Length independent generalization bounds for deep SSM architectures

Dániel Rácz, Mihály Petreczky, Bálint Daróczy

ICML 2024 Next Generation of Sequence Modeling Architectures Workshop 1–22

Length independent generalization bounds for deep SSM architectures

author names withheld

Under Review for NGSM 2024

Abstract

Many state-of-the-art models trained on long-range sequences, for example S4, S5 or LRU, are made of sequential blocks combining State-Space Models (SSMs) with neural networks. In this paper we provide a PAC bound that holds for these kind of architectures with *stable* SSM blocks and the bound does not depend on the length of the input sequence. Imposing stability of the SSM blocks is a standard practice in the literature, and it is known to help performance. Our results provide a theoretical justification for the use of stable SSM blocks as the proposed PAC bound decreases as the degree of stability of the SSM blocks increases.

1. Introduction

The problem of modeling long-range sequences, i.e. sequences with large number of time-steps, is an especially challenging task of the field. Recently, several novel architectures (etc. S4 Gu et al. (2021), S4D Gu et al. (2022), S5 Smith et al. (2022), LRU Orvieto et al. (2023)) have been published that are outperforming previous models by a significant margin. The common basis of these models are, combined with some nonlinearity, the so-called Structured State-Space Models (SSMs), which are basically dynamical systems of either continuous or discrete time. One key point of these models is that they are equipped with some form of stability constraints. This motivates the question: *What is the role of stability in the success of deep SSM architectures for long-range sequences?*

Contribution. In this paper, we focus on this question and provide a theoretical framework to analyze the model's generalization behavior in a rigorous manner by showing, to our knowledge, the first generalization bound for deep SSMs. We show that stability of deep SSM architectures has an influence on their Rademacher complexity, resulting in a generalization bound that does not depend on the length of the input sequence.

Related work. Bounds for general RNNs are related to SSMs as Linear Time-Invariant (LTI) dynamical systems are essential elements for almost all SSMs and they are a special class of RNNs. There are several PAC bounds for either discrete or continuous-time RNNs in Koiran and Sontag (1998); Sontag (1998); Hanson et al. (2021) by using VC dimension usually through covering numbers. PAC bounds for RNNs based on Rademacher complexity were presented in Wei and Ma (2019); Akpinar et al. (2020); Joukovsky et al. (2021); Chen et al. (2020), while in Zhang et al. (2018) the authors developed PAC-Bayesian bounds. As all of these results tend to infinity with the integration time (number of time steps) they are not meaningful in case of long-range sequences. In (Hanson and Raginsky, 2024) the authors propose a PAC bound based on Rademacher complexity for input-affine non-linear

systems, however their bound is still exponential in the length of the integration interval. The generalization bound for single vanilla RNNs in (Chen et al., 2020, Theorem 2) is an upper bound of the H_1 norm proposed in this paper, see e.g. Chellaboina et al. (1999), thus our results based on the H_2 norm is even tighter. In Golowich et al. (2018) the authors derived a depth independent bound under the condition of bounded Schatten p-norm and a bound with polynomial dependence on depth for Rademacher complexity for DNNs with ReLU activations by applying contraction. In a recent paper Truong (2022b) the author extends this bound for non ReLU activations and show that the new, non-vacuous bound is depth independent. Lastly, we mention that in Trauger and Tewari (2024) the authors propose a sequence length independent Rademacher complexity bound for a single layer transformer architecture. For multi layer transformers they improve slightly the result in Edelman et al. (2022) however the bound grows logarithmically with the sequence length.

2. Preliminaries

Σ denotes a dynamical system specified in the context. The constant n_{in} refers to the dimension of the input sequence, T refers to its length in time, while n_{out} is the dimension of the output (not necessarily a sequence). Denote by $\ell_T^{2,2}(\mathbb{R}^n)$ and $\ell_T^{\infty,\infty}(\mathbb{R}^n)$ the Banach spaces generated by the all finite sequences over $\mathbb{R}^n$ of length T with the norm $\|\mathbf{u}\|^2_{\ell_T^{2,2}(\mathbb{R}^n)} = \sum_{k=0}^{T-1} \|\mathbf{u}[k]\|_2^2$ and $\|\mathbf{u}\|_{\ell_T^{\infty,\infty}(\mathbb{R}^n)} = \sup_{k=0,\dots,T-1} \|\mathbf{u}[k]\|_\infty$ respectively. For a Banach space $\mathcal{X}$, $B_{\mathcal{X}}(r) = \{x \in \mathcal{X} \mid \|x\|_{\mathcal{X}} \leq r\}$ denotes the ball of radius $r > 0$ centered in zero.

Generalization gap. We consider the usual supervised learning framework for sequential input data. The considered models, parametrized by θ, are of the form $f_\theta : \ell_T^{2,2}(\mathbb{R}^{n_{\text{in}}}) \to \mathbb{R}^{n_{\text{out}}}$. In this paper, we are agnostic regarding the origin of θ. A dataset is an i.i.d sample of the form $S = \{(\mathbf{u}_i, \mathbf{y}_i)\}_{i=1}^N$ from some distribution $\mathcal{D}$ over $\ell_T^{2,2}(\mathbb{R}^{n_{\text{in}}}) \times \mathbb{R}$. An elementwise loss function is of the form $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Let $\mathcal{L}^S_{emp}(f) = \frac{1}{N}\sum_{i=1}^N \ell(f(\mathbf{u}_i), y_i)$ denote the empirical loss of a model f w.r.t a dataset S. We denote the true error by $\mathcal{L}(f) = \mathbf{E}_{(\mathbf{u},y)\sim\mathcal{D}}[\ell(f(\mathbf{u}), y)]$. The generalization error or gap of a model f is defined as $|\mathcal{L}^S_{emp}(f) - \mathcal{L}(f)|$.

SSMs. A *State-Space Model (SSM)* is a discrete-time linear dynamical system of the form

$$\Sigma \begin{cases} \mathbf{x}[k] = A\mathbf{x}[k-1] + B\mathbf{u}[k], \ \mathbf{x}[0] = 0 \\ \mathbf{y}[k] = C\mathbf{x}[k] + D\mathbf{u}[k] \end{cases} \tag{1}$$

where $A \in \mathbb{R}^{n_x \times n_x}, B \in \mathbb{R}^{n_x \times n_u}, C \in \mathbb{R}^{n_y \times n_x}$ and $D \in \mathbb{R}^{n_y \times n_u}$ are matrices, $\mathbf{u}[k], \mathbf{x}[k]$ and $\mathbf{y}[k]$ are the input, the state and the output signals respectively for $k = 1, 2, \dots, T$, where T is the number of time steps. We consider the value of T to be fixed to handle pooling. We emphasize that the generalization bound in Theorem 3 is independent of T.

Stability. We call the SSM (1) *stable*, if the matrix A is Schur, i.e., the moduli of all its eigenvalues are smaller than 1. Intuitively, stable SSMs are robust to perturbations, i.e., their state and outputs are continuous in the initial state and input, see for instance Antoulas (2005) for a more detailed discussion. In particular, a sufficient (but not necessary) condition for stability is that A is a contraction, i.e. $\|A\|_2 < 1$. For more details, see Appendix A.

