$$

It will also be necessary to allow for a more general version of this remainder, parametrized by an integer $0 \leq s \leq [p] - |\tau|$, defined as

$$\overline{R}^{\tau,s}_{k,l} := f_\tau(\mathbf{x}_l) - \sum_{|\sigma| \leq s} \frac{1}{\Sigma(\sigma)} f_{\tau \curvearrowright \sigma}(\mathbf{x}_k) \overline{\mathbb{W}}^\sigma_{k,l}, \tag{13}$$

The general idea is that such remainder should have be measured in the $\frac{p}{s+1}$-variation norm.

One last remark: from now on we continue working with $\overline{\mathbb{W}}$ instead of $\mathbb{W}$. This causes no harm since by definition

$$\sum_{\tau \in \mathcal{T}} f_\tau(\mathbf{x}_k) \overline{\mathbb{W}}^\tau_{k,l} = \sum_{u \in \mathfrak{A}} f_{\mathfrak{a}_c^\dagger(u)}(\mathbf{x}_k) \mathbb{W}^u_{k,l},$$

so any estimate we prove using the sum on the right immediately implies that the same bound holds when we put the sum on the right in its place. In the same vein, we define $\overline{\mathbb{W}}^{(n)} := (\overline{\mathbb{W}}^\tau : |\tau| = n)$ and

$$|\!|\!| \overline{\mathbb{W}} |\!|\!|_{p;[k,l]} := \max_{n=1,\dots,[p]} \|\overline{\mathbb{W}}^{(n)}\|^{1/n}_{p/n;[k,l]},$$

and remark that since $\mathfrak{a}_c(H_n) \subset \mathcal{T}_n$ for all $n \geq 0$, the bounds

$$|\!|\!| \overline{\mathbb{W}} |\!|\!|_{p;[k,l]} \lesssim |\!|\!| \mathbb{W} |\!|\!|_{p;[k,l]} \lesssim |\!|\!| \overline{\mathbb{W}} |\!|\!|_{p;[k,l]}$$

hold uniformly on k, l for all $p \geq 1$.

Proposition 4.12. *The bound*

$$\left|\mathbf{x}_{k,l} - \sum_{0<|\tau|\leq[p]} \frac{1}{\Sigma(\tau)} f_\tau(\mathbf{x}_k)\overline{\mathbb{W}}^\tau_{k,l}\right| \leq C_p \sum_{0<|\tau|\leq[p]} \|R^\tau\|_{p/([p]+1-|\tau|);[k,l]} |||\overline{\mathbb{W}}|||^{|\tau|}_{p;[k,l]}$$

holds for all $0 \leq k < l \leq N$, *where* $C_p := 2^{([p]+1)/p}\zeta\left(\frac{[p]+1}{p}\right)$.

Proof. By eq. (4)

$$\delta\Xi_{k,l,m} = -\sum_{0<|\tau|\leq[p]} \frac{1}{\Sigma(\tau)}\left((f_\tau(\mathbf{x}) - f_\tau(\mathbf{x}_k))\overline{\mathbb{W}}^\tau_{l,m} - \sum_{(\tau)} f_\tau(\mathbf{x}_k)\overline{\mathbb{W}}^{\tau_1}_{k,l}\overline{\mathbb{W}}^{\tau_2}_{l,m}\right)$$

Thanks to eq. (10), we rewrite it as

$$\sum_{\sigma\in\mathcal{F},\rho\in\mathcal{T}}\sum_{0<|\tau|\leq[p]} c(\sigma,\rho;\tau) f_\tau(\mathbf{x}_k)\overline{\mathbb{W}}^\sigma_{k,l}\overline{\mathbb{W}}^\rho_{l,m} = \sum_{|\rho|\leq[p]}\sum_{\sigma\in\mathcal{F}} f_{\tau\curvearrowright\sigma}(\mathbf{x}_k)\overline{\mathbb{W}}^\sigma_{k,l}\overline{\mathbb{W}}^\rho$$

since, by definition

$$\rho \curvearrowleft \sigma = \sum_{\tau\in\mathcal{T}} c(\sigma,\rho;\tau)\tau.$$

