Proceedings of Machine Learning Research vol 144:1–11, 2021

# Learning Recurrent Neural Net Models of Nonlinear Systems

**Joshua Hanson** JMH4@ILLINOIS.EDU
*University of Illinois*
*Urbana, IL 61801*

**Maxim Raginsky** MAXIM@ILLINOIS.EDU
*University of Illinois*
*Urbana, IL 61801*

**Eduardo Sontag** E.SONTAG@NORTHEASTERN.EDU
*Northeastern University*
*Boston, MA 02115*

## Abstract

We consider the following learning problem: Given sample pairs of input and output signals generated by an unknown nonlinear system (which is not assumed to be causal or time-invariant), we wish to find a continuous-time recurrent neural net with hyperbolic tangent activation function that approximately reproduces the underlying i/o behavior with high confidence. Leveraging earlier work concerned with matching output derivatives up to a given finite order (Sontag, 1998), we reformulate the learning problem in familiar system-theoretic language and derive quantitative guarantees on the sup-norm risk of the learned model in terms of the number of neurons, the sample size, the number of derivatives being matched, and the regularity properties of the inputs, the outputs, and the unknown i/o map.

**Keywords:** empirical risk minimization, recurrent neural nets, dynamical systems, continuous time, system identification, statistical learning theory, generalization bounds

## 1. Introduction

We consider a learning-theoretic framework for system identification, where the goal is to approximate an unknown nonlinear input/output (i/o) map by a continuous-time recurrent neural net (RNN) on the basis of a finite collection of input/output pairs. The approximation criterion is the $\mathcal{L}^\infty$ norm of the difference between the ground-truth output and the one predicted by the learned model.

Earlier work (Sontag, 1998) has addressed a related problem of reproducing output $k$-jets as a function of input $(k-1)$-jets via RNNs, where a $k$*-jet* is defined as the vector of derivatives up to order $k$ of the output (respectively, input) signal evaluated at the initialization time. Equivalently, $k$-jets can be defined as optimal degree-$k$ polynomial approximations, i.e., truncated Taylor series. In this work, we build on the aforementioned result about matching $k$-jets, but work in a familiar system-theoretic setting using the language of i/o maps.

The training procedure used both in that work and here differs to some extent from the standard "backpropagation through time" algorithm often used for training RNNs. Rather than forward-propagating training inputs through the net or setting up an adjoint equation for the net parameters, we "pull back" the input signals to the initialization time and compute the corresponding output $k$-jets, which can be expressed explicitly as a function of the weights of the neural net, its initial state, and the $(k-1)$-jets of the input. The loss function is then evaluated with the predicted and ground-truth output jets. At training time, there is no requirement to advance inputs through the network. In essence, the RNN can be trained using conventional nonlinear regression as if it were a single-layer feedforward net.

In the following section, we describe the learning problem, state our assumptions, and introduce the needed notation. Proceeding in Section 3, we develop the setting of jets and use this language to formulate the proposed learning procedure and state our main results. Some directions for future research are also outlined. Section 4.1 contains some technical lemmas, followed by the proof of the main theorems in Sections 4.2 and 4.3.

## 2. Problem formulation

In this work, a *system* (or an *i/o map*) is a nonlinear operator $\mathsf{F} : C([0,T]) \to C([0,T])$, where $C([0,T])$ is the Banach space of continuous functions $u : [0,T] \to \mathbb{R}$ equipped with the sup norm

$$\|u\|_\infty := \sup_{t\in[0,T]} |u(t)|.$$

The learning problem can be phrased as follows: Let $N$ pairs $(u^1, y^1), \dots, (u^N, y^N)$ be given, where $u^i \in C([0,T])$ are the inputs and $y^i = \mathsf{F}u^i \in C([0,T])$ are the corresponding outputs. We wish to construct a system $\hat{\mathsf{F}} : C([0,T]) \to C([0,T])$ that approximately reproduces the unknown system $\mathsf{F}$ on a given class of inputs $\mathcal{U} \subset C([0,T])$. (We focus on single-input, single-output systems mainly for simplicity; our approach easily extends to multiple inputs and outputs.) Next, we impose a set of minimal assumptions on the system $\mathsf{F}$ and specify the accuracy criterion.

We first recall the definition of the *modulus of continuity* of a function $u \in C([0,T])$ (see, e.g., Ch. 2 of DeVore and Lorentz (1993)):

$$\omega_u(\delta) := \sup_{\substack{t_1,t_2\in[0,T]\\ |t_1-t_2|\le\delta}} |u(t_1) - u(t_2)|.$$

The function $\delta \mapsto \omega_u(\delta)$ is nondecreasing, and, since any $u \in C([0,T])$ is uniformly continuous, $\lim_{\delta\downarrow 0} \omega_u(\delta) = \omega_u(0) = 0$. (In fact, we will refer to any function with these properties that majorizes $\omega_u$ as a modulus of continuity of $u$.) We impose the following assumption on the class of inputs $\mathcal{U}$:

**Assumption 1** *The inputs in $\mathcal{U}$ are uniformly bounded, i.e., $R := \sup_{u\in\mathcal{U}} \|u\|_\infty < \infty$, and equicontinuous with common modulus of continuity $\omega_{\mathcal{U}}(\delta)$:*

$$\sup_{u\in\mathcal{U}} \omega_u(\delta) \le \omega_{\mathcal{U}}(\delta),$$

*where $\omega_{\mathcal{U}}(\delta)$ is a nondecreasing function that satisfies $\lim_{\delta\downarrow 0} \omega_{\mathcal{U}}(\delta) = \omega_{\mathcal{U}}(0) = 0$.*

We also use moduli of continuity to describe the regularity of $\mathsf{F}$:

**Assumption 2** *The output space $\mathcal{Y} = \mathsf{F}(\mathcal{U})$, i.e., the image of $\mathcal{U}$ under $\mathsf{F}$, is equicontinuous with a common modulus of continuity $\omega_{\mathcal{Y}}(\delta)$.*

Finally, we assume that $\mathsf{F}$ has the bounded-in, bounded-out property:

**Assumption 3** $\gamma_{\mathsf{F}}(R) := \sup_{\substack{u \in C[0,T] \\ \|u\|_\infty \le R}} \|\mathsf{F}u\|_\infty < \infty.$

Examples of classes of inputs that satisfy Assumption 1 include:

- finite combinations of Fourier terms
$$u(t) = \sum_{i=1}^{m} c_i \sin(\omega_i t + \alpha_i), \tag{1}$$
for all $m \ge 1$, provided the coefficients $c_i$ and the frequencies $\omega_i$ satisfy the inequalities $\sum_{i=1}^m |c_i| \le R$ and $\sum_{i=1}^m |c_i \omega_i| \le L$ for some fixed finite constants $R$ and $L$;
- polynomial inputs
$$u(t) = \sum_{i=0}^{m} c_i t^i \tag{2}$$
for all $m \ge 0$, provided the coefficients $c_i$ satisfy the inequalities $\sum_{i=0}^m |c_i| T^i \le R$ and $\sum_{i=1}^m i|c_i|T^{i-1} \le L$ for some $R, L < \infty$.

In both cases, Assumption 1 holds with $\omega_{\mathcal{U}}(\delta) = L\delta$. As an example of a system satisfying our hypotheses, consider a state-space model of the form

$$\begin{aligned} \dot{x}(t) &= f(x(t)) + g(x(t))u(t), \\ y(t) &= h(x(t)) \end{aligned}$$

where $x(t) \in \mathbb{R}^n$ is an internal state with a given initial condition $x(0) = \xi$. Let $\mathsf{F}$ be the induced i/o map that sends the input $u : [0,T] \to \mathbb{R}$ to the output $y : [0,T] \to \mathbb{R}$. Then it is not hard to verify that Assumptions 2 and 3 will hold if the functions $f : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^n$, and $h : \mathbb{R}^n \to \mathbb{R}$ are all Lipschitz continuous.

We assume the sample inputs $u^i$ are independent and identically distributed (i.i.d.) random elements of $C([0,T])$ drawn according to a fixed Borel probability measure $\mu$ supported on $\mathcal{U}$. Thus, the input-output pairs $(u^1, y^1), \dots, (u^N, y^N)$, with $y^i = \mathsf{F}u^i$, are themselves i.i.d. random elements of the product space $\mathcal{U} \times \mathcal{Y}$. For instance, $\mathsf{F}$ could be a model of a two-terminal electronic device whose terminals are connected through a switch to an excitation circuit consisting of linear and nonlinear elements and current and/or voltage sources (Chua, 1980). Suppose this excitation circuit is specified by a vector of parameters $\theta \in \mathbb{R}^d$; e.g., it could be used to generate sinusoidal or polynomial inputs, as in (1) or (2). We can then generate $N$ i.i.d. samples $\theta^1, \dots, \theta^N$ according to a fixed Borel probability measure $\nu$ on $\mathbb{R}^d$. Each sample $\theta^i$ corresponds to a random realization of the excitation circuit so that, when the switch is closed at time $t = 0$ and open at time $t = T$, we can take the input $u^i : [0,T] \to \mathbb{R}$ to be the voltage waveform across $\mathsf{F}$ and the output $y^i : [0,T] \to \mathbb{R}$ to be the current waveform through $\mathsf{F}$. (This assumes that $\mathsf{F}$ is voltage-controlled.) Thus, the probability measure on the input-output space $\mathcal{U} \times \mathcal{Y}$ is well-defined but specified *indirectly* through $\nu$, the structure of the excitation circuit, and $\mathsf{F}$. At any rate, given $\mathsf{F}$ and an approximating system $\hat{\mathsf{F}}$, we define the $\mathcal{L}^\infty$ risk

$$\mathcal{L}(\hat{\mathsf{F}}) := \mathbf{E}_\mu\big[\|\hat{\mathsf{F}}u - \mathsf{F}u\|_\infty\big],$$

where the expectation is taken with respect to the probability measure $\mu$ on the input space $\mathcal{U}$. The goal is to generate, on the basis of the observed input-output pairs $(u^i, y^i)$, an approximate system $\hat{\mathsf{F}}$ from a given model class $\mathcal{F}$, so that the risk $\mathcal{L}(\hat{\mathsf{F}})$ (which is a random variable due to its dependence on the training data) is small with high probability.

## 3. The proposed learning procedure and its performance

**Recurrent neural nets.** We first specify the model class $\mathcal{F}$ from which our learning procedure will select the approximation $\hat{\mathsf{F}}$. The systems in $\mathcal{F}$ are described by differential equations of the form

$$\dot{x}(t) = \sigma^{(n)}(Ax(t) + bu(t)) \tag{3a}$$

$$y(t) = c^T x(t) \tag{3b}$$

for $t \in [0, T]$ with initial condition $x(0) = \xi \in \mathbb{R}^n$. Here, $A \in \mathbb{R}^{n \times n}$, $b, c \in \mathbb{R}^n$, and $\sigma^{(n)} : \mathbb{R}^n \to \mathbb{R}^n$ is the diagonal map $\sigma^{(n)}((x_1, \dots, x_n)^T) := (\sigma(x_1), \dots, \sigma(x_n))^T$ with $\sigma(r) = \tanh(r)$. The system (3) is a continuous-time *recurrent neural net* (RNN) with $n$ neurons. Each pair $(\Sigma, \xi)$, where $\Sigma := (A, b, c)$, specifies an input-output map $\mathsf{F}_{\Sigma,\xi}$ that sends an input $u \in C([0, T])$ to the corresponding output $y$ given by

$$y(t) = c^T \xi + \int_0^t c^T \sigma^{(n)}(Ax(\tau) + bu(\tau))\, \mathrm{d}\tau, \qquad t \in [0, T].$$