Input-output maps of SSMs as operators on $\ell_T^{p,p}$,$p = \infty, 2$. An SSM (1) induces an input-output map, which maps every input sequence $\mathbf{u}[0], \dots, \mathbf{u}[T-1]$ to output sequences

$\mathbf{y}[0], \ldots, \mathbf{y}[T-1]$, which can be interpreted as a linear operator $\mathcal{S}_{\Sigma,T}$ from $\ell_T^{p,p}(\mathbb{R}^{n_u}) \to \ell_T^{r,r}(\mathbb{R}^{n_y})$, for any choice $p, r \in \{\infty, 2\}$. In particular, $\mathcal{S}_{\Sigma,T}$ has a well-defined induced norm as a linear operator, defined in the usual way, $\|\mathcal{S}_{\Sigma,T}\|_{p,r} = \sup_{\mathbf{u} \in \ell_T^{p,p}(\mathbb{R}^{n_u})} \frac{\|\mathcal{S}_{\Sigma,T}(\mathbf{u})\|_{\ell_T^{r,r}(\mathbb{R}^{n_y})}}{\|\mathbf{u})\|_{\ell_T^{r,r}(\mathbb{R}^{n_y})}}$. If Σ is internally stable, then it is a standard result in control theory that the norms $\{\|\mathcal{S}_{\Sigma,T}\|\}_{T=1}^{\infty}$ are bounded, i.e., $\|\Sigma\|_{p,r} = \sup_{T>0} \|\mathcal{S}_{\Sigma,T}\|_{p,r}$ exists and it is finite, see Antoulas (2005). Moreover, $\|\Sigma\|_{p,r}$ can be viewed as the norm of the extension of the input-output operators $\mathcal{S}_{\Sigma,T}$ to the Banach space generated by $\{\ell_T^{p,p}(\mathbb{R}^{n_u}\}_{T=1}^{\infty}$ Antoulas (2005). In this paper, we will use the norms $\|\Sigma\|_{2,\infty}$ and $\|\Sigma\|_{\infty,\infty}$ to upper bound the Rademacher complexity. For more details about the norms see Appendix A.

Relationship with continuous-time models. SSMs are often derived by discretizing a continuous-time linear differential equation in time (e.g. Gu and Dao (2023) and references therein). If the discretization step Δ is a fixed constant, then we obtain time-invariant linear system of the form (1). Now let us define a single discrete-time SSM block.

Definition DT-SSM. *A DT-SSM block (or simply SSM block) is a function $f^{DTB} : \ell_T^{p,q}(\mathbb{R}^{n_u}) \to \ell_T^{r,s}(\mathbb{R}^{n_u})$ that is composed of a stable SSM followed by a nonlinear transformation that is constant in time. That is, $f^{DTB}(\mathbf{u})[k] = g(\mathcal{S}_{\Sigma,T}(\mathbf{u})[k]) + \alpha \mathbf{u}[k]$ for some $\alpha \in [0,1]$ and $g : \mathbb{R}^{n_u} \to \mathbb{R}^{n_u}$ for all $k \in [T]$.*

We incorporate α so that the definition covers residual connections. A deep SSM model consists of SSM blocks along with an encoder, and a decoder transformation preceded by a time-pooling layer. We present an overview of the various architectures found in the literature in Table 1, Appendix B. Now we may define deep SSMs. For precise details about the particular elements see Appendix B.

Definition DT deep SSM. *A discrete time deep SSM model for classification is a function $f : \ell_T^{p,q}(\mathbb{R}^{n_{\text{in}}}) \to \mathbb{R}^{n_{\text{out}}}$ of the form $f = f^{\text{Dec}} \circ f^{\text{Pool}} \circ f^{\text{B}_L} \circ \ldots \circ f^{\text{B}_1} \circ f^{\text{Enc}}$, where $\circ$ denotes composition of functions. The functions f^{Enc} and f^{Dec} are linear transformations which are constant in time, while f^{B_i} is a DT-SSM block for all i. By pooling we mean the operation $f^{\text{Pool}}(\mathbf{u}) = \frac{1}{T}\sum_{k=1}^{T} \mathbf{u}[k]$, an average pooling over the time axis.*

3. Rademacher contraction of deep SSMs

Before we state our main theorem we introduce a property of functions, referred to as Rademacher Contraction, that is universal enough to include functions represented by both deep SSMs and neural networks.

Definition 1 ((μ, c)-Rademacher Contraction) *Let X_1 and X_2 be subsets of Banach spaces $\mathcal{X}_1, \mathcal{X}_2$, with norms $\|\cdot\|_{\mathcal{X}_1}$ and $\|\cdot\|_{\mathcal{X}_2}$, and let $\mu \geq 0$ and $c \geq 0$. A set of functions $\Phi = \{\varphi : X_1 \to X_2\}$ is said to be (μ, c)-Rademacher Contraction, or (μ, c)-RC in short., if for all $n \in \mathbb{N}^+$ and $Z \subseteq X_1^n$ we have*

$$\mathbb{E}_\sigma \left[\sup_{\varphi \in \Phi} \sup_{\{\mathbf{u}_i\}_{i=1}^n \in Z} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \varphi(\mathbf{u}_i) \right\|_{\mathcal{X}_2} \right] \leq \mu \mathbb{E}_\sigma \left[\sup_{\{\mathbf{u}_i\}_{i=1}^n \in Z} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{u}_i \right\|_{\mathcal{X}_1} \right] + \frac{c}{\sqrt{N}}, \quad (2)$$

where σ_i are i.i.d. Rademacher random variables, $i \in [N]$, i.e. $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$.

The RC property can be used to upper bound the Rademacher complexity of the hypothesis class. In fact, special cases of Definition 1 were used in the literature to this effect Golowich et al. (2018); Truong (2022a); Trauger and Tewari (2024) for deep neural networks and transformers respectively.

All blocks of deep SSMs have the RC property, for the proofs see Appendix D.1. In particular, SSM components of SSM blocks are $(\mu, 0)$-RC, and the constants μ depends on the control-theoretic system norm of the SSM component. Moreover, the RC property is preserved under composition of layers.

Lemma 2 (Composition lemma) *Let $\Phi_1 = \{\varphi_1 : X_1 \to X_2\}$ be (μ_1, c_1)-RC and $\Phi_2 = \{\varphi_2 : X_2 \to X_3\}$ be a (μ_2, c_2)-RC. Then the set of compositions $\Phi_2 \circ \Phi_1 := \{\varphi_2 \circ \varphi_1 : X_1 \to X_3 \mid \varphi_1 \in \Phi_1, \varphi_2 \in \Phi_2\}$ is a $(\mu_1\mu_2, \mu_2 c_1 + c_2)$-RC.*

The proof is in Appendix D.2. Lemma (2) allows us to establish the RC property for any deep structure, in particular, for deep SSMs. Next we describe our assumptions for deep SSMs.

Let $\mathcal{F}$ be a set of deep SSM models, namely let $f \in \mathcal{F}$ has the form $f = f^{\mathrm{Dec}} \circ f^{\mathrm{Pool}} \circ f^{\mathrm{B}_L} \circ \ldots \circ f^{\mathrm{B}_1} \circ f^{\mathrm{Enc}}$. Let us assume that **(1)** there exist constants $K_{\mathbf{u}}$ and K_y such that with probability one w.r.t. the data distribution $\mathcal{D}$, for any input-label pair $(\mathbf{u}, y)$, the $\ell_T^{2,2}$ norm of $\mathbf{u}$ and the absolute value of y are bounded from above $K_{\mathbf{u}}$ and K_y respectively, **(2)** the elementwise loss is L_l-Lipschitz continuous, **(3)** the SSM component of Σ_i of each SSM block in the parametrization is stable and its *norm $\|\Sigma_i\|_{p,q}$ is bounded by a constant $K_{p,q}$, $p, q = 2, \infty$*, **(4)** the norms of the weights of the encoder f^{Enc} and decoder f^{Dec} are bounded by $K_{\mathrm{Enc}}, K_{\mathrm{Dec}}$ respectively, **(5)** the non-linear component g_i of the ith SSM block g_i *is (μ_{g_i}, c_{g_i})-RC*. Internal stability of the SSM components is a standard assumption in the literature, and it is shown in Appendix D that the commonly used non-linear components are RC. We present the assumptions formally in Appendix C. Now we state our main theorem.

