Proceedings of Machine Learning Research vol 242:758–769, 2024 6th Annual Conference on Learning for Dynamics and Control

Rademacher Complexity of Neural ODEs via Chen–Fliess Series

Joshua Hanson JMH4@ILLINOIS.EDU
University of Illinois at Urbana-Champaign, 1308 W Main St, Urbana, IL 61801

Maxim Raginsky MAXIM@ILLINOIS.EDU
University of Illinois at Urbana-Champaign, 1308 W Main St, Urbana, IL 61801

Editors: A. Abate, M. Cannon, K. Margellos, A. Papachristodoulou

Abstract

We show how continuous-depth neural ODE models can be framed as single-layer, infinite-width nets using the Chen–Fliess series expansion for nonlinear ODEs. In this net, the output "weights" are taken from the signature of the control input — a tool used to represent infinite-dimensional paths as a sequence of tensors — which comprises iterated integrals of the control input over a simplex. The "features" are taken to be iterated Lie derivatives of the output function with respect to the vector fields in the controlled ODE model. The main result of this work applies this framework to derive compact expressions for the Rademacher complexity of ODE models that map an initial condition to a scalar output at some terminal time. The result leverages the straightforward analysis afforded by single-layer architectures. We conclude with some examples instantiating the bound for some specific systems and discuss potential follow-up work.

Keywords: Neural ODE, Chen–Fliess series, Rademacher complexity, generalization bound

1. Introduction

Several recent works have examined continuous-depth idealizations of deep neural nets, viewing them as continuous-time ordinary differential equation (ODE) models with either fixed or time-varying parameters. Traditional discrete-layer nets can be recovered by applying an appropriate temporal discretization scheme, e.g., the Euler or Runge-Kutta methods. In applications, this perspective has resulted in advantages concerning regularization (Kelly et al., 2020; Pal et al., 2021; Kobyzev et al., 2021), efficient parameterization (Queiruga et al., 2020), convergence speed (Chen et al., 2023), applicability to non-uniform data (Sahin and Kozat, 2019), among others. As a theoretical tool, continuous-depth idealizations have lead to better understanding of the contribution of depth to model expressiveness and generalizability (Massaroli et al., 2020; Marion, 2023), new or improved training strategies via framing as an optimal control problem (Corbett and Kangin, 2022), and novel model variations (Jia and Benson, 2019; Peluchetti and Favaro, 2020).

Considered as generic control systems, continuous-depth nets can admit a number of distinct input-output configurations depending on how the control system "anatomy" is delegated. Controlled neural ODEs (Kidger et al., 2020) and continuous-time recurrent neural nets (Fermanian et al., 2021) treat the (time-varying) control signal as the input to the model; the initial condition is either fixed or treated as a trainable parameter; the (time-varying) output signal is the model output; and any free parameters of the vector fields (weights) are held constant in time. One may instead consider the initial condition to be the input; the output signal at a fixed terminal time as the model output; and the (fixed or time-varying) control signal as a representative for (depth-varying) model

parameters, which yields a typical neural ODE (Chen et al., 2018). Here the input is a static finite-dimensional vector rather than a sequence or function of time, and this is the setting we consider in this work. Recent work of Marion (2023), discussed below, also considers this setting.

New results and insight have emerged from studying neural nets in the infinite-width or "mean-field" limit (Lu et al., 2020; Jacot et al., 2018). We can apply similar methods to neural ODEs by representing them using an infinite series expansion. One such example is the Chen–Fliess series (Chen, 1957; Fliess, 1981), which represents the output of a system as a sum of iterated Lie derivatives of the output function multiplied by corresponding iterated integrals of the control input, eliminating any recursive dependence on the state. This sequence of iterated integrals of the control is called the signature, which has been used in rough path theory for approximating and reconstructing stochastic signals using finite-dimensional data (Fermanian et al., 2023), and has also appeared in the control literature as a tool for studying small-time asymptotic behavior of control trajectories (Sussmann, 1983). Expressed as an infinite series using this formalism, the small-time initial-condition-to-output map is linear in the iterated integrals of the control input (i.e., elements of the input signature), and the remaining terms depend only on the initial condition.

The Chen–Fliess series can be interpreted as an infinitely wide, single-layer neural net where each node uses a different activation function computing the appropriate iterated Lie derivative in the series. In this net, the input weights are set to unity and the output weights are the corresponding signature elements. Applying series expansions for nonlinear ODEs in this way allows us to analyze continuous-depth nets using well-established techniques for single-layer, infinite-width nets, which are often comparatively simpler. Our goal in this work is to demonstrate a compact generalization bound for neural ODEs using these techniques. Marion (2023) also gives bounds on the generalization error of neural ODEs using covering number estimates for parameterized ODE model classes. By contrast, the generalization bounds in this work made use of Rademacher complexities and can complement those of Marion (2023).

The remainder of the paper is organized as follows. In Section 2, we describe the Chen–Fliess series and how it is used to generate a tractable model architecture. Section 3 defines the learning problem, then states the main Rademacher complexity bound and proof. In Section 4 we give some concrete examples to instantiate the bound, with conclusions and discussion provided in Section 5.