For $0 < M < \infty$, we define the class of systems

$$\mathcal{F}(M) := \left\{ \mathsf{F}_{\Sigma,\xi} = \mathsf{F}_{(A,b,c),\xi} : \|A\|, |b|, |c|, |\xi| \leq M \right\},$$

where $\|A\|$ is the spectral norm of $A$, and $|b|, |c|, |\xi|$ are the $\ell^2$ norms of $b, c, \xi$. Our learning procedure will generate a pair $(\hat{\Sigma}, \hat{\xi})$ based on the data $\{(u^i, y^i)\}$, and output the model $\hat{\mathsf{F}} = \mathsf{F}_{\hat{\Sigma},\hat{\xi}} \in \mathcal{F}(M)$. With a slight abuse of notation, we will often write $(\Sigma, \xi) \in \mathcal{F}(M)$ instead of $\mathsf{F}_{\Sigma,\xi} \in \mathcal{F}(M)$.

**Jets.** While it is well-known that RNNs of the form (3) are universal approximators for i/o maps $\mathsf{F}$ that admit smooth nonlinear state-space realizations $\dot{x} = f(x, u)$, $y = g(x)$ (Sontag, 1992; Funahashi and Nakamura, 1993; Hanson and Raginsky, 2020), here we are *not* assuming that $\mathsf{F}$ admits such a realization (in fact, we are not even requiring $\mathsf{F}$ to be causal or time-invariant). Nevertheless, we will show that, provided Assumptions 1–3 hold for our learning problem, we will be able to approximate $\mathsf{F}$ by a recurrent net model which will have the properties of causality and time invariance by construction. Our approach proceeds by way of reducing the infinite-dimensional problem of learning the i/o map $\mathsf{F}_{\hat{\Sigma},\hat{\xi}}$ to a certain finite-dimensional problem (Sontag, 1998). To that end, consider a system of the form (3) fed with an input $u$ which has at least $k - 1$ derivatives at $t = 0$. Then the output $y = \mathsf{F}_{\Sigma,\xi} u$ will have at least $k$ derivatives at $t = 0$, which can be computed explicitly as

$$y^{(0)}(0) = c^T \xi, \quad y^{(1)}(0) = c^T \sigma^{(n)}(A\xi + bu(0)), \quad \dots.$$

We can then define the map $Y_{k,\Sigma,\xi} : \mathbb{R}^k \to \mathbb{R}^{k+1}$ according to

$$Y_{k,\Sigma,\xi}((u(0), u'(0), \dots, u^{(k-1)}(0))^T) := (y(0), y'(0), \dots, y^{(k)}(0))^T,$$

where $y^{(\ell)}(0) = \frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell}\big|_{t=0} \mathsf{F}_{\Sigma,\xi}u(t)$ for $0 \le \ell \le k$. We can also phrase this in terms of *jets*, where the $k$-jet at $t = 0$ of a function $f : \mathbb{R} \to \mathbb{R}$ which is $C^k$ in some neighborhood of $t = 0$ is the degree-$k$ polynomial

$$J_0^k f(s) := \sum_{\ell=0}^{k} f^{(\ell)}(0) \frac{s^\ell}{\ell!}.$$

Then, for inputs that are $C^{k-1}$ in some neighborhood of $t = 0$, the map $Y_{k,\Sigma,\xi} : \mathbb{R}^k \to \mathbb{R}^{k+1}$ can be lifted to a map from $(k-1)$-jets to $k$-jets via $J_0^{k-1}u \mapsto J_0^k \mathsf{F}_{\Sigma,\xi}u$, where the vector of coefficients of $J_0^k \mathsf{F}_{\Sigma,\xi}u$ is the image of the vector of coefficients of $J_0^{k-1}u$ under $Y_{k,\Sigma,\xi}$. Thus, at least for inputs $u$ that are of class $C^{k-1}$, the i/o map $\mathsf{F}_{\Sigma,\xi}$ will be a good approximation to $\mathsf{F}$ provided (a) the outputs $\mathsf{F}u$ and $\mathsf{F}_{\Sigma,\xi}u$ can be accurately approximated by their $k$-jets $J_0^k\mathsf{F}u$ and $J_0^k\mathsf{F}_{\Sigma,\xi}u$ (e.g., if $k$ is sufficiently large) and (b) the $k$-jets $J_0^k\mathsf{F}u$ and $J_0^k\mathsf{F}_{\Sigma,\xi}u$ are close to one another in sup norm on $[0,T]$.

**The learning procedure.** Since the inputs in $\mathcal{U}$ and the corresponding outputs in $\mathcal{Y}$ are only assumed to be continuous, we will first approximate them by functions that have as many derivatives at $t = 0$ as needed. To that end, we start by defining for each $k \in \mathbb{N}$ two linear maps $\mathsf{S}_k : C([0,T]) \to \mathbb{R}^k$ and $\mathsf{S}_k^* : \mathbb{R}^k \to C([0,T])$ according to

$$\mathsf{S}_k(u) := \left( B_{k-1}(u,0), \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} B_{k-1}(u,t), \dots, \frac{\mathrm{d}^{k-1}}{\mathrm{d}t^{k-1}}\Big|_{t=0} B_{k-1}(u,t) \right)^{\scriptscriptstyle T},$$

and

$$\mathsf{S}_k^*((a_0,\dots,a_{k-1})^{\scriptscriptstyle T})(t) := \sum_{\ell=0}^{k-1} a_\ell \frac{t^\ell}{\ell!},$$