Theorem 3 (Informal theorem) *The following PAC inequality holds*

$$\mathbb{P}_{S\sim\mathcal{D}^N}\left[\forall f \in \mathcal{F} \quad \mathcal{L}(f) - \mathcal{L}^S_{emp}(f) \leq \frac{\mu K_{\mathbf{u}} L_l + c L_l}{\sqrt{N}} + K_l\sqrt{\frac{2\log(4/\delta)}{N}}\right] > 1 - \delta,$$

where the constants μ and c depend on the hypothesis class $\mathcal{F}$ and they satisfy

$$\mu \leq K_{\mathrm{Enc}} K_{\mathrm{Dec}} \prod_{i=1}^{L} \left(\mu_{g_i} K + \alpha_i\right), \qquad c \leq K_{\mathrm{Dec}} \sum_{j=1}^{L} \left[\prod_{i=j+1}^{L} \left(\mu_{g_i} K + \alpha_i\right)\right] c_{g_j}$$

and the constant $K_l > 0$ such that $|l(\cdot,\cdot)| \leq K_l$, while $K = \max\{K_{2,\infty}, K_{\infty,\infty}\}$.

The formal counterpart of Theorem 3 and its proof can be found in Appendix E. Notice that the bound does not depend on T. While the bound grows with the depth of the deep SSM, its growth can be controlled by choosing the state-space blocks with a small system norm. In turn, the system norm of the state-space models depends not only on the number and magnitude of its parameters, but on the degree of stability, i.e., systems with large number of weights with possibly large parameter norms can still have a small system norm. In practice, for popular deep SSM architectures, e.g. S4, S4D, S5 or LRU, the stability conditions are naturally met due their constrained parametrizations. This indicates that stability induced norms are crucial for deep SSMs.

Appendix A. Stability of SSMs

We recite the definition of internally stable dynamical systems of the form eq. 1.

Definition 4 (Antoulas (2005)) *SSM of the form (1) is internally stable, if the matrix A is Schur, meaning all the eigenvalues of A are inside the complex unit disk.*

In particular, a sufficient (but not necessary) condition for stability is that A is a contraction, i.e. $\|A\|_2 < 1$. A stable SSM Σ is not only robust to perturbations, but its input-output map can be extended to act on the Banach spaces of infinite sequences, generated by $\ell_T^{p,q}(\mathbb{R}^{n_{in}})$, $T \geq 0$.

More precisely, denote by $\ell^{2,2}(\mathbb{R}^n)$ and $\ell^{\infty,\infty}(\mathbb{R}^n)$ the Banach spaces generated by the all inifinite sequences over $\mathbb{R}^n$ such that the quantities $\|\mathbf{u}\|^2_{\ell^{2,2}(\mathbb{R}^n)} = \sum_{k=0}^{\infty} \|\mathbf{u}[k]\|_2^2$ and $\|\mathbf{u}\|_{\ell^{\infty,\infty}(\mathbb{R}^n)} = \sup_k \|\mathbf{u}[k]\|_\infty$ are well defined and finite. If $\mathbf{u} = \mathbf{u}[0]\ldots,\mathbf{u}[T-1]$ is a finite sequence of length T, then we can interpret it as an infinite sequence $\mathbf{u} = \mathbf{u}[0]\ldots,\mathbf{u}[T-1],0,0,\ldots$; elements of which are zero after the Tth element. With this identification, $\ell_T^{p,q}(\mathbb{R}^n)$ is a close subspace of $\subseteq \ell^{p,q}(\mathbb{R}^n)$, and $\ell^{p,q}(\mathbb{R}^n)$ contains no proper closed subspace containing $\bigcup_{T\geq 0} \ell_T^{p,q}(\mathbb{R}^n)$, i.e., $\bigcup_{T\geq 0} \ell_T^{p,q}(\mathbb{R}^n)$ generates $\ell^{p,q}(\mathbb{R}^n)$.

A stable SSM Σ is not only robust to perturbations, but its input-output map can be extended to a linear operator $\mathcal{S}_\Sigma : \ell^{p,p}(\mathbb{R}^{n_u}) \to \ell^{r,r}(\mathbb{R}^{n_y})$, for any choice $p, r \in \{\infty, 2\}$. More precisely, define $\mathcal{S}_\Sigma(\mathbf{u})[T-1] = \mathcal{S}_{\Sigma,T}(\mathbf{u}[0]\cdots\mathbf{u}[T-1])[T-1]$ for all $T > 0$. It then follows that for any $\mathbf{u} \in \ell^{p,p}(\mathbb{R}^{n_u})$, $\mathcal{S}_\Sigma(\mathbf{u}) \in \ell^{r,r}(\mathbb{R}^{n_y})$, $p, r = 2, \infty$, see Antoulas (2005). In particular, $\mathcal{S}_\Sigma$ has a well-defined induced norm as a linear operator, defined in the usual way, $\|\mathcal{S}_\Sigma\|_{p,r} = \sup_{\mathbf{u}\in\ell^{p,p}(\mathbb{R}^{n_u})} \frac{\|\mathcal{S}_\Sigma(\mathbf{u})\|_{\ell^{r,r}(\mathbb{R}^{n_y})}}{\|\mathbf{u})\|_{\ell^{r,r}(\mathbb{R}^{n_y})}}$. In the sequel, by a slight abuse of notation, we will denote by $\|\Sigma\|_{p,r}$ the induced norm $\|\mathcal{S}\|_{p,r}$.

As it was mentioned above

$$\|\Sigma\|_{p,r} = \sup_{T\geq 0} \|\mathcal{S}_{\Sigma,T}\|_{p,r}$$

Indeed, for any $\mathbf{u} \in \ell^{p,p}(\mathbb{R}^{n_u})$, $\|\mathcal{S}_\Sigma(\mathbf{u})\|_{\ell^{r,r}(\mathbb{R}^{n_y})} = \lim_{T\to\infty} \|\mathcal{S}_{\Sigma,T}(\mathbf{u}[0]\cdots\mathbf{u}[T-1])\|_{\ell_T^{r,r}(\mathbb{R}_y^n)}$, and $\|\mathbf{u}\|_{\ell^{p,p}(\mathbb{R}^{n_u})} = \lim_{T\to\infty} \|\mathbf{u}\|_{\ell_T^{p,p}(\mathbb{R}^{n_u})}$ and hence $\sup_{T>0} \|\mathcal{S}_{\Sigma,T}\|_{p,r} \geq \|\Sigma\|_{p,r}$. Moreover, for any $\mathbf{u} \in \ell_T^{p,p}(\mathbb{R}^{n_u})$, $\|\mathcal{S}_\Sigma(\mathbf{u})\|_{\ell^{r,r}(\mathbb{R}^{n_y})} \geq \|\mathcal{S}_{\Sigma,T}(\mathbf{u}[0]\cdots\mathbf{u}[T-1])\|_{\ell_T^{r,r}(\mathbb{R}_y^n)}$ and $\|\mathbf{u}\|_{\ell^{p,p}(\mathbb{R}^{n_u})} = \|\mathbf{u}\|_{\ell_T^{p,p}(\mathbb{R}^{n_u})}$, hence $\|\Sigma\|_{p,r} \geq \|\mathcal{S}_{\Sigma,T}\|_{p,r}$.