2. Chen–Fliess series

We are interested in maps $\varphi : \mathsf{X} \subset \mathbb{R}^n \to \mathsf{Y} \subset \mathbb{R}$ that can be described by sending an initial condition $x_0 \in \mathsf{X}$ to the resulting output $y(T) \in \mathsf{Y}$ at time T of a fixed control system

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)), \qquad x(0) = x_0 \\ y(t) &= h(x(t), u(t)) \end{aligned} \tag{1}$$

with a fixed control input $u : [0, T] \to \mathsf{U}$, where U is an arbitrary subset of some finite-dimensional vector space.

2.1. Control-affine systems

Consider the generic nonlinear control system (1). For the purposes of this work, we can restrict our attention to control-affine systems with linear outputs, which we will justify below. Assume that $f(\cdot, u) : \mathbb{R}^n \to \mathbb{R}^n$ is continuous for every $u \in \mathsf{U}$. Then we can always find continuous vector fields

$g_1, \ldots, g_m : \mathbb{R}^n \to \mathbb{R}^n$ such that $\{f(x,u) : u \in \mathsf{U}\} \subseteq \operatorname{span}\{g_1(x), \ldots, g_m(x)\}$ for every $x \in \mathsf{X}$ (Sussmann, 2008). Now for each $x \in \mathbb{R}^n$ define

$$G := \begin{bmatrix} g_1(x) & \cdots & g_m(x) \end{bmatrix} \in \mathbb{R}^{n \times m}, \qquad P := \begin{bmatrix} I & 0 \end{bmatrix} \in \mathbb{R}^{r \times m},$$

where $r \leq m$ is the rank of G. Assume without loss of generality that $g_{r+1}(x) = \cdots = g_m(x) = 0$, so then $GP^{\mathsf{T}} \in \mathbb{R}^{n \times r}$ has rank r. By construction, $f(x,u)$ is in the column space of G, thus for each $u \in \mathsf{U}$ there exists $v \in \mathbb{R}^m$ such that $Gv = f(x,u)$. In this case, we can take

$$v = P^{\mathsf{T}}(PG^{\mathsf{T}}GP^{\mathsf{T}})^{-1}PG^{\mathsf{T}}f(x,u).$$

Then the trajectory of the control-affine system

$$\dot{x}(t) = \sum_{i=1}^{m} v_i(t) g_i(x(t)), \qquad x(0) = x_0 \tag{2}$$

is that same as the trajectory of (1). Considered as an open-loop control system, the set of admissible solutions of (2) subsumes the set of admissible solutions of (1). The map $(x,u) \mapsto v$ may be discontinuous, but this is without consequence. From another angle that avoids state feedback, we could have instead considered appending the dynamics with an input integrator to yield

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)), \qquad x(0) = x_0 \\ \dot{u}(t) &= v(t), \end{aligned} \tag{3}$$

which is another control-affine system with the same trajectory as (1), provided that u is at least weakly differentiable. Furthermore, differentiating the output yields the equation

$$\dot{y}(t) = \frac{\partial h}{\partial x}(x(t), u(t)) f(x(t), u(t)) + \frac{\partial h}{\partial u}(x(t), u(t)) v(t),$$

which can be appended to the state dynamics (preceding the construction of (2) or (3) above) so that the output map becomes linear in the (augmented) state. Thus for the purposes of characterizing complexity, we restrict our attention to control-affine systems with linear output maps of the form

$$\begin{aligned} \dot{x}(t) &= f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t)), \qquad x(0) = x_0 \\ y(t) &= c^{\mathsf{T}} x(t), \end{aligned} \tag{4}$$

where $f, g_1, \ldots, g_m : \mathbb{R}^n \to \mathbb{R}^n$ and $c \in \mathbb{R}^n$. Lastly, we can disguise the drift if necessary by setting $g_0 \equiv f/M$ and $u_0 \equiv M$ for some constant $M \neq 0$, so without loss of generality we will only consider the driftless case (i.e., $f \equiv 0$) which will be convenient for certain calculations later.

2.2. Chen–Fliess series

To keep the paper self-contained, we give here a formal derivation of the Chen–Fliess series; see Sussmann (1983); Isidori (1995); Beauchard et al. (2023) for rigorous expositions, including the

analysis of convergence and truncation errors. By the fundamental theorem of calculus, the output at time $t \geq 0$ can be written

$$\begin{aligned} y(t) &= c^{\mathsf{T}} x_0 + \int_0^t c^{\mathsf{T}} \dot{x}(s) \, \mathrm{d}s \\ &= c^{\mathsf{T}} x_0 + \int_0^t c^{\mathsf{T}} \left(\sum_{i=1}^m u_i(s) g_i(x(s)) \right) \mathrm{d}s \\ &= c^{\mathsf{T}} x_0 + \sum_{i=1}^m \int_0^t u_i(s) L_{g_i} c^{\mathsf{T}} x(s) \, \mathrm{d}s. \end{aligned} \tag{5}$$

where $L_{g_i} c^{\mathsf{T}} x(s) = c^{\mathsf{T}} g_i(x(s))$ is the Lie derivative of the output function with respect to the vector field $g_i : \mathbb{R}^n \to \mathbb{R}^n$ at time $s \geq 0$. Using a similar trick, we can rewrite this Lie derivative as