where

$$B_m(u,t) := \sum_{i=0}^{m} u\left(\frac{iT}{m}\right) \binom{m}{i} \left(\frac{t}{T}\right)^i \left(1 - \frac{t}{T}\right)^{m-i}, \qquad t \in [0,T]$$

is the degree-$m$ *Bernstein polynomial* of $u \in C([0,T])$ (DeVore and Lorentz, 1993, Ch. 1). Whenever no confusion will arise, we will also write $B_m u$ instead of $B_m(u,\cdot)$. Observe that $\mathsf{S}_k \circ \mathsf{S}_k^* = \mathrm{id}_{\mathbb{R}^k}$ and $(\mathsf{S}_k^* \circ \mathsf{S}_k)u = B_{k-1}(u,\cdot)$. To motivate the introduction of these objects, we give the following bound on the expected risk of any recurrent net model $\mathsf{F}_{\Sigma,\xi}$:

**Theorem 4** *Under Assumptions 1 and 2, the following holds for any $\Sigma = (A,b,c)$, $\xi$, and $k \ge 2$:*

$$\begin{aligned}\mathcal{L}(\mathsf{F}_{\Sigma,\xi}) \le 2\omega_{\mathcal{Y}}\left(\frac{T}{\sqrt{k}}\right) + 2|c||b|e^{\|A\|T}\omega_{\mathcal{U}}\left(\frac{2T}{\sqrt{k}}\right) + |c|Te^{\|A\|T}\sqrt{\frac{n}{k}} \\ + \mathbf{E}_\mu\left[\|(\mathsf{S}_{k+1}^* \circ Y_{k,\Sigma,\xi} \circ \mathsf{S}_k)u - (\mathsf{S}_{k+1}^* \circ \mathsf{S}_{k+1} \circ \mathsf{F})u\|_\infty\right].\end{aligned} \tag{4}$$

Note that the last term on the right-hand side of (4) is the expectation, with respect to $u \sim \mu$, of the sup norm of the difference between the degree-$k$ Bernstein polynomial $B_k\mathsf{F}u$ and the degree-$k$ Bernstein polynomial $B_k\mathsf{F}_{\Sigma,\xi}B_{k-1}u$.

We are now ready to present our learning procedure. For $1 \le j \le k$, let $t_j := jT/k$, and consider the following Empirical Risk Minimization (ERM) scheme:

$$(\hat{\Sigma}, \hat{\xi}) \in \operatorname*{arg\,min}_{(\Sigma,\xi)\in\mathcal{F}(M)} \frac{1}{N}\sum_{i=1}^{N} \max_{1\le j\le k} \left|(\mathsf{S}^*_{k+1}\circ Y_{k,\Sigma,\xi}\circ \mathsf{S}_k)u(t_j) - (\mathsf{S}^*_{k+1}\circ \mathsf{S}_{k+1})y(t_j)\right|. \tag{5}$$

The objective being minimized in (5) is simply the empirical expectation of the maximum absolute difference between the Bernstein polynomials $B_k\mathsf{F}u$ and $B_k\mathsf{F}_{\Sigma,\xi}B_{k-1}u$ on the finite grid $\{T/k, 2T/k, \dots, (k-1)T/k, T\} \subset [0,T]$.

**Theorem 5** *Suppose Assumptions 1–3 are satisfied, and $N \ge k(6n^6 + 10n^3\log_2 k)$. Then with probability at least $1-\delta$,*

$$\begin{aligned}\mathcal{L}(\hat{\mathsf{F}}) \le 4\omega_{\mathcal{Y}}\left(\frac{T}{\sqrt{k}}\right) + 2M^2e^{MT}\omega_{\mathcal{U}}\left(\frac{2T}{\sqrt{k}}\right) + 3MTe^{MT}\sqrt{\frac{n}{k}}\\ + \bar{\mathcal{L}}^* + c\left(M(M+\sqrt{n}T) + \gamma_{\mathsf{F}}(R)\right)\sqrt{\frac{k(n^6+n^3\log_2 k)\log N + \log(\frac{1}{\delta})}{N}},\end{aligned} \tag{6}$$

*where $c > 0$ is an absolute constant and*

$$\bar{\mathcal{L}}^* := \inf_{(\Sigma,\xi)\in\mathcal{F}(M)} \mathbf{E}_\mu\left[\max_{1\le j\le k}\left|(\mathsf{S}^*_{k+1}\circ Y_{k,\Sigma,\xi}\circ \mathsf{S}_k)u(t_j) - (\mathsf{S}^*_{k+1}\circ \mathsf{S}_{k+1})y(t_j)\right|\right]. \tag{7}$$

The first three terms in (6) can be made arbitrarily small by choosing $k$ sufficiently large. The exponential dependence on $M$ and $T$ is an unavoidable artifact of the Grönwall lemma, and can be removed under appropriate stability assumptions (Hanson and Raginsky, 2020). The remaining two terms are, respectively, the approximation error and the estimation error of the ERM procedure (5). While the latter can be made arbitrarily small by increasing the size $N$ of the training set, the former is an intrinsic measure of the ability of recurrent neural nets to approximate output $k$-jets for a randomly chosen input. This minimal error value can be decreased by considering larger, more expressive nets; a quantitative analysis of this term is a promising direction for further work. Note also that, for a fixed $k$, we only need to collect input and output samples on an equispaced grid $\{0, T/k, 2T/k, \dots, T\}$. On the other hand, in many applications (e.g., medical or electronic system modeling), the input/output data are often available at non-uniformly spaced times $0 \le t_1 < t_2 < \dots < t_k \le T$ that are not under the learner's control. Extending the approach of this paper to non-uniform (or even random) sampling of time instants is another interesting future direction.