In this paper, we will use the induced norms $\|\Sigma\|_{2,\infty}$ and $\|\Sigma\|_{\infty,\infty}$ to upper bound the Rademacher complexity. In turn, these norms can be upper bounded by the following two standard control-theoretical norms defined on SSMs.

Definition 5 (Chellaboina et al. (1999)) *For a SSM Σ of the form (1) define the ℓ_1 and H_2 norm of Σ, denoted by $\|\Sigma\|_1$ and $\|\Sigma\|_2$ respectively,*

$$\|\Sigma\|_1 := \max_{1\leq i\leq n_y} \|D_i\|_1 + \sum_{k=0}^{\infty} \left\|C_i A^k B\right\|_1, \quad \|\Sigma\|_2 := \sqrt{\|D\|_F^2 + \sum_{k=0}^{\infty} \|CA^kB\|_F^2}$$

Lemma 6 (Chellaboina et al. (1999)) *For a system of form (1) $\|\Sigma\|_{\infty,\infty} \leq \|\Sigma\|_1$ and $\|\Sigma\|_{2,\infty} \leq \|\Sigma\|_2$.*

The norms defined above will play a cruicial role in in the main result of the paper, as they will allow us to bound the Rademacher complexity of the deep SSM model.

Remark 7 (Computing $\|\Sigma\|_i$,$i = 1, 2$) *The norm $\|\Sigma\|_2$ can be computed by solving Sylvester equations, for which standard numerical algorithms exist Antoulas (2005). The computation of $\|\Sigma\|_1$ is more involved, it can be computed by taking a sufficiently large finite sum instead of the infinite sum used in its definition. If $\|A\|_2 < \beta < 1$, then an easy calculation reveals that $\|\Sigma\|_1 \leq \|D\|_2 + \frac{\|B\|_2\|C\|_2}{1-\beta}$ and $\|\Sigma\|_2 \leq \sqrt{\|D\|_F^2 + \frac{n_y\|B\|_2^2\|C\|_2^2}{1-\beta^2}}$.*

Appendix B. Elements of deep SSM models

As it was mentioned above, the principal components of deep SSM models are SSM blocks. Various SSM models used in the literature differ from each other in the way the SSM components are parametrized and in the choice of the non-linear component of SSM blocks, see Table 1 for a summary. Note that in some papers, the matrices of the SSM component

Model	SSM	Block
S4 Gu et al. (2021)	$A = \Lambda - PQ^*$ block-diagonal	SSM + nonlinear activation
S4D Gu et al. (2022)	$A = -\exp(A_{Re}) + i \cdot A_{Im}$ block-diagonal	SSM + nonlinear activation
S5 Smith et al. (2022)	diagonal A	SSM + nonlinear activation
LRU Orvieto et al. (2023)	diagonal A exponential parametrization	SSM + MLP skip connection

Table 1: Summary of some popular deep SSM models.

are allowed to be complex valued, but such linear dynamical systems can be replaced by linear dynamical systems defined using real matrices, by doubling the dimension of the state-space and that of the input and the output space.

Remark 8 (Stability assumptions in the SSM literature) *In some of the cited papers, the discrete-time SSM components were obtained by discretizing internally stable continuous-time linear time-invariant dynamical model in time, using a fixed discretization time step. It is well-known in control theory that the thus obtained discrete-time linear systems of the form (1) are also internally stable. In this way, the majority of literature considers deep SSM models for which the SSM components are internally stable, at least as far as the parametrization used for learning is concerned, as they are internally stable in continuous-time.*

The Encoder and Decoder layers are given by the weight matrices W^{Enc} and W^{Dec}. Therefore $f^{\text{Enc}}(\mathbf{u})[k] = W^{\text{Enc}}\mathbf{u}[k]$ and $f^{\text{Dec}}(\mathbf{u})[k] = W^{\text{Dec}}\mathbf{u}[k]$ for all $k \in [T]$. We use the slightly abused notations $f^{\text{Enc}} \equiv \langle W^{\text{Enc}}, \cdot \rangle$ and $f^{\text{Dec}} \equiv \langle W^{\text{Dec}} \cdot \rangle$.

As for the Neural Network components of an SSM block, we consider the following two variants.

Definition 9 (MLP layer) *An MLP layer is a function from $\ell^{\infty,\infty}(\mathbb{R}^{n_u})$ to $\ell^{\infty,\infty}(\mathbb{R}^{n_u})$ that is induced by applying a deep neural network $f : \mathbb{R}^{n_u} \to \mathbb{R}^{n_u}$ for each timestep. A neural network of L layer is a function of the form $f = f_{W_1,\mathbf{b}_1} \circ \ldots \circ f_{W_L,\mathbf{b}_L} \circ g_{W_{L+1},\mathbf{b}_{L+1}}$, where $f_{W,\mathbf{b}}(\mathbf{x}) = \rho(g_{W,\mathbf{b}}(\mathbf{x}))$ is called a hidden layer, $g_{W,\mathbf{b}}(\mathbf{x}) = W\mathbf{x} + \mathbf{b}$ is called preactivation and ρ is the activation function, which is identical for all layers of the network and is either sigmoid or ReLU. The matrices are of the size $W_i \in \mathbb{R}^{n_{i+1} \times n_i}$ and $\mathbf{b} \in \mathbb{R}^{n_i}$ such that $n_1 = n_u$ and $n_{L+1} = n_u$. By slight abuse of notation, for $\mathbf{u} \in \ell^{\infty,\infty}(\mathbb{R}^{n_u})$ let $f(\mathbf{u}) \in \ell^{\infty,\infty}(\mathbb{R}^{n_v})$ such that $f(\mathbf{u})[k] = f(\mathbf{u}[k])$ for all $1 \le k \le T$.*

Definition 10 (GLU layer Smith et al. (2022)) *A GLU layer is a function of the form $GLU : \ell^{\infty,\infty}(\mathbb{R}^{n_u}) \to \ell^{\infty,\infty}(\mathbb{R}^{n_v})$ parametrized by a linear operaetor W such that $GLU(\mathbf{u})[k] = GELU(\mathbf{u}[k]) \odot \sigma(W(GELU(\mathbf{u}[k])))$, where σ is the sigmoid function and $GELU$ is the Gaussian Error Linear Unit Hendrycks and Gimpel (2016).*

Note, that this definition of GLU layer differs from the original definition in Dauphin et al. (2017), because in deep SSM models GLU is usually applied individually for each time step, without any time-mixing operations. See Appendix G.1 in Smith et al. (2022). The linear operation W is usually represented by a convolution operation.