$$\begin{aligned} L_{g_i} c^{\mathsf{T}} x(s) &= L_{g_i} c^{\mathsf{T}} x_0 + \int_0^s c^{\mathsf{T}} \frac{\partial g_i}{\partial x}(x(r)) \dot{x}(r) \, \mathrm{d}r \\ &= L_{g_i} c^{\mathsf{T}} x_0 + \int_0^s c^{\mathsf{T}} \frac{\partial g_i}{\partial x}(x(r)) \left(\sum_{j=1}^m u_j(r) g_j(x(r)) \right) \mathrm{d}r \\ &= L_{g_i} c^{\mathsf{T}} x_0 + \sum_{j=1}^m \int_0^s u_j(r) L_{g_j} \circ L_{g_i} c^{\mathsf{T}} x(r) \, \mathrm{d}r, \end{aligned}$$

where $\frac{\partial g_i}{\partial x}(x(r)) \in \mathbb{R}^{n \times n}$ is the Jacobian of g_i evaluated at $x(r)$. Substituting this into (5) gives

$$\begin{aligned} y(t) &= c^{\mathsf{T}} x_0 + \sum_{i=1}^m \int_0^t u_i(s) \left(L_{g_i} c^{\mathsf{T}} x_0 + \sum_{j=1}^m \int_0^s u_j(r) L_{g_j} \circ L_{g_i} c^{\mathsf{T}} x(r) \, \mathrm{d}r \right) \mathrm{d}s \\ &= c^{\mathsf{T}} x_0 + \sum_{i=1}^m \left(\int_0^t u_i(s) \, \mathrm{d}s \right) L_{g_i} c^{\mathsf{T}} x_0 + \sum_{1 \leq i,j \leq m} \int_0^t \int_0^s u_i(s) u_j(r) L_{g_j} \circ L_{g_i} c^{\mathsf{T}} x(r) \, \mathrm{d}r \, \mathrm{d}s. \end{aligned} \tag{6}$$

Repeating this process for $L_{g_j} \circ L_{g_i} c^{\mathsf{T}} x(r)$ in (6) and for the resulting higher order Lie derivatives generates the so-called *Chen–Fliess series*

$$y(t) = \sum_{\substack{1 \leq i_1, \ldots, i_k \leq m \\ k \geq 0}} \left(\int_0^t \int_0^{\tau_k} \cdots \int_0^{\tau_2} u_{i_k}(\tau_k) \cdots u_{i_1}(\tau_1) \, \mathrm{d}\tau_1 \cdots \mathrm{d}\tau_{k-1} \, \mathrm{d}\tau_k \right) \left(L_{g_{i_1}} \circ \cdots \circ L_{g_{i_k}} c^{\mathsf{T}} x_0 \right).$$

It will be convenient later to have a more compact expression for this series. Denote the set of multi-indices by $\mathsf{W} := \{w = (i_1, \ldots, i_k) : 1 \leq i_1, \ldots, i_k \leq m, \ k \geq 0\}$. For a multi-index $w = (i_1, \ldots, i_k)$, let $L_w := L_{g_{i_1}} \circ \cdots \circ L_{g_{i_k}}$ and $u_w(\tau) := u_{i_1}(\tau_1) \cdots u_{i_k}(\tau_k)$. The region of integration is a k-simplex, which we denote by $\Delta^k(t) := \{(\tau_1, \ldots, \tau_k) : 0 \leq \tau_1 \leq \cdots \leq \tau_k \leq t\}$. Now we can write

$$y(t) = \sum_{w \in \mathsf{W}} \left(\int_{\Delta^{|w|}(t)} u_w(\tau) \, \mathrm{d}\tau \right) L_w c^{\mathsf{T}} x_0 \tag{7}$$

where $|w|$ is the length of the multi-index w. The term corresponding to the empty multi-index is simply the constant $c^{\mathsf{T}} x_0$.

2.3. Sequence-space embeddings

Consider a space of bounded, measurable inputs $\mathcal{U} \subset \{u : [0,T] \to \mathbb{R}^m : u_i(t) \in [-M, M]\}$. We can embed this space of functions $\mathcal{U}$ into the space of real-valued sequences $\mathbb{R}^{\mathsf{W}}$ via the so-called *signature* $S : \mathcal{U} \to \mathbb{R}^{\mathsf{W}}$. For each $w \in \mathsf{W}$, define

$$S^w(u) := \int_{\Delta^{|w|}(T)} u_w(\tau)\,\mathrm{d}\tau.$$

Observe that $|S^w(u)| \leq \frac{(MT)^{|w|}}{|w|!}$, since the integrand is bounded by $|u_w(\tau)| \leq M^{|w|}$ and the volume of the $|w|$-simplex $\Delta^{|w|}(T)$ is $\frac{T^{|w|}}{|w|!}$.