Therefore, since the internal symmetry factor is multiplicative,

$$\delta\Xi_{k,l,m} = -\sum_{0<|\tau|\leq[p]} \frac{1}{\Sigma(\tau)}\left(f_\tau(\mathbf{x}_l) - \sum_{|\sigma|\leq[p]-|\tau|} \frac{1}{\Sigma(\sigma)} f_{\tau\curvearrowright\sigma}\overline{\mathbb{W}}^\sigma_{k,l}\right)\overline{\mathbb{W}}^\tau_{l,m}.$$

In particular,

$$|\delta\Xi_{k,l,m}| \leq \sum_{0<|\tau|\leq[p]} \frac{1}{\Sigma(\tau)}|R^\tau_{k,l}||\overline{\mathbb{W}}^\tau_{l,m}| \leq \sum_{0<|\tau|\leq[p]} \frac{1}{\Sigma(\tau)}\omega^{([p]+1-|\tau|)/p}_{u;k,l}\hat{\omega}^{|\tau|/p}_{\tau;l,m}$$

with $\omega_{\tau;k,l} := \|R^\tau\|^{p/([p]+1-|\tau|)}_{p/([p]+1-|\tau|);[k,l]}$ and $\hat{\omega}_{\tau;k,l} = \|\overline{\mathbb{W}}^\tau\|^{p/|\tau|}_{p/|\tau|;[k,l]}$.

Since

$$\frac{[p]+1-|\tau|}{p} + \frac{|\tau|}{p} = \frac{[p]+1}{p} > 1,$$

the Sewing Lemma (Proposition 2.15) implies the desired bound. □

We now introduce the *error terms*

$$\mathcal{E}^{(s)}_{k,l} := \mathbf{x}_l - \sum_{|\tau|\leq s} \frac{1}{\Sigma(\tau)} f_\tau(\mathbf{x}_k)\overline{\mathbb{W}}^\tau_{k,l}$$

for $s = 0, \ldots, [p]$. We remark that since the elementary vector fields vanish on forests, the error term consists of a sum only over trees.

Lemma 4.13. *Let $\tau \in \mathcal{T}$ with $|\tau| \le [p]$ and n nodes, and $0 \le s \le [p] - |\tau|$ There is a control $w^{(\tau)}$, depending polynomially on the error terms $\mathcal{E}^{(s+n-r)}$, $r = n+1, \dots, n+s$, $\|\!|\overline{\mathbb{W}}|\!\|_p$ and $\overline{R}^{\rho,s'}$ for $|\rho| < |\tau|$ and some $s' < s$, such that*

$$\left\| \overline{R}^{\tau,s}_{k,l} \right\|^{p/(s+1)}_{p/(s+1);[k,l]} \lesssim \|\mathbf{x}\|^p_{p;[k,l]} + \|\!|\overline{\mathbb{W}}|\!\|^p_{p;[k,l]} + \omega^{(\tau,s)}_{k,l}$$

Moreover, the dependence on $\mathcal{E}^{(s+n-r)}$ is a polynomial of degree $r-n$ and a sum of homogeneous polynomials of degree $0, \dots, n$ on the $\overline{R}^{\tau,s'}$.

Proof. We proceed by induction on the number of nodes in τ, the case when $\tau = \emptyset$ corresponds to Proposition 4.12, since $R^{\emptyset}_{k,l} = \mathcal{E}^{([p])}_{k,l} = \mathbf{x}_{k,l} - \overline{\Xi}_{k,l}$. We also note that in general $n \le |\tau|$.

Suppose $\tau = B^+_a(\tau_1 \cdots \tau_n)$ for some trees $\tau_1, \dots, \tau_n \in \mathcal{T}$ and some $a \in \mathfrak{A}$. On the one hand, by performing a Taylor expansion up to order s we see that

$$\begin{aligned} f_\tau(\mathbf{x}_l) &= D^n f_a(\mathbf{x}_l) : (f_{\tau_1}, \dots, f_{\tau_n}) \\ &= \sum_{r=n}^{n+s} \frac{1}{(r-n)!} D^r f_a(\mathbf{x}_k)(f_{\tau_1}(\mathbf{x}_l), \dots, f_{\tau_n}(\mathbf{x}_l), \mathbf{x}_{k,l}, \dots \mathbf{x}_{k,l}) + B_{k,l} \end{aligned} \tag{14}$$

where the remainder term satisfies

$$|B_{k,l}| \le \frac{|\mathbf{x}_{k,l}|^{s+1}}{(s+1)!} \le \|\mathbf{x}\|^{s+1}_{p;[k,l]}.$$