Appendix C. Assumptions

Hereinafter we denote by $\mathcal{F}$ a set of deep SSM models represented by its direct product of its layerwise parameters. Furthermore, let $\mathcal{E}$ denote the set of all SSM models Σ for which there is a model $f \in \mathcal{F}$ such that Σ is an SSM layer of f. First, we restate our assumptions:

Assumption 11 *We assume the following properties hold.*

1. ***Scalar output.*** *Let $n_{\text{out}} = 1$.*
2. ***Lipschitz loss function.*** *Let the elementwise loss l be L_l-Lipschitz continuous.*
3. ***Bounded input.*** *There exist $K_{\mathbf{u}} > 0$ and $K_y > 0$ such that for any input trajectory $\mathbf{u}$ and label y sampled from $\mathcal{D}$, with probability 1 we have that $\|\mathbf{u}\|_{\ell_T^{2,2}(\mathbf{R}^{n_{\text{in}}})} \le K_{\mathbf{u}}$ and $|y| \le K_y$.*
4. ***Stability.*** *All $\Sigma \in \mathcal{E}$ are internally stable, implying $\|\Sigma\|_p < +\infty$ for $p = 1, 2$. Therefore we assume there exist $K_p > 0$ such that $\sup_{\Sigma \in \mathcal{E}} \|\Sigma\|_p < K_p$ for $p = 1, 2$.*
5. ***Bounded Encoder and Decoder.*** *We assume the Encoder and Decoder have bounded operator norms, i.e. $\sup_{W \in \mathcal{W}_{\text{Enc}}} \|W\|_{2,2} < K_{\text{Enc}}$ and $\sup_{W \in \mathcal{W}_{\text{Dec}}} \|W\|_{\infty,\beta} < K_{\text{Dec}}$ for $\beta \in \mathbb{N} \cup \{\infty\}$.*

6. ***Bounded MLPs and GLUs.*** *We assume that any considered MLP or GLU layer has a bounded parameters, i.e. for an L-layer model defined in Definition 9 we have* $\max_{1\le i\le L+1}\sup_{W\in\mathcal{W}_i}\|W\|_{\infty,\infty}<K_W$ *and* $\max_{1\le i\le L+1}\sup_{\mathbf{b}\in\mathcal{B}_i}\|\mathbf{b}\|_\infty<K_{\mathbf{b}}$*. Moreover, for any GLU layer defined in Definition 10 we have* $\sup_{W\in\mathcal{W}_{\mathrm{GLU}}}\|W\|_{\infty,\infty}<K_{\mathrm{GLU}}$.

Assumption 1: is not restrictive as we consider classification.
Assumption 2: The Lipschitzness holds for most of the loss functions used in practice. We mention that even the square-loss is Lipschitz on a bounded domain. From the practical aspect, the upped boundedness is also mild, as parameters along the learning algorithm's trajectory usually make l bounded. In the worst case, l is bounded on a bounded domain due to being Lipschitz.
Assumption 3: is yet again standard in the literature. Even in practical applications the input is usually normalized or standardized as a preprocessing step before learning.
Assumption 4: is the most important one as it plays a central role in our work. The motivation behind this assumption is twofold. First, in practical implementation of SSM based architectures, it is very common to apply some structured parametrization of the matrices of the systems, which leads to learning stable matrices. In many cases, the underlying intention is numerical stability of the learning algorithm, however we argue that the major advantage of such parametrizations is to ensure a stable behavior of the system. Second, similar stability assumptions are standard in control theory.
Assumption 5. and 6: are again fairly standard, as they require the weights of the encoder, decoder and network layers' to be bounded.

Appendix D. Technical results on Rademacher contractions

In this section we need to prove (μ,c)-RC property for linear (or affine) transformations which are constant in time, in many cases. For better readibility, we only do the calculations once and use it as a lemma.

Lemma 12 *Let* $\mathbf{u}\in\ell_T^{p,p}(\mathbb{R}^{n_u})=:X_1$ *and let* $f_{W,\underline{\mathbf{b}}}(\mathbf{u})=W(\mathbf{u})+\underline{\mathbf{b}}\in\ell_T^{q,q}(\mathbb{R}^{n_v})=:X_2$*, where* $W\in\mathcal{L}(X_1,X_2)$ *is a linear operator and* $\underline{\mathbf{b}}\in X_2$*. We consider the cases*

a) $p=q=2$,

b) $p=2,q=\infty$,

c) $p=q=\infty$.

Let us assume that $W\in\mathcal{W}$ *such that* $\sup_{W\in\mathcal{W}}\|W\|_{\mathrm{op}}<K_W$ *and* $\underline{\mathbf{b}}\in\mathcal{B}$ *such that* $\sup_{\underline{\mathbf{b}}\in\mathcal{B}}\|\underline{\mathbf{b}}\|_{q,q}<K_{\mathbf{b}}$ *for all the considered cases. Then the set of transformation* $\mathcal{F}=\{f_{W,\underline{\mathbf{b}}}\mid W\in\mathcal{W},\mathbf{b}\in\mathcal{B}\}$ *is* $(K_W,K_{\mathbf{b}})$*-RC in all three cases. Furthermore, the image of the ball* $B_{X_1}(r)$ *under* $f\in\mathcal{F}$ *is contained in* $B_{X_2}(K_Wr+K_{\mathbf{b}})$.

Remark 13 *For the special case of affine transformations that are constant in time, i.e.* $f(\mathbf{u})[k]=W\mathbf{u}[k]+\mathbf{b}$ *for a weight matrix* $W\in\mathbb{R}^{n_v\times n_u}$ *and bias term* $\mathbf{b}\in\mathbb{R}^{n_v}$ *for all* $k\in[T]$*, the operator norm equals to the corresponding matrix norm, i.e* $\|W\|_{\mathrm{op}}=\|W\|_{p,q}$*. In this case,* $\underline{\mathbf{b}}$ *is the sequence for which* $\underline{\mathbf{b}}[k]=\mathbf{b}$ *for all* $k\in[T]$*, thus* $\|\underline{\mathbf{b}}\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})}=\|\mathbf{b}\|_q$.

Proof First let us prove a simple fact about Rademacher random variables that we will need, namely if $\sigma = \{\sigma_i\}_{i=1}^N$ is a sequence of i.i.d. Rademacher variables, then

$$\mathbb{E}_\sigma \left[\left| \sum_{i=1}^N \sigma_i \right| \right] \leq \sqrt{N}. \tag{3}$$

This is true, because

$$\mathbb{E}_\sigma \left[\left| \sum_{i=1}^N \sigma_i \right| \right] = \sqrt{\left(\mathbb{E}_\sigma \left[\left| \sum_{i=1}^N \sigma_i \right| \right] \right)^2} \leq \sqrt{\mathbb{E}_\sigma \left[\left| \sum_{i=1}^N \sigma_i \right|^2 \right]}$$

$$= \sqrt{\mathbb{E}_\sigma \left[\sum_{i=1}^N \sigma_i^2 + 2 \sum_{i,j=1}^N \sigma_i \sigma_j \right]} = \sqrt{\sum_{i=1}^N \mathbb{E}_\sigma \left[\sigma_i^2 \right] + 2 \sum_{i,j=1}^N \mathbb{E}_\sigma \left[\sigma_i \sigma_j \right]} = \sqrt{N},$$

where the first inequality follows from Jensen's inequality and the last equality follows from the linearity of the expectation, and the facts that σ_i are Rademacher variables and form and i.i.d sample.

The proof is the same for all the considered cases of p and q. For $Z \in \ell_T^{p,p}(\mathbb{R}^{n_u})$ we have

$$\mathbb{E}_\sigma \left[\sup_{(W,\underline{\mathbf{b}}) \in \mathcal{W} \times \mathcal{B}} \sup_{\{\mathbf{u}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i (W(\mathbf{u}_i) + \underline{\mathbf{b}}) \right\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})} \right]$$

$$\leq \mathbb{E}_\sigma \left[\sup_{W \in \mathcal{W}} \sup_{\{\mathbf{u}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i W(\mathbf{u}_i) \right\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})} \right] + \mathbb{E}_\sigma \left[\sup_{\mathbf{b} \in \mathcal{B}} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i \underline{\mathbf{b}} \right\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})} \right]$$

$$= \mathbb{E}_\sigma \left[\sup_{W \in \mathcal{W}} \sup_{\{\mathbf{u}_i\}_{i=1}^N \in Z} \left\| W \left(\frac{1}{N} \sum_{i=1}^N \sigma_i \mathbf{u}_i \right) \right\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})} \right] + \mathbb{E}_\sigma \left[\sup_{\mathbf{b} \in \mathcal{B}} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i \underline{\mathbf{b}} \right\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})} \right]$$