Now consider a compact set of initial conditions $\mathsf{X} \subset \mathbb{R}^n$. In the same manner that $\mathcal{U}$ is embedded into a sequence space, we can also embed X into a sequence space via a map $\Phi : \mathsf{X} \to \mathbb{R}^{\mathsf{W}}$ that computes iterated Lie derivatives of the output map. For each $w \in \mathsf{W}$, define

$$\Phi^w(x) := L_w c^{\mathsf{T}} x.$$

The embeddings S and Φ pair naturally to recover the Chen–Fliess series (7) concisely as

$$y(T) = \big\langle S(u), \Phi(x_0) \big\rangle = \sum_{w \in \mathsf{W}} S^w(u)\Phi^w(x_0). \tag{8}$$

This representation of the output $y(T)$ admits a natural interpretation as a linear combination of nonlinear "features" $\Phi^w(x_0)$, where the "weights" are precisely the signature elements $S^w(u)$, which is a well-understood model architecture in learning theory. It is important to recognize that this (formal) series may fail to converge unless the time horizon T is sufficiently short, the control magnitude M is sufficiently small, and/or certain regularity assumptions on the vector fields $g_1, \ldots, g_m$ are satisfied. Such conditions are needed so that the map $x_0 \mapsto y(T)$ is well-defined. In a later section we will consider sufficient assumptions to guarantee convergence.

3. Main result

Suppose we have a sample of i.i.d. random vectors $(X_1, Y_1), \ldots, (X_N, Y_N)$ drawn according to a probability measure μ with compact support $\operatorname{supp}(\mu) = \mathsf{X} \times \mathsf{Y} \subset \mathbb{R}^n \times \mathbb{R}$. We seek to identify a function $\varphi : \mathsf{X} \to \mathsf{Y}$ that approximately reproduces this sample and generalizes to other identically distributed samples. Consider a class of such functions $\mathcal{F}$ given by

$$\mathcal{F} = \Big\{ x \mapsto \varphi(x) = \big\langle S(u), \Phi(x) \big\rangle : u \in \mathcal{U} \Big\}.$$

The vector fields $g_1, \ldots, g_m$ implicit in the definition of Φ are considered to be fixed, and the learnable parameters are represented by the control input u (or equivalently, the elements of the signature $S(u)$). Given a loss function $\ell : \mathsf{Y} \times \mathsf{Y} \to \mathbb{R}$, define the *expected risk*

$$L_\mu(\varphi) := \mathbf{E}_\mu\left[\ell(Y, \varphi(X))\right] = \int_{\mathsf{X}\times\mathsf{Y}} \ell(y, \varphi(x))\mu(\mathrm{d}x, \mathrm{d}y),$$

the *minimum risk* $L_\mu^*(\mathcal{F}) := \inf_{\varphi\in\mathcal{F}} L_\mu(\varphi)$, and the *empirical risk* $\frac{1}{N}\sum_{i=1}^N \ell(Y_i,\varphi(X_i))$. The empirical risk minimization (ERM) algorithm can be stated succinctly as

$$\hat{\varphi} \in \arg\min_{\varphi\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N \ell(Y_i,\varphi(X_i)),$$

and this minimization problem is usually solved numerically using e.g., gradient descent or its variations. Assuming that $0 \le \ell(y,\varphi(x)) \le B$ for all $(x,y) \in \mathsf{X}\times\mathsf{Y}$, $\varphi\in\mathcal{F}$, then the following excess risk guarantee holds with probably at least $1-\delta$ (see e.g., Hajek and Raginsky (2021)):

$$L_\mu(\hat{\varphi}) - L_\mu^*(\mathcal{F}) \le 4\mathbf{E}\mathcal{R}_N(\ell\circ\mathcal{F}) + B\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{N}}, \tag{9}$$

where the quantity $\mathcal{R}_N(\ell\circ\mathcal{F})$ is the empirical Rademacher complexity conditioned on the data. This is given by

$$\mathcal{R}_N(\ell\circ\mathcal{F}) := \mathbf{E}_\epsilon\left[\sup_{\varphi\in\mathcal{F}}\frac{1}{N}\left|\sum_{i=1}^N \epsilon_i\ell(Y_i,\varphi(X_i))\right|\right], \tag{10}$$

where the expectation is taken with respect to the sequence $\epsilon_1,\ldots,\epsilon_N$ of i.i.d. Rademacher random variables that are independent of the data. We can see that if the right-hand side of (9) is small, then the expected risk of the ERM map $\hat{\varphi}$ is close to the minimum risk, in which case we would say that $\hat{\varphi}$ generalizes well. Hence to study generalizability of $\mathcal{F}$, we seek to bound $\mathcal{R}_N(\ell\circ\mathcal{F})$.