On the other hand, thanks to Lemma 4.7 we see that

$$\sum_{|\sigma| \le s} f_{\tau \curvearrowleft \sigma}(\mathbf{x}_k) \overline{\mathbb{W}}^{\sigma}_{k,l} = \sum_{r=n}^{n+s} \frac{1}{(r-n)!} \sum_{\rho,\gamma} D^r f_a(\mathbf{x}_k) : (f_{\tau_1 \curvearrowleft \rho_1}, \dots, f_{\tau_n \curvearrowleft \rho_n}, f_{\gamma_1}, \dots, f_{\gamma_{n-r}}) \overline{\mathbb{W}}^{\rho\gamma}_{k,l} \tag{15}$$

where now the inner sum is over forests $\rho_1, \dots, \rho_n, \gamma_1, \dots, \gamma_{n-r}$ such that the sum of their weights is less than or equal to s, and $|\gamma_j| > 0$.

For a fixed $n \le r \le n+s$, we replace in eq. (14) the identities

$$f_{\tau_j}(\mathbf{x}_l) = \sum_{|\rho| \le s'_j} f_{\tau_j \curvearrowleft \rho} \overline{\mathbb{W}}^{\rho}_{k,l} + R^{\tau_j, s'_j}_{k,l}$$

and

$$\mathbf{x}_{k,l} = \sum_{0 < |\gamma| \le s+n-r} f_\gamma(\mathbf{x}_k) \overline{\mathbb{W}}^{\gamma}_{k,l} + \mathcal{E}^{(s+n-r)}_{k,l}.$$

where $0 \le s'_j \le [p] - |\tau_j|$ will be chosen later. In the end, we see that the corresponding term in eq. (14) equals

$$\sum_{\rho',\gamma'} D^r f_a(\mathbf{x}_k) : (f_{\tau_1 \curvearrowleft \rho'_1}, \dots, f_{\tau_n \curvearrowleft \rho'_n}, f_{\gamma'_1}, \dots, f_{\gamma'_{r-n}}) \overline{\mathbb{W}}^{\rho'\gamma'}_{k,l} + \mathcal{R}^{\tau}_{k,l}$$

where $\mathcal{R}^{\tau}_{k,l}$ is the sum of all the terms that contain at least one factor from the set

$$\left\{ R^{\tau_1, s'_1}_{k,l}, \dots, R^{\tau_n, s'_n}_{k,l}, \mathcal{E}^{(s+n-r)}_{k,l} \right\}$$

and $\mathcal{E}_{k,l}^{(s+n-r)}$ can appear between zero and $r-n$ times. Also, $|\rho_j'| \le s_j$ and $0 < |\gamma_j'| \le s+n-r$. In order to have enough terms to cancel out eq. (15) we need that $\sum s_j' \ge s-(r-n)$ and so we choose $0 \le s_j' \equiv [\frac{s+n-r}{n}] < s$. Given this condition, we can be sure that the difference between eqs. (14) and (15) will be of the form

$$R_{k,l}^{\tau} = \mathcal{R}_{k,l}^{\tau,s} + \mathcal{X}_{k,l}^{\tau} + B_{k,l}$$

where now $\mathcal{X}^{\tau}$ contains all the terms like those in eq. (15) with $s < |\rho|+|\gamma| \le (s+n-r)(r-n+1)$. This term is bounded by

$$|\mathcal{X}_{k,l}^{\tau}| \le \sum_{j=1}^{(s+n-r-1)(r-n)} |||\overline{\mathbb{W}}|||_{p;[k,l]}^{s+j} \lesssim |||\overline{\mathbb{W}}|||_{p;[k,l]}^{s+1}$$

by the smallness assumption on $|||\overline{\mathbb{W}}|||_{p;[k,l]}$.

Finally, we analyse the remainder term $\mathcal{R}^{\tau,s}$ more closely. A generic term looks, modulo a combinatorial factor, like

$$\sum_{\rho,\gamma} D^r f_a(\mathbf{x}_k) : \left\{ R_{k,l}^{\tau_{i_1},s_r'}, \dots, R_{k,l}^{\tau_{i_j},s_r'}, f_{\tau_{\ell_1}\curvearrowleft\rho_{\ell_1}}, \dots, f_{\tau_{\ell_{n-j}}\curvearrowleft\rho_{\ell_{n-j}}}, \left(\mathcal{E}_{k,l}^{(s+n-r)}\right)^t, f_{\gamma_1}, \dots, f_{\gamma_{r-n-t}} \right\} \overline{\mathbb{W}}_{k,l}^{\sigma\gamma}$$