$$\leq \mathbb{E}_\sigma \left[\sup_{W \in \mathcal{W}} \|W\|_{\mathrm{op}} \sup_{\{\mathbf{u}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i \mathbf{u}_i \right\|_{\ell_T^{p,p}(\mathbb{R}^{n_u})} \right] + \mathbb{E}_\sigma \left[\frac{1}{N} \left| \sum_{i=1}^N \sigma_i \right| \sup_{\mathbf{b} \in \mathcal{B}} \|\underline{\mathbf{b}}\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})} \right]$$

$$\leq \sup_{W \in \mathcal{W}} \|W\|_{\mathrm{op}} \, \mathbb{E}_\sigma \left[\sup_{\{\mathbf{u}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i \mathbf{u}_i \right\|_{\ell_T^{p,p}(\mathbb{R}^{n_u})} \right] + \sup_{\mathbf{b} \in \mathcal{B}} \|\underline{\mathbf{b}}\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})} \, \mathbb{E}_\sigma \left[\frac{1}{N} \left| \sum_{i=1}^N \sigma_i \right| \right]$$

$$\leq \sup_{W \in \mathcal{W}} \|W\|_{\mathrm{op}} \, \mathbb{E}_\sigma \left[\sup_{\{\mathbf{u}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i \mathbf{u}_i \right\|_{\ell_T^{p,p}(\mathbb{R}^{n_u})} \right] + \frac{1}{\sqrt{N}} \sup_{\mathbf{b} \in \mathcal{B}} \|\underline{\mathbf{b}}\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})}$$

where the first inequality follows from the triangle inequality, the first equality is the linearity of W, the second inequality follows from the definition of the operator norm, while the third and fourth inequalities refer only to the bias term and follow from the absolute homogenity of the norm and equation (3).

The values of (p, q) influence the terms $\|W\|_{\mathrm{op}}$ and $\|\underline{\mathbf{b}}\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})}$. As a result, for the separate cases of a), b) and c) it is enough to separately bound these norms with constants

similar to K_W. In the statement of the Lemma we assumed universal constants K_W and $K_{\mathbf{b}}$ that do not depend on the values of (p, q).

We can see that the calculations hold if the transformations are restricted to the ball $B_{X_1}(r)$ for any choice of X_1 we consider. The radius can grow as

$$\|W(\mathbf{u}) + \mathbf{b}\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})} \leq \|W(\mathbf{u})\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})} + \|\underline{\mathbf{b}}\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})} \leq \|W\|_{\mathrm{op}} \|\mathbf{u}\|_{\ell_T^{p,p}(\mathbb{R}^{n_u})} + \|\underline{\mathbf{b}}\|_{\ell_T^{q,q}(\mathbb{R}^{n_v})}.$$

Remark 13 is straightforward from the definitions of the considered norms. ■

D.1. Proofs for elements of deep SSMs

What we have left are the following. First, in light of the previous corollary, we need to show that each component of a deep SSM model is (μ, c)-RC for some μ and c w.r.t. compatible normed spaces. Second, we need to show that the Rademacher complexity of a (μ, c)-RC model set are bounded in terms of μ and c. We start with the first one.

Lemma 14 *Let $\mathcal{W}_{\mathrm{Enc}}$, $\mathcal{W}_{\mathrm{Dec}}$, and $\mathcal{E}$ denote some sets of parameters of some fixed Encoder, Decoder and SSM layers, respectively. Moreover, let $\{\mathcal{W}_i \times \mathcal{B}_i\}_{i=1}^{L+1}$ the parameter set of an MLP layer defined in Definition 9, and let $\mathcal{W}_{\mathrm{GLU}}$ be the parameter set of a GLU layer defined in Definition 10. The corresponding function sets (in line with Assumption 11) are*

- $\mathcal{F}_{\mathrm{Enc}} = \{f^{\mathrm{Enc}} = \langle W, \cdot \rangle \mid W \in \mathcal{W}_{\mathrm{Enc}}, \sup_{W \in \mathcal{W}_{\mathrm{Enc}}} \|W\|_{2,2} < K_{\mathrm{Enc}}\}$,
- $\mathcal{F}_{\mathrm{Dec}} = \{f^{\mathrm{Dec}} = \langle W, \cdot \rangle \mid W \in \mathcal{W}_{\mathrm{Dec}}, \sup_{W \in \mathcal{W}_{\mathrm{Dec}}} \|W\|_{\infty,\infty} < K_{\mathrm{Dec}}\}$,
- $\mathcal{F}_{\mathrm{SSM}} = \{\mathcal{S}_\Sigma \mid \Sigma \in \mathcal{E}, \sup_{\Sigma \in \mathcal{E}} \|\Sigma\|_p < K_p, p = 1, 2\}$,
- $$\mathcal{F}_{\mathrm{MLP}}^{\rho} = \left\{ f_{deep} = f_{W_1,\mathbf{b}_1} \circ \ldots \circ f_{W_L,\mathbf{b}_L} \circ g_{W_{L+1},\mathbf{b}_{L+1}} \;\middle|\; \begin{array}{l} (W_i, \mathbf{b}_i) \in \mathcal{W}_i \times \mathcal{B}_i, \\ \sup_{W \in \mathcal{W}_i} \|W\|_{\infty,\infty} < K_W, \\ \sup_{\mathbf{b} \in \mathcal{B}_i} \|\mathbf{b}\|_\infty < K_{\mathbf{b}}, \\ 1 \leq i \leq L+1 \end{array} \right\},$$
- $$\mathcal{F}_{\mathrm{GLU}} = \left\{ f_{GLU} = GELU(\cdot) \odot \sigma \circ \langle W, GELU(\cdot) \rangle \;\middle|\; \begin{array}{l} W_{\mathrm{GLU}} \in \mathcal{W}_{\mathrm{GLU}}, \\ \sup_{W \in \mathcal{W}_{\mathrm{GLU}}} \|W\|_{\infty,\infty} < K_{\mathrm{GLU}} \end{array} \right\},$$

where $\mathcal{S}_\Sigma$ denotes the input-output map of the dynamical system Σ, ρ is either the sigmoid or ReLU activations, and σ denotes the sigmoid functions in the definition of GLU. Then all of these function sets are (μ, c)-RC according to the following table, where for any Banach space $\mathcal{X}$, $B_{\mathcal{X}}(t) = \{x \in \mathcal{X} \mid \|x\|_{\mathcal{X}} \leq r\}$ denotes the ball of radius r centered in zero for arbitrary r.

	μ	c	X_1	X_2
$\mathcal{F}_{\text{Enc}}$	K_{Enc}	0	$B_{\ell_T^{2,2}(\mathbb{R}^{n_{\text{in}}})}(r)$	$B_{\ell_T^{2,2}(\mathbb{R}^{n_u})}(K_{\text{Enc}}r)$
$\mathcal{F}_{\text{Dec}}$	K_{Dec}	0	$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r)$	$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_{\text{out}}})}(K_{\text{Dec}}r)$
	K_{Dec}	0	$B_{(\mathbb{R}^{n_u},\|\cdot\|_\infty)}(r)$	$B_{(\mathbb{R}^{n_{\text{out}}},\|\cdot\|_\beta)}(K_{\text{Dec}}r)$
$\mathcal{F}_{\text{SSM}}$	K_2	0	$B_{\ell_T^{2,2}(\mathbb{R}^{n_u})}(r)$	$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_y})}(K_2r)$
	K_1	0	$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r)$	$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_y})}(K_1r)$
$\mathcal{F}_{\text{MLP}}^{ReLU}$	$4K_W(L+1)$	$4K_{\mathbf{b}}\sum_{q=1}^{L}(4K_W)^q$	$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r)$	$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}\left(K_W^{L+1}r + K_{\mathbf{b}}\sum_{q=1}^{L-1}K_W^q\right)$
$\mathcal{F}_{\text{MLP}}^{sigmoid}$	$K_W(L+1)$	$(K_{\mathbf{b}}+0.5)\sum_{q=1}^{L}(K_W)^q$	$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r)$	$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(K_Wr + K_{\mathbf{b}})$
$\mathcal{F}_{\text{GLU}}$	$16(r(K_{\text{GLU}}+1)^2$ $+K_{\text{GLU}}+1)$	0	$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r)$	$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r)$

Furthermore, the operation of f^{Pool}, according the definition of deep SSMs, is $(1,0)$-RC between $X_1 = B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r)$ and $X_2 = B_{(\mathbb{R}^{n_u},\|\cdot\|_\infty)}(r)$.