## 4. Proofs

### 4.1. Technical lemmas

In this section, we collect a few results that will be used in the proofs of Theorems 4 and 5. The proofs of the lemmas are omitted due to space limitations, and can be found in the full version of this paper (Hanson et al., 2021).

**Lemma 6** *For any $u \in C([0,T])$ with modulus of continuity $\omega$ and any $k \in \mathbb{N}$, the Bernstein polynomial $B_k(u,\cdot)$ has modulus of continuity $2\omega$ and satisfies*

$$\|u - B_k(u,\cdot)\|_\infty \le 2\omega\left(\frac{T}{\sqrt{k}}\right). \tag{8}$$

**Lemma 7** *Let* $\mathsf{G} : C([0,T]) \to C([0,T])$ *be a Lipschitz-continuous i/o map, i.e.,*

$$\|\mathsf{G}\|_{\mathrm{Lip}} := \sup_{\substack{u_1,u_2 \in C([0,T]) \\ u_1 \neq u_2}} \frac{\|\mathsf{G}u_1 - \mathsf{G}u_2\|_\infty}{\|u_1 - u_2\|_\infty} < \infty.$$

*Then, for any integer* $k \ge 2$ *and any* $u \in C([0,T])$*,*

$$\|\mathsf{G}u - (\mathsf{S}^*_{k+1} \circ \mathsf{S}_{k+1} \circ \mathsf{G} \circ \mathsf{S}^*_k \circ \mathsf{S}_k)u\|_\infty \le 2\|\mathsf{G}\|_{\mathrm{Lip}}\omega_u\left(\frac{2T}{\sqrt{k}}\right) + 2\omega_{\mathsf{G}B_{k-1}u}\left(\frac{T}{\sqrt{k}}\right).$$

**Lemma 8** *Let* $\mathsf{G}$ *be the i/o map of the differential dynamical system*

$$\dot{x}(t) = f(x(t), u(t)), \quad y(t) = h^{\scriptscriptstyle T} x(t); \qquad t \in [0,T],\ x(0) = \xi \tag{9}$$

*with input* $u(t) \in \mathbb{R}$*, state* $x(t) \in \mathbb{R}^n$*, and output* $y(t) \in \mathbb{R}$*. Suppose that* $f(\cdot,\cdot)$ *is bounded, i.e.,* $\sup_{(x,u)\in\mathbb{R}^n\times\mathbb{R}} |f(x,u)| < \infty$*, and Lipschitz-continuous, i.e.,*

$$|f(x_1,u_1) - f(x_2,u_2)| \le L_X|x_1 - x_2| + L_U|u_1 - u_2|.$$

*Then, for any input* $u \in C([0,T])$*, the output* $y = \mathsf{G}u$ *is bounded with*

$$\|y\|_\infty \le |h|\left(|\xi| + T\sup_{x,u}|f(x,u)|\right), \tag{10}$$

*and has modulus of continuity*

$$\omega_y(\delta) \le \frac{|h|}{L_X}\sup_{x,u}|f(x,u)|\left(e^{L_X\delta} - 1\right). \tag{11}$$

*Moreover, the i/o map* $\mathsf{G}$ *is Lipschitz-continuous, with*

$$\|\mathsf{G}\|_{\mathrm{Lip}} \le \frac{|h|L_U}{L_X}\left(e^{L_X T} - 1\right). \tag{12}$$

We also need the following result on the Rademacher averages of VC-subgraph classes (Farrell et al., 2020). Recall that a class $\mathcal{G}$ of measurable functions $g : \mathbb{R}^d \to \mathbb{R}$ is a *VC-subgraph class* (Giné and Nickl, 2016, Sec. 3.6.2) if the class of all sets of the form $\{(x,r) \in \mathbb{R}^d \times \mathbb{R} : g(x) \ge r\}$ with $g \in \mathcal{G}$ is a Vapnik–Chervonenkis (or VC) class, i.e., there exists a finite $D \in \mathbb{N}$, such that, for each $m \le D$, there exist $m$ points $(x^1,r^1),\dots,(x^m,r^m)$ that are *shattered by* $\mathcal{G}$, i.e.,

$$\{(\mathbf{1}_{\{g(x^1)\ge r^1\}},\dots,\mathbf{1}_{\{g(x^m)\ge r^m\}}) : g \in \mathcal{G}\} = \{0,1\}^m,$$

and no such $m$-tuple of points exists for $m > D$. This $D$ is called the *VC-subgraph dimension* (or *pseudo-dimension*) of $\mathcal{G}$, and is denoted by $\mathrm{vc}(\mathcal{G})$.

**Lemma 9** *Let* $\mathcal{G}$ *be VC subgraph class of real-valued measurable functions* $g : \mathbb{R}^d \to [0,B]$*. Let* $\boldsymbol{x} = (x^1,\dots,x^N)$ *be an arbitrary* $N$*-tuple of points in* $\mathbb{R}^d$*, and define the* Rademacher average

$$R_N(\mathcal{G};\boldsymbol{x}) := \frac{1}{N}\mathbf{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^N \varepsilon^i g(x^i)\right|\right],$$

*where* $\varepsilon^1,\dots,\varepsilon^N$ *are i.i.d. random variables with* $\mathbf{P}[\varepsilon^i = \pm 1] = \frac{1}{2}$*. Then, for any* $N \ge \mathrm{vc}(\mathcal{G})$*,*

$$R_N(\mathcal{G};\boldsymbol{x}) \le cB\sqrt{\frac{\mathrm{vc}(\mathcal{G})\log N}{N}}$$