for some $r \in \{n, \dots, n+s\}$, and integers $j, t \ge 0$ such that $j \le n$ and $t \le r-n$, and the sum is over forests ρ_{ℓ_j}, γ_j. Its absolute value is therefore bounded by

$$\sum_{t=0}^{n-r} \sum_{k=1}^{n} \sum_{\ell=r-n-t}^{\hat{\ell}} \|\mathcal{E}^{(s+n-r)}\|_{p/(s+n-r+1);[k,l]}^{t} |||\overline{\mathbb{W}}|||_{p;[k,l]}^{\ell} S_k\left(\|\overline{R}^{\tau_1,s_r'}\|_{p/(s_r'+1)}, \dots, \|\overline{R}^{\tau_n,s_r'}\|_{p/(s_r'+1)} \right).$$

where

$$S_k(x_1, \dots, x_n) := \sum_{1 \le j_1 < \dots < j_k \le n} x_{j_1} \cdots x_{j_k}$$

is the elementary symmetric polynomial of degree k, and

$$\hat{\ell} := (n-k)([\tfrac{s+n-r}{n}]+1) + (r-n-t)(s+n-r)$$

.

Finally we note that for each term the sum of exponents is

$$\frac{t(s+n-r+1)+\ell+ks_r'+k}{p} \ge \frac{t(s+n-r+1)+(r-n-t)+[(s+n-r)/n]}{p} \ge \frac{s+1}{p}$$

□

Proof of Theorem 4.11. First, we claim that for all $r = 1, \dots, [p]$, there exists a control $w^{(r)}$ such that

$$\|\mathcal{E}^{([p]-r)}\|_{p/([p]-r+1);[k,l]}^{p/([p]-r+1)} \lesssim w_{k,l}^{(r)} + \sum_{r<|\tau|\le[p]} \|\overline{R}^{\tau}\|_{p/([p]-|\tau|+1);[k,l]}^{p/([p]-|\tau|+1)} |||\overline{\mathbb{W}}|||_{p;[k,l]}^{p(|\tau|-r)/([p]-r+1)}.$$

Moreover, $w^{(r)}$ can be chonsen to have the form

$$\omega_{k,l}^{(r)} = \sum_{j=1}^{r} \|\mathbf{x}\|_{p;[k,l]}^{jp} + \hat{w}_{k,l}^{(r)}$$

with $\hat{w}^{(r)}$ depending polynomially on $\mathcal{E}^{([p]-r-j)}, j = 1, \dots, [p]-r$ and $\|\!|\overline{\mathbb{W}}|\!\|^p_{p;[k,l]}$. Indeed, if $r = 1$, we have

$$|\mathcal{E}^{([p]-1)}_{k,l}| \leq |\mathbf{x}_{k,l} - \overline{\Xi}_{k,l}| + \|\!|\overline{\mathbb{W}}|\!\|^{[p]}_{p;[k,l]}.$$

Therefore, by Proposition 4.12 we obtain

$$|\mathcal{E}^{([p]-1)}_{k,l}|^{p/[p]} \lesssim 2^{p/[p]-1} C_p^{p/[p]} \sum_{0<|\tau|\leq[p]} \|\overline{R}^\tau\|^{p/[p]}_{p/([p]-|\tau|+1);[k,l]} \|\!|\overline{\mathbb{W}}|\!\|^{p|\tau|/[p]}_{p;[k,l]} + 2^{p/[p]-1} \|\!|\overline{\mathbb{W}}|\!\|^p_{p;[k,l]}.$$

Consider the term with $|\tau| = 1$. By the Lemma 4.13, the bound

$$\|R^\tau\|^{p/[p]}_{p/[p];[k,l]} \leq \|\mathcal{E}^{([p]-1)}\|^{p/[p]}_{p/[p];[k,l]} + w^{(1)}_{k,l} + \|\mathbf{x}\|^p_{p;[k,l]}$$

holds for all trees with $|\tau| = 1$, and some control $w^{(1)}_{k,l}$ depending polynomialy on $\mathcal{E}^{([p]-r)}_{k,l}$ for $r = 2, \dots, [p]$.