Proof

Encoder and Decoder. The Encoder is case **a)**, while the Decoder is case **b)** in Lemma 12 along with Remark 13.

SSM. As discussed in Appendix B, an SSM is equivalent to a linear transformation called its input-output map. Therefore, by Lemma 12, the SSM is $(\mu,0)$-RC in both cases, where μ is the operator norm of the input-output map. Combining this with Lemma 6 yields the result.

Remark 15 *As the value of T is fixed, the input-output map can be described by the so-called Toeplitz matrix of the system. In this case, the operator norm equals to the appropriate induced matrix norm of the Toeplitz matrix. For the case of $T=\infty$, the input-output map still exists and is a linear operator. The proof of Lemma 12 holds in this case as well for operator norms.*

Pooling. For any $Z \subseteq \ell_T^{\infty,\infty}(\mathbb{R}^{n_u})$ we have

$$\begin{aligned}
&\mathbb{E}_\sigma\left[\sup_{\{\mathbf{z}_i\}_{i=1}^N\in Z}\left\|\frac{1}{N}\sum_{i=1}^N\sigma_i f^{\text{Pool}}(\mathbf{z}_i)\right\|_\infty\right]\\
&=\mathbb{E}_\sigma\left[\sup_{\{\mathbf{z}_i\}_{i=1}^N\in Z}\sup_{1\le j\le n_u}\left|\frac{1}{N}\sum_{i=1}^N\sigma_i\left(\frac{1}{T}\sum_{k=1}^T\mathbf{z}_i^{(j)}[k]\right)\right|\right]\\
&=\mathbb{E}_\sigma\left[\sup_{\{\mathbf{z}_i\}_{i=1}^N\in Z}\sup_{1\le j\le n_u}\left|\frac{1}{T}\sum_{k=1}^T\left(\frac{1}{N}\sum_{i=1}^N\sigma_i\mathbf{z}_i^{(j)}[k]\right)\right|\right]\\
&\le\mathbb{E}_\sigma\left[\sup_{\{\mathbf{z}_i\}_{i=1}^N\in Z}\frac{1}{T}\sum_{k=1}^T\sup_{1\le j\le n_u}\left|\frac{1}{N}\sum_{i=1}^N\sigma_i\mathbf{z}_i^{(j)}[k]\right|\right]
\end{aligned}$$

$$= \mathbb{E}_\sigma \left[\sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \frac{1}{T} \sum_{k=1}^{T} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{z}_i[k] \right\|_\infty \right]$$
$$\leq \mathbb{E}_\sigma \left[\sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{z}_i \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right]$$

MLP with sigmoid activations. Consider a single hidden layer $f(\mathbf{x}) = \rho(g(\mathbf{x}))$, where $g(x) = W\mathbf{x} + \mathbf{b}$ is the preactivation and let $\mathcal{G} = \{g(x) = W\mathbf{x} + \mathbf{b} \mid W \in \mathcal{W}, \mathbf{b} \in \mathcal{B}\}$ denote the set of possible preactivation functions. As compared to Definition 9, we omit the subscript from the notation of g. For an input sequence $\mathbf{z} \in \ell_T^{\infty,\infty}(\mathbb{R}^{n_u})$ let $g(\mathbf{z}) \in \ell_T^{\infty,\infty}(\mathbb{R}^{n_v})$ mean that we apply g for each timestep independently, i.e. $g(\mathbf{z})[k] = g(\mathbf{z}[k])$. We have

$$\mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \rho(g(\mathbf{z}_i)) \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right]$$
$$= \mathbb{E}_\sigma \left[\sup_{(W,\mathbf{b}) \in \mathcal{W} \times \mathcal{B}} \sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \sup_{1 \leq k \leq T} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \rho(W\mathbf{z}_i[k] + \mathbf{b}) \right\|_\infty \right]$$

Let $\mathbf{x}_i = i$, $i = 1, \ldots, N$ and let $\mathcal{H} = \{h_{W,\mathbf{b},\underline{z},k} \mid (W, \mathbf{b}, \underline{z}, k) \in \mathcal{W} \times \mathcal{B} \times (Z \cup \{0\}) \times [T]\}$ such that $h_{W,\mathbf{b},\underline{z},k}(\mathbf{x}_i) = g(\mathbf{z}_i[k])$. Under the assumption that $\mathcal{H}$ is symmetric to the origin, meaning that $h \in \mathcal{H}$ implies $-h \in \mathcal{H}$ (equivalently $(W, \mathbf{b}) \in \mathcal{W} \times \mathcal{B}$ implies $(-W, -\mathbf{b}) \in \mathcal{W} \times \mathcal{B}$), we can apply (Truong, 2022b, Theorem 9) for the sigmoid activation and hence $\rho - \rho(0)$ being odd, as follows.

$$\mathbb{E}_\sigma \left[\sup_{(W,\mathbf{b}) \in \mathcal{W} \times \mathcal{B}} \sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \sup_{1 \leq k \leq T} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \rho(W\mathbf{z}_i[k] + \mathbf{b}) \right\|_\infty \right]$$
$$= \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \rho(h(\mathbf{x}_i)) \right\|_\infty \right]$$
$$\leq \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i h(\mathbf{x}_i) \right\|_\infty \right] + \frac{1}{2\sqrt{N}}$$
$$= \mathbb{E}_\sigma \left[\sup_{(W,\mathbf{b}) \in \mathcal{W} \times \mathcal{B}} \sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \sup_{1 \leq k \leq T} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i (W\mathbf{z}_i[k] + \mathbf{b}) \right\|_\infty \right] + \frac{1}{2\sqrt{N}}$$
$$= \mathbb{E}_\sigma \left[\sup_{(W,b) \in \mathcal{W} \times \mathcal{B}} \sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i (W\mathbf{z}_i + \mathbf{b}) \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_v})} \right] + \frac{1}{2\sqrt{N}},$$

because the sigmoid is 1-Lipschitz and $\rho(0) = 0.5$. Now we can apply Lemma 12 (see Remark 13) to get that

$$\mathbb{E}_\sigma \left[\sup_{(W,\mathbf{b})\in\mathcal{W}\times\mathcal{B}} \sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N}\sum_{i=1}^N \sigma_i(W\mathbf{z}_i + \mathbf{b}) \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_v})} \right] + \frac{1}{2\sqrt{N}}$$
$$\leq \sup_{W\in\mathcal{W}} \|W\|_{\infty,\infty} \mathbb{E}_\sigma \left[\sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N}\sum_{i=1}^N \sigma_i \mathbf{z}_i \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right] + \frac{1}{\sqrt{N}} \sup_{\mathbf{b}\in\mathcal{B}} \|\mathbf{b}\|_\infty + \frac{1}{2\sqrt{N}}$$

Therefore, the sigmoid MLP layer is $(K_W, K_\mathbf{b} + 0.5)$-RC. The restriction of an MLP to the ball $B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r)$ maps to the ball $B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_v})}(1)$, because of the elementwise sigmoid activation. For the deep model the result is starightforward from Lemma 2 along with Lemma 12, Remark 13 and Remark 19.