If the loss function ℓ is well-behaved, we can often bound $\mathcal{R}_N(\ell\circ\mathcal{F})$ directly in terms of $\mathcal{R}_N(\mathcal{F})$. For instance, let $\ell(y,\varphi(x)) = (y-\varphi(x))^2$ and assume that $\sup_{y\in\mathsf{Y}}|y| \le M_1$ and $\sup_{x\in\mathsf{X}}\sup_{\varphi\in\mathcal{F}}|\varphi(x)| \le M_2$. Then by observing that $\big(y-\varphi(x)\big) \mapsto \ell(y,\varphi(x))$ is $2(M_1+M_2)$-Lipschitz and using the contraction principle (Ledoux and Talagrand, 1991), we have

$$\mathcal{R}_N(\ell\circ\mathcal{F}) \le 4\left(M_1+M_2\right)\left(\frac{M_1}{\sqrt{N}} + \mathcal{R}_N(\mathcal{F})\right).$$

If instead the loss is given by $\ell(y,\varphi(x)) = |y-\varphi(x)|$, which is 1-Lipschitz as a function of $\big(y-\varphi(x)\big)$, then using the contraction principle gives

$$\mathcal{R}_N(\ell\circ\mathcal{F}) \le \frac{2M_1}{\sqrt{N}} + 2\mathcal{R}_N(\mathcal{F}).$$

With this in mind, for the remainder of this section we will focus on bounding $\mathcal{R}_N(\mathcal{F})$.

Theorem 1 *The empirical Rademacher complexity of $\mathcal{F}$ is bounded by*

$$\mathcal{R}_N(\mathcal{F}) \le \frac{1}{\sqrt{N}}\sum_{k\ge 0}\frac{(mMT)^k}{k!}\Lambda_k, \tag{11}$$

where $\Lambda_k := \sup\left\{\left|L_w c^{\mathsf{T}} x\right| : x\in\mathsf{X},\ w\in\mathsf{W},\ |w|=k\right\}$.

Proof We will use the following lemma (we omit the proof, which is an elementary application of Jensen's inequality):

Lemma 2 *Let $\psi : \mathsf{X} \to \mathbb{R}$ be an arbitrary function. Then*

$$\mathbf{E}_\epsilon \left[\left| \sum_{i=1}^{N} \epsilon_i \psi(X_i) \right| \right] \leq \sqrt{N} \sup_{x \in \mathsf{X}} |\psi(x)|. \tag{12}$$

The Rademacher complexity of $\mathcal{F}$ is defined by

$$\mathcal{R}_N(\mathcal{F}) = \frac{1}{N} \mathbf{E}_\epsilon \left[\sup_{\varphi \in \mathcal{F}} \left| \sum_{i=1}^{N} \epsilon_i \varphi(X_i) \right| \right] = \frac{1}{N} \mathbf{E}_\epsilon \left[\sup_{u \in \mathcal{U}} \left| \sum_{i=1}^{N} \epsilon_i \sum_{w \in \mathsf{W}} S^w(u) \Phi^w(X_i) \right| \right].$$

We can bound the expression inside the expectation using the triangle inequality:

$$\begin{aligned} \sup_{u \in \mathcal{U}} \left| \sum_{i=1}^{N} \epsilon_i \sum_{w \in \mathsf{W}} S^w(u) \Phi^w(X_i) \right| &\leq \sup_{u \in \mathcal{U}} \sum_{w \in \mathsf{W}} \left| S^w(u) \right| \left| \sum_{i=1}^{N} \epsilon_i \Phi^w(X_i) \right| \\ &\leq \sum_{w \in \mathsf{W}} \frac{(MT)^{|w|}}{|w|!} \left| \sum_{i=1}^{N} \epsilon_i L_w c^{\mathsf{T}} X_i \right|. \end{aligned}$$

Applying Lemma 2 with $\psi \leftarrow L_w c^{\mathsf{T}}$ gives

$$\mathbf{E}_\epsilon \left[\left| \sum_{i=1}^{N} \epsilon_i L_w c^{\mathsf{T}} X_i \right| \right] \leq \sqrt{N} \sup_{x \in \mathsf{X}} |L_w c^{\mathsf{T}} x|.$$

Putting everything together, we have

$$\begin{aligned} \mathcal{R}_N(\mathcal{F}) &\leq \frac{1}{N} \mathbf{E}_\epsilon \left[\sup_{u \in \mathcal{U}} \left| \sum_{i=1}^{N} \epsilon_i \sum_{w \in \mathsf{W}} S^w(u) \Phi^w(X_i) \rangle \right| \right] \\ &\leq \frac{1}{N} \mathbf{E}_\epsilon \left[\sum_{w \in \mathsf{W}} \frac{(MT)^{|w|}}{|w|!} \left| \sum_{i=1}^{N} \epsilon_i L_w c^{\mathsf{T}} X_i \right| \right] \\ &\leq \frac{1}{N} \sum_{w \in \mathsf{W}} \frac{(MT)^{|w|}}{|w|!} \mathbf{E}_\epsilon \left[\left| \sum_{i=1}^{N} \epsilon_i L_w c^{\mathsf{T}} X_i \right| \right] \\ &\leq \frac{1}{\sqrt{N}} \sum_{w \in \mathsf{W}} \frac{(MT)^{|w|}}{|w|!} \sup_{x \in \mathsf{X}} \left| L_w c^{\mathsf{T}} x \right| \\ &\leq \frac{1}{\sqrt{N}} \sum_{k \geq 0} \frac{(mMT)^k}{k!} \Lambda_k. \end{aligned}$$

We are able to exchange the order of the infinite sum and the expectation above using Tonelli's theorem, because the summand/integrand are uniformly non-negative and measurable, and the rest follows from definitions. ∎

Now instantiating the Rademacher complexity bound comes down to bounding the norm of iterated Lie derivatives Λ_k, which we explore in some examples in the following section.