*for some universal constant* $c$*.*

### 4.2. Proof of Theorem 4

Fix any $u \in \mathcal{U}$ and any $(\Sigma, \xi)$. Then

$$\begin{aligned}
&\|\mathsf{F}u - \mathsf{F}_{\Sigma,\xi}u\|_\infty \\
&\quad \le \|\mathsf{F}u - (\mathsf{S}^*_{k+1} \circ \mathsf{S}_{k+1} \circ \mathsf{F})u\|_\infty + \|\mathsf{F}_{\Sigma,\xi}u - (\mathsf{S}^*_{k+1} \circ \mathsf{S}_{k+1} \circ \mathsf{F}_{\Sigma,\xi} \circ \mathsf{S}^*_k \circ \mathsf{S}_k)u\|_\infty \\
&\qquad + \|(\mathsf{S}^*_{k+1} \circ \mathsf{S}_{k+1} \circ \mathsf{F}_{\Sigma,\xi} \circ \mathsf{S}^*_k \circ \mathsf{S}_k)u - (\mathsf{S}^*_{k+1} \circ \mathsf{S}_{k+1} \circ \mathsf{F})u\|_\infty \\
&\quad =: T_1 + T_2 + T_3.
\end{aligned} \tag{13}$$

Since $(\mathsf{S}^*_{k+1} \circ \mathsf{S}_{k+1})u = B_k u$, we can estimate $T_1$ using Lemma 6 and Assumption 2:

$$T_1 \le \sup_{u \in \mathcal{U}} \|\mathsf{F}u - B_k(\mathsf{F}u, \cdot)\|_\infty = \sup_{y \in \mathcal{Y}} \|y - B_k(y, \cdot)\|_\infty \le 2\omega_{\mathcal{Y}}\left(\frac{T}{\sqrt{k}}\right). \tag{14}$$

For $T_2$, Lemma 7 gives

$$T_2 \le 2\|\mathsf{F}_{\Sigma,\xi}\|_{\mathrm{Lip}}\omega_u\left(\frac{2T}{\sqrt{k}}\right) + 2\omega_{\mathsf{F}_{\Sigma,\xi}B_{k-1}(u,\cdot)}\left(\frac{T}{\sqrt{k}}\right) \tag{15}$$

Now, the system (3) has the form (9) with $f(x, u) = \sigma^{(n)}(Ax + bu)$ and $h = c$; thus, applying Lemma 8 to the i/o map $\mathsf{F}_{\Sigma,\xi}$ with $\Sigma = (A, b, c)$, we get

$$\|\mathsf{F}_{\Sigma,\xi}\|_{\mathrm{Lip}} \le |c||b|e^{\|A\|T} \qquad \text{and} \qquad \omega_{\mathsf{F}_{\Sigma,\xi}B_{k-1}(u,\cdot)}(\delta) \le \sqrt{n}|c|e^{\|A\|T}\delta.$$

Substituting these estimates into (15) and invoking Assumption 1, we obtain

$$T_2 \le 2|c||b|e^{\|A\|T}\omega_{\mathcal{U}}\left(\frac{2T}{\sqrt{k}}\right) + 2|c|Te^{\|A\|T}\sqrt{\frac{n}{k}}. \tag{16}$$

Finally, using the fact that $(\mathsf{S}^*_{k+1} \circ \mathsf{F}_{\Sigma,\xi} \circ \mathsf{S}^*_k \circ \mathsf{S}_k)u = (\mathsf{S}^*_{k+1} \circ Y_{k,\Sigma,\xi} \circ \mathsf{S}_k)u$, we can write

$$T_3 = \|(\mathsf{S}^*_{k+1} \circ Y_{k,\Sigma,\xi} \circ \mathsf{S}_k)u - (\mathsf{S}^*_{k+1} \circ \mathsf{S}_{k+1} \circ \mathsf{F})u\|_\infty. \tag{17}$$

Using (14), (16), and (17) in (13) and taking expectation with respect to $\mu$, we obtain (4).

### 4.3. Proof of Theorem 5

For each pair $(\Sigma, \xi)$, define the function $g_{\Sigma,\xi} : \mathbb{R}^k \times \mathbb{R}^{k+1} \to \mathbb{R}_+$ according to

$$g_{\Sigma,\xi}(v, z) := \max_{1 \le j \le k} \left|(\mathsf{S}^*_{k+1} \circ Y_{k,\Sigma,\xi})v(t_j) - \mathsf{S}^*_{k+1}z(t_j)\right|,$$

where $t_j = jT/k$. Let $\bar{\mu}$ denote the joint probability law of $\mathsf{S}_k u$ and $\mathsf{S}_{k+1}y$ when $u \sim \mu$ and $y = \mathsf{F}u$. Then $\bar{\mu}$ is a Borel probability measure on $\mathbb{R}^k \times \mathbb{R}^{k+1}$, and we can define the expected risk

$$\bar{\mathcal{L}}(\Sigma, \xi) := \mathbf{E}_{\bar{\mu}}[g_{\Sigma,\xi}(v, z)].$$

Given the i/o data $(u^i, y^i) \overset{\text{i.i.d.}}{\sim} \mu$, the points $(v^i, z^i)$ with $v^i = \mathsf{S}_k(u^i)$ and $z^i = \mathsf{S}_{k+1}(y^i)$ are i.i.d. samples from $\bar{\mu}$, and our learning procedure selects any minimizer $(\hat{\Sigma}, \hat{\xi})$ of the empirical risk

$$\bar{\mathcal{L}}_N(\Sigma, \xi) := \frac{1}{N}\sum_{i=1}^N g_{\Sigma,\xi}(v^i, z^i)$$

among all $\Sigma = (A, b, c)$ and $\xi$ satisfying the constraint $\|A\|, |b|, |c|, |\xi| \le M$.