Choosing $k < l$ such that $2C_p \|\!|\overline{\mathbb{W}}|\!\|_{p;[k,l]} \leq 1$ we obtain

$$\|\mathcal{E}^{([p]-1)}\|^{p/[p]}_{p/[p];[k,l]} \lesssim \frac{1}{2}\|\mathcal{E}^{([p]-1)}\|^{p/[p]}_{p/[p];[k,l]} + \frac{1}{2} \sum_{1<|\tau|\leq[p]} \|\overline{R}^\tau\|^{p/[p]}_{p/([p]-|\tau|+1);[k,l]} \|\!|\overline{\mathbb{W}}|\!\|^{p(|\tau|-1)/[p]}_{p;[k,l]} + \omega^{(1)}_{k,l}$$

and the claim is proven in the base case.

Now, assume it is true for all integers from 1 up to $r-1$. Then, since

$$\mathcal{E}^{([p]-r)}_{k,l} = \mathcal{E}^{([p]-r+1)}_{k,l} + \sum_{|\tau|=[p]-r+1} f_\tau(\mathbf{x}_k)\overline{\mathbb{W}}^\tau_{k,l}$$

we obtain

$$|\mathcal{E}^{([p]-r)}_{k,l}|^{p/([p]-r+1)} \lesssim \|\mathcal{E}^{([p]-r+1)}\|^{p/([p]-r+1)}_{p/([p]-r+2);[k,l]} + \|\!|\overline{\mathbb{W}}|\!\|^p_{p;[k,l]}.$$

By the induction hypothsis,

$$\|\mathcal{E}^{([p]-r+1)}\|^{p/([p]-r+2)}_{p/([p]-r+2);[k,l]} \lesssim w^{(r-1)}_{k,l} + \sum_{r-1<|\tau|\leq[p]} \|\overline{R}^\tau\|^{p/([p]-r+2)}_{p/([p]-|\tau|+1);[k,l]} \|\!|\overline{\mathbb{W}}|\!\|^{p(|\tau|-r+1)/([p]-r+2)}_{p;[k,l]}.$$

Raising both sides to the power of $\frac{[p]-r+2}{[p]-r+1} > 1$, we get

$$|\mathcal{E}^{([p]-r)}_{k,l}|^{p/([p]-r+1)} \lesssim \left(w^{(r-1)}_{k,l}\right)^{\frac{[p]-r+2}{[p]-r+1}} + \sum_{r-1<|\tau|\leq[p]} \|\overline{R}^\tau\|^{p/([p]-r+1)}_{p/([p]-|\tau|+1);[k,l]} \|\!|\overline{\mathbb{W}}|\!\|^{p(|\tau|-r+1)/([p]-r+1)}_{p;[k,l]}.$$

Isolate from the right-hand side the term with $|\tau| = r$. According to Lemma 4.13, for all such trees remainder satisfies

$$\|\overline{R}^\tau\|^{p/([p]-r+1)}_{p/([p]-r+1);[k,l]} \lesssim \|\mathcal{E}^{([p]-r)}\|^{p/([p]-r+1)}_{p/([p]-r+1);[k,l]} + \|\mathbf{x}\|^p_{p;[k,l]} + \tilde{w}^{(r)}_{k,l}.$$

Again by Lemma 4.13, the dependence of $w^{(r-1)}$ on $\mathcal{E}^{([p]-r)}$ is such that

$$\left(w^{(r-1)}_{k,l}\right)^{\frac{[p]-r+2}{[p]-r+1}} \lesssim \sum_{j=2}^{r} \|\mathcal{E}^{([p]-r)}\|^{jp/([p]-r+1)}_{p/([p]-r+1)} + \hat{w}^{(r-1)}_{k,l}$$

so that

$$\|\mathcal{E}^{([p]-r)}\|^{p/([p]-r+1)}_{p/([p]-r+1);[k,l]} \lesssim \frac{1}{2}\|\mathcal{E}^{([p]-r)}\|^{p/([p]-r+1)}_{p/([p]-r+1);[k,l]} + \frac{1}{2}\sum_{j=2}^{r}\|\mathcal{E}^{([p]-r)}\|^{jp/([p]-r+1)}_{p/([p]-r+1)} + \omega^{(r)}_{k,l}$$

This polynomial inequality is of the form

$$Q_n(\lambda) := \lambda^n + \lambda^{n-1} + \cdots + \lambda^2 - \lambda + c \geq 0$$

for some small constant $c > 0$. Note that $Q_n(0) = c > 0$ and $Q_n(2c) = -c + o(c) < 0$ if c is sufficiently small. [3] Therefore, given that c is small enough, the smallest possitive root of Q_n is smaller than $2c$. More precisely, we need that

$$\frac{(2c - 2^n c^n)}{(1-2c)} < \frac{1}{2}.$$

So the claim is proven.