4. Examples

In the following examples, let $r := \sup_{x \in \mathsf{X}} |x|$ and assume $|c^\mathsf{T}| = 1$. The output is always given by $y(t) = c^\mathsf{T} x(t)$. We continue to use M, T as they are as defined in Section 2.3. Recall that some conditions are needed on M, T and/or the vector fields $g_1, \ldots, g_m$ so that the series (8) converges and the map $x_0 \mapsto y(T)$ is well-defined. Looking at error estimates for the Chen–Fliess expansion (Beauchard et al., 2023), it appears natural to suggest a condition like $\Lambda_k \leq C^k k!$ for some constant $C > 0$ depending on $M, T, g_1, \ldots, g_m$, which yields a geometric series. However for some systems, this condition is overly restrictive and we can in fact achieve convergence for any M, T even if this condition is violated. On the other hand, this condition excludes certain systems of interest that still admit convergent series expansions for some M, T which we will see in the following examples.

Example 1 Consider the class of bilinear systems

$$\dot{x}(t) = \left(\sum_{i=1}^m A_i u_i(t) \right) x(t), \qquad x(0) = x_0$$

where $A_1, \ldots, A_m \in \mathbb{R}^{n \times n}$. Let $a := \max_{i=1,\ldots,m} \sigma_{\max}(A_i)$ be the maximum spectral norm of the matrices $A_1, \ldots, A_m$. The Lie derivative of a linear function with respect to a linear vector field is simple to compute, leading to the following bound:

$$\begin{aligned} \Lambda_k &= \sup \left\{ \left| L_w c^\mathsf{T} x \right| : x \in \mathsf{X},\ w \in \mathsf{W},\ |w| = k \right\} \\ &= \sup \left\{ \left| c^\mathsf{T} A_{i_1} \cdots A_{i_k} x \right| : x \in \mathsf{X},\ w = (i_1, \ldots, i_k) \right\} \\ &\leq r \max_{i_1, \ldots, i_k} \|A_{i_1}\| \cdots \|A_{i_k}\| \leq r a^k. \end{aligned}$$

Substituting this into Theorem 1 yields

$$\mathcal{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}} \sum_{k \geq 0} \frac{(mMT)^k}{k!} r a^k = \frac{r}{\sqrt{N}} \exp\left(mMTa\right),$$

which is defined for all M, T.

Example 2 Consider the class of control-affine systems

$$\dot{x}(t) = \sum_{i=1}^m u_i(t) g_i(x(t)), \qquad x(0) = x_0$$

where $g_1, \ldots, g_m : \mathbb{R}^n \to \mathbb{R}^n$ are analytic vector fields. Let $\tilde{g}_1, \ldots, \tilde{g}_m : \mathbb{C}^n \to \mathbb{C}^n$ represent analytic continuations of $g_1, \ldots, g_m$ and let $\tilde{L}_w := L_{\tilde{g}_{i_1}} \circ \cdots \circ L_{\tilde{g}_{i_k}}$. Denote a closed polydisc by

$$P(\xi, \rho) := \{(z_1, \ldots, z_n) \in \mathbb{C}^n : |z_i - \xi_i| \leq \rho,\ 1 \leq i \leq n\}.$$

It is evident that $\iota(\mathsf{X}) \subset P(\iota(x), 2r)$ for any $x \in \mathsf{X}$, where $\iota : \mathbb{R}^n \hookrightarrow \mathbb{C}^n$ is the inclusion map. Define the component-wise maximum modulus of the complex vector fields $\tilde{g}_1, \ldots, \tilde{g}_m$ by

$$a(r) := \max_{i=1,\ldots,m} \max_{j=1,\ldots,n} \sup_{x \in \mathsf{X}} \sup_{z \in P(\iota(x), 2r)} |\tilde{g}_i^j(z)|,$$

where $\tilde{g}_i^j : \mathbb{C}^n \to \mathbb{C}$ is the jth component of $\tilde{g}_i : \mathbb{C}^n \to \mathbb{C}^n$. We can apply Lemma 3.8 in Lesiak and Krener (1978), which is based on Cauchy estimates from complex analysis, to bound Λ_k as follows:

$$\begin{aligned}
\Lambda_k &= \sup\left\{\left|L_w c^{\mathsf{T}} x\right| : x \in \mathsf{X},\ w \in \mathsf{W},\ |w| = k\right\} \\
&\leq \sup\left\{\left|\tilde{L}_w c^{\mathsf{T}} \iota(x)\right| : x \in \mathsf{X},\ z \in P(\iota(x), 2r),\ w \in \mathsf{W},\ |w| = k\right\} \\
&= \sup\left\{\left|\left(\sum_{j=1}^n \tilde{g}_{i_k}^j(z)\frac{\partial}{\partial z_j}\right)\cdots\left(\sum_{j=1}^n \tilde{g}_{i_1}^j(z)\frac{\partial}{\partial z_j}\right) c^{\mathsf{T}}\iota(x)\right| : x \in \mathsf{X},\ z \in P(\iota(x), 2r),\ 1 \leq i_1, \ldots, i_k \leq m\right\} \\
&\leq k!\left(\frac{2^n n a(r)}{r}\right)^k \left(1 + 2\sqrt{n}\right) r
\end{aligned}$$