We next show that, with high probability, the excess risk $\bar{\mathcal{L}}(\hat{\Sigma}, \hat{\xi}) - \bar{\mathcal{L}}^*$ is small, where the minimum risk $\bar{\mathcal{L}}^*$ is defined in (7). By Lemma 8 applied to any i/o map $\mathsf{F}_{\Sigma,\xi} \in \mathcal{F}(M)$, and for any $v \in C([0,T])$,

$$\|B_k \mathsf{F}_{\Sigma,\xi} v\|_\infty \le \|\mathsf{F}_{\Sigma,\xi} v\|_\infty \le |c| \left(|\xi| + \sqrt{n}T\right) \le M(M + \sqrt{n}T).$$

Moreover, for any $u \in \mathcal{U}$,

$$\|(\mathsf{S}^*_{k+1} \circ \mathsf{S}_{k+1} \circ \mathsf{F})u\|_\infty = \|B_k(\mathsf{F}u, \cdot)\|_\infty \le \|\mathsf{F}u\|_\infty \le \gamma_{\mathsf{F}}(R).$$

Therefore, we may assume without loss of generality that $0 \le g_{\Sigma,\xi}(\cdot) \le M(M + \sqrt{n}T) + \gamma_{\mathsf{F}}(R) =: B$. Then the usual ERM analysis (see, e.g., Cor. 6.1 in Hajek and Raginsky (2019)) guarantees that, with probability at least $1 - \delta$,

$$\bar{\mathcal{L}}(\hat{\Sigma}, \hat{\xi}) \le \bar{\mathcal{L}}^* + 4\mathbf{E}R_N(\mathcal{G}) + B\sqrt{\frac{2\log(\frac{1}{\delta})}{N}},$$

where $R_N(\mathcal{G}) = R_N(\mathcal{G}; ((v^1, z^1), \dots, (v^N, z^N)))$ is the Rademacher average of the function class $\mathcal{G} := \left\{g_{\Sigma,\xi} : (\Sigma, \xi) \in \mathcal{F}(M)\right\}$. Using Lemma 9, we then see that

$$\bar{\mathcal{L}}(\hat{\Sigma}, \hat{\xi}) \le \bar{\mathcal{L}}^* + cB\sqrt{\frac{\mathrm{vc}(\mathcal{G})\log N + \log(\frac{1}{\delta})}{N}} \tag{18}$$

with probability at least $1 - \delta$, provided $N \ge \mathrm{vc}(\mathcal{G})$, where $c > 0$ is an absolute constant and $\mathrm{vc}(\mathcal{G})$ is the VC-subgraph dimension (or pseudo-dimension) of $\mathcal{G}$.

**Pseudo-dimension estimate.** For any $v \in \mathbb{R}^k$ and $z = (z_0, \dots, z_k)^\intercal \in \mathbb{R}^{k+1}$, we can write

$$g_{\Sigma,\xi}(v, z) = \max_{j \in [k]} |h^{(j)}_{\Sigma,\xi}(v, z)|,$$

where

$$h^{(j)}_{\Sigma,\xi}(v, z) := \sum_{\ell=0}^{k} \left(y^{(\ell)}_{k,\Sigma,\xi}(v) - z_\ell\right)\frac{t_j^\ell}{\ell!},$$

with $y^{(\ell)}_{k,\Sigma,\xi}(v)$ denoting the $\ell$th coordinate of $Y_{k,\Sigma,\xi}(v)$. Let $\mathcal{H}^{(j)}$ denote the class of all functions of the form $h^{(j)}_{\Sigma,\xi}$. We make the following two claims:

1. $\mathcal{H}^{(1)}, \dots, \mathcal{H}^{(k)}$ are VC-subgraph classes with the same pseudo-dimension $d \le 3n^6 + 5n^3 \log_2 k$;
2. $\mathcal{G}$ is a VC-subgraph class with $\mathrm{vc}(\mathcal{G}) \le 2kd$.

We prove the second claim first. For each $j \in [k]$, consider the class $\mathcal{C}^{(j)}$ of subsets of $\mathbb{R}^k \times \mathbb{R}^{k+1} \times \mathbb{R}$ of the form

$$\left\{(v, z, r) \in \mathbb{R}^k \times \mathbb{R}^{k+1} \times \mathbb{R} : |h^{(j)}_{\Sigma,\xi}(v, z)| \ge r\right\}. \tag{19}$$

Then evidently each set $\{(v,z,r) : g_{\Sigma,\xi}(v,z) \ge r\}$ is of the form $C^{(1)} \cup \ldots \cup C^{(k)}$ with $C^{(j)} \in \mathcal{C}^{(j)}$. By the standard VC dimension estimates (Giné and Nickl, 2016, Prop. 3.6.7), then, $\mathrm{vc}(\mathcal{G}) \le \mathrm{vc}(\mathcal{H}^{(1)}) + \ldots + \mathrm{vc}(\mathcal{H}^{(k)})$. On the other hand, each set in (19) is itself the union of the sets $\{(v,z,r) : h^{(j)}_{\Sigma,\xi}(v,z) \ge r\}$ and $\{(v,z,r) : h^{(j)}_{\Sigma,\xi} \le -r\}$, and therefore $\mathrm{vc}(\mathcal{H}^{(j)}) \le 2d$. It remains to prove the first claim.