At the end of the induction we have that

$$\|\mathbf{x}\|^p_{p;[k,l]} \lesssim \omega^{([p])}_{k,l}$$

and the inductive argument shows that in fact

$$\|\mathbf{x}\|^p_{p;[k,l]} \lesssim \frac{1}{2}\sum_{j=1}^{[p]}\|\mathbf{x}\|^{jp}_{p;[k,l]} + |\!|\!|\overline{\mathbb{W}}|\!|\!|^p_{p;[k,l]}.$$

We again arrive at a polynomial inequation of the previous form, so the same argument gives

$$\|\mathbf{x}\|^p_{p;[k,l]} \lesssim 2^p C_p^p |\!|\!|\overline{\mathbb{W}}|\!|\!|^p_{p;[k,l]}.$$

The conclusion is obtained by an application of Lemma 2.12 and scaling back $\|F\|_{C^{[p]+1}}$. □

Corollary 4.14. *Let $(\mathbf{x}_k : 0, \dots, N)$ be the complete evolution of the input features $\mathbf{x}_0$ through a trained ResNet with weights $\mathbf{w}$ and activation functions σ; that is, the values of $\mathbf{x}_k$, $k = 1, \dots, N$ are obtained from eq.* (6)*. For $p \geq 1$, suppose that $\sigma \in C^{[p]+1}_b$ and consider the discrete signature lift $\mathbb{W}$ of $\mathbf{w}$. The inequality*

$$\|\mathbf{x}\|_{p;[k,l]} \leq \left(C_p^{p-1}\|\sigma\|^p_{C^{[p]+1}_b}|\!|\!|\mathbb{W}|\!|\!|^p_{p;[k,l]} \vee \|\sigma\|_{C^{[p]+1}_b}|\!|\!|\mathbb{W}|\!|\!|_{p;[k,l]}\right)$$

holds uniformly over all $0 \leq k < l \leq N$, with $C_p := 2^{([p]+1)/p}\zeta\left(\frac{[p]+1}{p}\right)$. In particular

$$|\mathbf{x}_N - \mathbf{x}_0| \leq \inf_{p\in[1,\infty)}\left(C_p^{p-1}\|\sigma\|^p_{C^{[p]+1}_b}|\!|\!|\mathbb{W}|\!|\!|^p_{p;[k,l]} \vee \|\sigma\|_{C^{[p]+1}_b}|\!|\!|\mathbb{W}|\!|\!|_{p;[k,l]}\right). \tag{16}$$

Proof. Let d be the width of the network. Apply Theorem 4.11 with $f_i = \sigma$ for all $i = 1, \dots, d$ and $f_{[i_1 \cdots i_n]} \equiv 0$ for all $n > 0$. This proves the first inequality. The second result is obtained from Equation (3). □

[3] In fact, by Decarte's rule of signs, the polynomial Q_n has either 0 or 2 positive roots, for any $n \geq 2$.

5 Conclusion and outlook

We have shown how to control the total p-variation of the evolution of features through a Deep ResNet by the p-variation of the weights (trained or not). When compared to the classical $p = 1$ setting, the improvement was seen to be significant and exhibits a structural feature of deep neural networks, whose importance, to the best of our knowledge, has not yet been sufficiently appreciated: weights are indeed *rough*, so that a proper stability analysis much benefits from our discrete rough-path approach.

Work in progress deals with a variation of these estimates to prove stability under retraining. Concretely, we expect a bound of the form

$$|\mathbf{x}_N - \tilde{\mathbf{x}}_N| \leq \inf_{p\in[1,\infty)} \left\{ K_p \left(|\mathbf{x}_0 - \tilde{\mathbf{x}}_0| + |||\mathbb{W} - \widetilde{\mathbb{W}}|||_p \right) e^{K_p \left(|||\mathbb{W}|||^p_{p-\mathrm{var}} + |||\widetilde{\mathbb{W}}|||^p_p \right)} \right\}, \tag{17}$$

for some explicitly computable constant $K_p > 0$, depending only on p and $\|\sigma\|$,uniformly in the network depth N. Here is some numerical evidence suggesting that considering $p \geq 1$, i.e. going beyond the Lipschitz theory can provide an advantage.