where we have used that $\left|c^{\mathsf{T}} z\right| \leq |c|\left(\left|\iota(x)\right| + \sqrt{(2r)^2 n}\right) \leq \left(1 + 2\sqrt{n}\right) r$ for $z \in P(\iota(x), 2r)$. Assuming that $2^n nmMTa(r) < r$, substituting this into Theorem 1 yields

$$\begin{aligned}
\mathcal{R}_N(\mathcal{F}) &\leq \frac{\left(1 + 2\sqrt{n}\right) r}{\sqrt{N}} \sum_{k \geq 0} \frac{(mMT)^k}{k!} k! \left(\frac{2^n n a(r)}{r}\right)^k \\
&= \frac{\left(1 + 2\sqrt{n}\right) r}{\sqrt{N}} \sum_{k \geq 0} \left(\frac{2^n nmMTa(r)}{r}\right)^k \\
&= \frac{\left(1 + 2\sqrt{n}\right) r}{\sqrt{N}} \frac{r}{r - 2^n nmMTa(r)}.
\end{aligned}$$

Example 3 Consider a class of Hopfield nets

$$\dot{x}(t) = u(t)\sigma(x(t)) = \sum_{1 \leq i,j \leq n} u_{ij}(t)\sigma(x_j(t))e_i,$$

where $e_i \in \mathbb{R}^n$ is the ith unit vector, $u : [0, T] \to \mathbb{R}^{n \times n}$ is a matrix-valued control input, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a sigmoidal nonlinearity. Suppose the derivatives of σ satisfy the bound $\sup_{x \in \mathbb{R}} |\sigma^{(k)}(x)| \leq b a^k k!$ for some $a, b > 0$, which holds for many common sigmoidal activation functions. If $k = 0$, then $\Lambda_k \leq r$, so suppose $k \geq 1$. Then

$$\begin{aligned}
\Lambda_k &= \sup\left\{\left|L_w c^{\mathsf{T}} x\right| : x \in \mathsf{X},\ w \in \mathsf{W},\ |w| = k\right\} \\
&= \sup\left\{\left|\left(\sigma(x_{j_1}) e_{i_1}^{\mathsf{T}} \frac{\partial}{\partial x}\right)\cdots\left(\sigma(x_{j_k}) e_{i_k}^{\mathsf{T}} \frac{\partial}{\partial x}\right) c^{\mathsf{T}} x\right| : x \in \mathsf{X},\ 1 \leq i_1, j_1, \ldots, i_k, j_k \leq n\right\} \\
&= \sup\left\{\left|\left(\sigma(x_{j_1}) \frac{\partial}{\partial x_{i_1}}\right)\cdots\left(\sigma(x_{j_k}) \frac{\partial}{\partial x_{i_k}}\right) c^{\mathsf{T}} x\right| : x \in \mathsf{X},\ 1 \leq i_1, j_1, \ldots, i_k, j_k \leq n\right\} \\
&\leq \sup\left\{\binom{k}{n_1, \ldots, n_k}\left|\sigma^{(n_1)}(x)\cdots\sigma^{(n_k)}(x)\right| : x \in [-r, r],\ n_1 + \cdots + n_k = k - 1\right\} \\
&\leq \sup\left\{\binom{k}{n_1, \ldots, n_k}(b a^{n_1} n_1!)\cdots(b a^{n_k} n_k!) : n_1 + \cdots + n_k = k - 1\right\} \\
&\leq \gamma(k) b^k a^{k-1},
\end{aligned}$$

where $\gamma(k) = \frac{k!}{2}\binom{2k}{k}$, which comes from a counting argument; see Appendix B.5 in Fermanian et al. (2021). Assuming that $4n^2MTba < 1$, substituting this into Theorem 1 yields

$$\begin{aligned}\mathcal{R}_N(\mathcal{F}) &\leq \frac{1}{\sqrt{N}}\left(r + \sum_{k\geq 1}\frac{\left(n^2MT\right)^k}{k!}\frac{k!}{2}\binom{2k}{k}b^k a^{k-1}\right)\\ &= \frac{1}{\sqrt{N}}\left(r + \frac{1}{2a}\sum_{k\geq 1}\binom{2k}{k}\left(n^2MTba\right)^k\right)\\ &= \frac{1}{\sqrt{N}}\left(r - \frac{1}{2a} + \frac{1}{2a\sqrt{1-4n^2MTba}}\right).\end{aligned}$$

The final expression above comes from the generating function for the central binomial coefficients

$$\frac{1}{\sqrt{1-4x}} = \sum_{k\geq 0}\binom{2k}{k}x^k,$$

which can be derived by applying the generalized binomial theorem with $n = -\frac{1}{2}$.

5. Conclusion

Using the Chen–Fliess series for nonlinear ODEs, we have shown how continuous-depth nets (a.k.a. neural ODEs) can be viewed as a kind of single-layer, infinite-width net, where the "weights" are the iterated integrals (signature elements) of the control input, and the "features" are the iterated Lie derivatives of the output function. This approach facilitates compact expressions for the generalization performance of ODE models based on the comparatively simpler analysis of single-layer architectures. These bounds are also straightforward to instantiate given various assumptions about the structure of the ODE model, which we have demonstrated through some examples.