To that end, fix some $\Sigma = (A,b,c)$ and $\xi$ and let $\theta$ be a vector of dimension $n^2 + 3n$ obtained by listing the entries of $A, b, c, \xi$ in some fixed order. Then an induction argument together with the fact that the function $\sigma(r) = \tanh r$ satisfies the identity $\sigma'(r) = 1 - \sigma^2(r)$ can be used to show that each function $h^{(j)}_{\Sigma,\xi}$ can be written in the form

$$P(\sigma(R_1(\theta,v,z)), \ldots, \sigma(R_n(\theta,v,z)), \theta, v, z),$$

where $P$ is a polynomial of degree at most $3k-1$ and each $R_i$ is a polynomial of degree at most 2 (Sontag, 1998). This implies, in turn, that

$$\mathrm{vc}(\mathcal{H}^{(1)}) = \ldots = \mathrm{vc}(\mathcal{H}^{(k)}) \le 3n^6 + 5n^3 \log_2 k$$

(Sontag, 1998, Cor. 7). Substituting the resulting estimate of $\mathrm{vc}(\mathcal{G})$ into (18), we see that

$$\bar{\mathcal{L}}(\hat{\Sigma}, \hat{\xi}) \le \bar{\mathcal{L}}^* + c(M(M + \sqrt{n}T) + \gamma_{\mathsf{F}}(R)) \sqrt{\frac{k(n^6 + n^3 \log_2 k) \log N + \log(\frac{1}{\delta})}{N}} \tag{20}$$

with probability at least $1 - \delta$.

**The final risk bound.** For any $\Sigma, \xi$ and any $u \in \mathcal{U}$,

$$\begin{aligned}
&\|(\mathsf{S}^*_{k+1} \circ Y_{k,\Sigma,\xi} \circ \mathsf{S}_k)u - (\mathsf{S}^*_{k+1} \circ \mathsf{S}_{k+1} \circ \mathsf{F})u\|_\infty \\
&= \|B_k(\mathsf{F}_{\Sigma,\xi} B_{k-1}u, \cdot) - B_k(\mathsf{F}u, \cdot)\|_\infty \\
&= \sup_{t \in [0,T]} |B_k(\mathsf{F}_{\Sigma,\xi} B_{k-1}u, t) - B_k(\mathsf{F}u, t)| \\
&\le \max_{1 \le j \le k} |B_k(\mathsf{F}_{\Sigma,\xi} B_{k-1}u, t_j) - B_k(\mathsf{F}u, t_j)| \\
&\qquad + \sup_{t \in [0,T]} \min_{1 \le j \le k} |B_k(\mathsf{F}_{\Sigma,\xi} B_{k-1}u, t_j) - B_k(\mathsf{F}_{\Sigma,\xi} B_{k-1}u, t)| \\
&\qquad + \sup_{t \in [0,T]} \min_{1 \le j \le k} |B_k(\mathsf{F}u, t_j) - B_k(\mathsf{F}u, t)|,
\end{aligned}$$

where the last two terms can be upper-bounded using the moduli of continuity of $B_k(\mathsf{F}_{\Sigma,\xi} B_{k-1}u, \cdot)$ and $B_k(\mathsf{F}u, \cdot)$, which, by Lemma 6, are upper-bounded by twice the moduli of continuity of $\mathsf{F}_{\Sigma,\xi} B_{k-1}u$ and $\mathsf{F}u$, respectively. Thus, using Assumption 2 and Lemma 8, we get

$$\begin{aligned}
&\|(\mathsf{S}^*_{k+1} \circ Y_{k,\Sigma,\xi} \circ \mathsf{S}_k)u - (\mathsf{S}^*_{k+1} \circ \mathsf{S}_{k+1} \circ \mathsf{F})u\|_\infty \\
&\qquad \le \max_{1 \le j \le k} |B_k(\mathsf{F}_{\Sigma,\xi} B_{k-1}u, t_j) - B_k(\mathsf{F}u, t_j)| + \frac{2}{k}\sqrt{n} M T e^{MT} + 2\omega_{\mathcal{Y}}\left(\frac{T}{k}\right).
\end{aligned}$$

Taking expectation of both sides with respect to $\mu$ yields, for any $(\Sigma, \xi) \in \mathcal{F}(M)$,

$$\mathbf{E}_\mu\left[\|(\mathsf{S}^*_{k+1} \circ Y_{k,\Sigma,\xi} \circ \mathsf{S}_k)u - (\mathsf{S}^*_{k+1} \circ \mathsf{S}_{k+1} \circ \mathsf{F})u\|_\infty\right] \le \bar{\mathcal{L}}(\Sigma, \xi) + \frac{2}{k}\sqrt{n} M T e^{MT} + 2\omega_{\mathcal{Y}}\left(\frac{T}{k}\right).$$

Using this and (20) in the bound of Theorem 4, we get the statement of the theorem.

## Acknowledgments

The work of J. Hanson and M. Raginsky was supported in part by the Illinois Institute for Data Science and Dynamical Systems (iDS$^2$), an NSF HDR TRIPODS institute, under award CCF-1934986 and by the Center for Advanced Electronics through Machine Learning (CAEML) NSF I/UCRC award CNS-16-24811.

## References

Leon O. Chua. Device modeling via basic nonlinear circuit elements. *IEEE Transactions on Circuits and Systems*, CAS-27(11):1014–1044, November 1980.

Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*. Springer-Verlag, New York, 1993.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 2020. URL https://arxiv.org/abs/1809.09953. To appear.

Ken-Ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6(6):801 – 806, 1993.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

Bruce Hajek and Maxim Raginsky. ECE 543: Statistical Learning Theory. University of Illinois lecture notes, 2019. URL http://maxim.ece.illinois.edu/teaching/SLT/SLT.pdf.

Joshua Hanson and Maxim Raginsky. Universal simulation of stable dynamical systems by recurrent neural nets. In *Proc. 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 384–392, 2020.

Joshua Hanson, Maxim Raginsky, and Eduardo Sontag. Learning recurrent neural net models of nonlinear systems, 2021. URL https://arxiv.org/abs/2011.09573.

Eduardo D. Sontag. Neural nets as systems models and controllers. In *Proc. Seventh Yale Workshop on Adaptive and Learning Systems*, pages 73–79, 1992.

Eduardo D. Sontag. A learning result for continuous-time recurrent neural networks. *Systems and Control Letters*, 34:151–158, 1998.