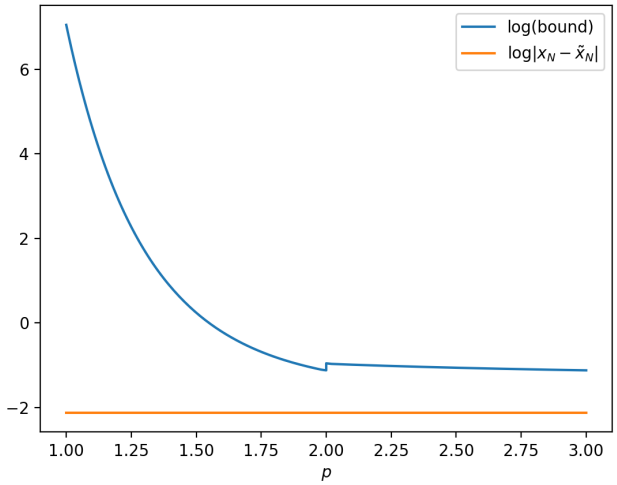


Figure 3: Bound on eq. (17) for different values of $p \in [1, 2]$ vs. the output value of two different values of a single feature.

References

[1] Y. Bruned, C. Curry, and K. Ebrahimi-Fard, "Quasi-shuffle algebras and renormalisation of rough differential equations," *Bulletin of the London Mathematical Society* **52** no. 1, (Nov, 2019) 43–63.

[2] F. Chapoton and M. Livernet, "Pre-Lie algebras and the rooted trees operad," *International Mathematics Research Notices* **2001** no. 8, (2001) 395.

[3] A. M. Davie, "Differential equations driven by rough paths: an approach via discrete approximation," *Applied Mathematics Research Express. AMRX* (2008) .

[4] J. Diehl, K. Ebrahimi-Fard, and N. Tapia, "Time warping invariants of multidimensional time series," *Acta Appl. Math.* (2020) , arXiv:1906.05823 [math.RA].

[5] W. E, "A proposal on machine learning via dynamical systems," *Communications in Mathematics and Statistics* **5** no. 1, (2017) 1–11.

[6] K. Ebrahimi-Fard, F. Fauvet, and D. Manchon, "A comodule-bialgebra structure for word-series substitution and mould composition," *Journal of Algebra* **489** (2017) 552 – 581.

[7] D. Feyel and A. de La Pradelle, "Curvilinear Integrals Along Enriched Paths," *Electronic Journal of Probability* **11** no. 0, (2006) 860–892.
http://dx.doi.org/10.1214/EJP.v11-356.

[8] M. Gubinelli, "Controlling rough paths," *Journal of Functional Analysis* **216** no. 1, (2004) 86–140.

[9] M. Gubinelli, "Ramification of rough paths," *Journal of Differential Equations* **248** no. 4, (2010) 693–721.

[10] E. Haber and L. Ruthotto, "Stable architectures for deep neural networks," *Inverse Problems. An International Journal on the Theory and Practice of Inverse Problems, Inverse Methods and Computerized Inversion of Data* **34** no. 1, (2018) 014004, 22.

[11] E. Haber, L. Ruthotto, E. Holtham, and S.-H. Jun, "Learning Across Scales—Multiscale Methods for Convolution Neural Networks," in *AAAI Conference on Artificial Intelligence*. 2018.

[12] M. Hairer and D. Kelly, "Geometric versus non-geometric rough paths," *Ann. Inst. Henri Poincaré Probab. Stat.* **51** no. 1, (2015) 207–251.

[13] 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)*. IEEE, 2016.

[14] M. E. Hoffman, "Quasi-shuffle products," *Journal of Algebraic Combinatorics* **11** no. 1, (2000) 49–68.

[15] T. J. Lyons, "Differential equations driven by rough signals.," *Revista Matemática Iberoamericana* **14** no. 2, (1998) 215–310.

[16] S. Marcus, "Modeling and analysis of stochastic differential equations driven by point processes," *IEEE Transactions on Information Theory* **24** no. 2, (Mar, 1978) 164–172.

[17] J.-M. Oudom and D. Guin, "On the Lie enveloping algebra of a pre-Lie algebra," *Journal of K-Theory* **2** no. 1, (May, 2008) 147–167.