One barrier to applying this technique in more generality is that the Chen–Fliess series converges only for sufficiently small time horizons and/or small control magnitudes. One could attempt to circumvent this issue by dividing the time horizon into slices and considering the composition of several convergent series expansions of the flow map, possibly later taking the limit as the number of slices increases to infinity. However, this sacrifices the convenience of working with a single-layer architecture, as bounding the Rademacher complexity of composite function classes is typically either challenging or yields conservative results. An alternative approach to generalize the main result is to interpret the control input as a perturbation of some nominal control trajectory, and instead focus on obtaining margin bounds, which is an interesting direction for follow-up work.

Incorporating any special structure known about the system of interest would also likely give sharper results in cases where it applies. For example, the result here is agnostic to any information concerning system stability or dissipativity. Instead of bounding the Lie derivatives of the vector fields directly, one could likely obtain tighter bounds in the stable case by using the logarithmic norm of the iterated Jacobians of the vector fields instead of the operator norm, for instance, or otherwise specializing the analysis to incorporate any behavioral or structural knowledge of the ODE model under consideration.

Acknowledgments

This work was supported in part by the NSF under award CCF-2106358 ("Analysis and Geometry of Neural Dynamical Systems") and in part by the Illinois Institute for Data Science and Dynamical Systems (iDS^2), an NSF HDR TRIPODS institute, under award CCF-1934986.

References

Karine Beauchard, Jérémy Le Borgne, and Frédéric Marbach. On expansions for nonlinear systems error estimates and convergence issues. *Comptes Rendus. Mathématique*, 361(G1):97–189, 2023.

Kuo-Tsai Chen. Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula. *The Annals of Mathematics*, 65(1):163, 1957.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

Shuangshuang Chen, Sihao Ding, Yiannis Karayiannidis, and Mårten Björkman. Learning continuous normalizing flows for faster convergence to target distribution via ascent regularizations. In *The Eleventh International Conference on Learning Representations*, 2023.

Andrew Corbett and Dmitry Kangin. Imbedding deep neural networks. In *International Conference on Learning Representations*, 2022.

Adeline Fermanian, Pierre Marion, Jean-Philippe Vert, and Gérard Biau. Framing RNN as a kernel method: A neural ODE approach. In *Advances in Neural Information Processing Systems*, 2021.

Adeline Fermanian, Terry Lyons, James Morrill, and Cristopher Salvi. New directions in the applications of rough path theory. *IEEE BITS The Information Theory Magazine*, pages 1–18, 2023.

Michel Fliess. Fonctionnelles causales non linéaires et indéterminées non commutatives. *Bulletin de la Société Mathématique de France*, 109:3–40, 1981.

Bruce Hajek and Maxim Raginsky. Statistical learning theory. http://maxim.ece.illinois.edu/teaching/SLT/SLT.pdf, 2021.

Alberto Isidori. *Nonlinear Control Systems*. Springer-Verlag, New York, third edition, 1995.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Junteng Jia and Austin R Benson. Neural jump stochastic differential equations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jacob Kelly, Jesse Bettencourt, Matthew James Johnson, and David Duvenaud. Learning differential equations that are easy to solve. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, pages 6696–6707. Curran Associates, Inc., 2020.

Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2021.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

C. Lesiak and A. Krener. The existence and uniqueness of Volterra series for nonlinear systems. *IEEE Transactions on Automatic Control*, 23(6):1090–1095, 1978.

Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean field analysis of deep ResNet and beyond: Towards provable optimization via overparameterization from depth. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6426–6436. PMLR, 13–18 Jul 2020.

Pierre Marion. Generalization bounds for neural ordinary differential equations and deep residual networks. In *Advances in Neural Information Processing Systems*, 2023.

Stefano Massaroli, Michael Poli, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. Dissecting neural ODEs. In *Advances in Neural Information Processing Systems*, volume 33, pages 3952–3963. Curran Associates, Inc., 2020.

Avik Pal, Yingbo Ma, Viral Shah, and Christopher V Rackauckas. Opening the blackbox: Accelerating neural differential equations by regularizing internal solver heuristics. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8325–8335. PMLR, 18–24 Jul 2021.

Stefano Peluchetti and Stefano Favaro. Infinitely deep neural networks as diffusion processes. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1126–1136. PMLR, 26–28 Aug 2020.

Alejandro F. Queiruga, N. Benjamin Erichson, Dane Taylor, and Michael W. Mahoney. Continuous-in-depth neural networks, 2020.

Safa Onur Sahin and Suleyman Serdar Kozat. Nonuniformly sampled data processing using LSTM networks. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5):1452–1461, 2019.

Héctor J. Sussmann. Lie brackets and local controllability: A sufficient condition for scalar-input systems. *SIAM Journal on Control and Optimization*, 21(5):686–713, 1983.

Héctor J. Sussmann. Smooth distributions are globally finitely spanned. *Analysis and Design of Nonlinear Control Systems*, pages 3–8, 2008.