MLP with ReLU activations. Similarly to the sigmoid case, we assume the upper bounds K_W and $K_\mathbf{b}$ exist, but we don't assume the symmetry of the parameter set. The proof is the same as in the sigmoid case up to the first inequality. Here we can apply (Ledoux and Talagrand, 1991, Equation 4.20) (this is the same idea as in the proof of (Golowich et al., 2018, Lemma 2)) to get

$$\mathbb{E}_\sigma \left[\sup_{h\in\mathcal{H}} \left\| \frac{1}{N}\sum_{i=1}^N \sigma_i \rho(h(\mathbf{x}_i)) \right\|_\infty \right] \leq 4\mathbb{E}_\sigma \left[\sup_{h\in\mathcal{H}} \left\| \frac{1}{N}\sum_{i=1}^N \sigma_i h(\mathbf{x}_i) \right\|_\infty \right],$$

where we used that $\rho(x) = ReLU(x)$ is 1-Lipschitz and the same logic for the alternative definition of the Rademacher complexity as in the proof of Lemma 14, which results in a constant factor of 2. The constant 4 is then obtained by the additional constant factor 2 from Talagrand's lemma. The rest of proof is identical to the sigmoid case.

The restriction of an MLP to the ball $B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r)$ maps to the ball $B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_v})}(K_W r + K_\mathbf{b})$, because the elementwise ReLU does not increase the infinity norm, hence we can apply Lemma 12 and Remark 13. Again, for the deep model the result is straightforward from Lemma 2 along with Lemma 12, Remark 13 and Remark 19.

GLU. First of all, we show that the function $h : (\mathbb{R}^2, \|\cdot\|_2) \to (\mathbb{R}, |\cdot|)$ defined as $h(\mathbf{x}) = x_1 \cdot \sigma(x_2)$ is $\sqrt{2}(K+1)$-Lipschitz on a bounded domain, where $|x_i| \leq K$ for all $\mathbf{x} \in \mathbb{R}^2$ we consider. We will later specify the value of K in relation to Assumption 11. By the sigmoid being 1-Lipschitz, we have

$$|h(\mathbf{x}) - h(\mathbf{y})| = |x_1\sigma(x_2) - y_1\sigma(x_2) + y_1\sigma(x_2) - y_1\sigma(y_2)| \leq$$
$$|(x_1 - y_1)\sigma(x_2)| + |y_1(\sigma(x_2) - \sigma(y_2))| \leq |x_1 - y_1| + |y_1||x_2 - y_2|$$
$$\leq \sqrt{2}(K+1)\,\|\mathbf{x} - \mathbf{y}\|_2$$

Second, we recall Corollary 4 in Maurer (2016).

Theorem 16 (Maurer (2016)) *Let $\mathcal{X}$ be any set, $(\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathcal{X}^N$, let $\mathcal{F}$ be a set of functions $f : \mathcal{X} \to \ell_T^2(\mathbb{R}^m)$ and let $h : \ell_T^2(\mathbb{R}^m) \to \mathbb{R}$ be an L-Lipschitz function. Under*

f_k denoting the k-th component function of f and σ_{ik} being a doubly indexed Rademacher variable, we have

$$\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{N}\sigma_i h(f(\mathbf{x}_i))\right] \leq \sqrt{2}L\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{N}\sum_{k=1}^{m}\sigma_{ik}f_k(\mathbf{x}_i)\right].$$

We wish to apply Theorem 16 to GLU layers. For any $Z \subseteq \ell_T^{\infty,\infty}(\mathbb{R}^{n_u})$, by letting $GLU_W(\mathbf{z}) = f_{GLU}(\mathbf{z})$ we have

$$\begin{aligned}
&\mathbb{E}_\sigma\left[\sup_{W\in\mathcal{W}}\sup_{\{\mathbf{z}_i\}_{i=1}^N\in Z}\left\|\frac{1}{N}\sum_{i=1}^{N}\sigma_i GLU_W(\mathbf{z}_i)\right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}\right] \\
&= \mathbb{E}_\sigma\left[\sup_{W\in\mathcal{W}}\sup_{\{\mathbf{z}_i\}_{i=1}^N\in Z}\sup_{1\leq k\leq T}\sup_{1\leq j\leq n_u}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i GLU_W^{(j)}(\mathbf{z}_i)[k]\right|\right]
\end{aligned}$$

Now this is an alternative version of the Rademacher complexity, where we take the absolute value of the Rademacher average. In order to apply Theorem 16, we reduce the problem to the usual Rademacher complexity. In turn, we can apply the last chain of inequalities in the proof of Proposition 6.2 in Hajek and Raginsky (2019). Concretely, by denoting $\mathbf{O} = \{\mathbf{0}\}_{i=1}^N$ and noticing that $GLU_W(0) = 0$, we have

$$\begin{aligned}
&\mathbb{E}_\sigma\left[\sup_{W\in\mathcal{W}}\sup_{\{\mathbf{z}_i\}_{i=1}^N\in Z}\sup_{1\leq k\leq T}\sup_{1\leq j\leq n_u}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i GLU_W^{(j)}(\mathbf{z}_i)[k]\right|\right] \\
&\leq 2\mathbb{E}_\sigma\left[\sup_{W\in\mathcal{W}}\sup_{\{\mathbf{z}_i\}_{i=1}^N\in Z\cup\{\mathbf{O}\}}\sup_{1\leq k\leq T}\sup_{1\leq j\leq n_u}\frac{1}{N}\sum_{i=1}^{N}\sigma_i GLU_W^{(j)}(\mathbf{z}_i)[k]\right]
\end{aligned}$$

Let $\mathbf{x}_i = i$, $i = 1,\ldots,N$ and let $\mathcal{F} = \{f_{W,\underline{z},k,j} \mid (W,\underline{z},k,j) \in \mathcal{W}\times(Z\cup\{0\})\times[T]\times[n_u]\}$ such that $f_{W,\underline{z},k,j}(\mathbf{x}_i) = \begin{bmatrix} GELU(\mathbf{z}_i[k])^{(j)} & (W(GELU(\mathbf{z}_i[k])))^{(j)}\end{bmatrix}^T$ for $\underline{z} = \{\mathbf{z}_i\}_{i=1}^N \in Z$. Since $Z \subseteq (B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(K_\mathbf{u}))^N$, it follows for all $\{\mathbf{z}_i\}_{i=1}^N \in Z$ and for all $k\in\mathbb{N}$ that $\|\mathbf{z}_i[k]\|_\infty \leq K_\mathbf{u}$, and hence $|GELU(\mathbf{z}_i[k])^{(j)}| < K_\mathbf{u}$, leading to $|W(GELU(\mathbf{z}_i[k]))^{(j)}| < \sup_{W\in\mathcal{W}}\|W\|_{\infty,\infty}\cdot K_\mathbf{u}$. In particular, $GLU_W^{(j)}(\mathbf{z}_i)[k] = h(f_{W,\underline{z},k,j}(\mathbf{x}_i)) = h|_B(f_{W,\underline{z},k,j}(\mathbf{x}_i))$, where $h|_B$ is the restriction of h to $B = \{x\in\mathbb{R}^2 \mid \|x\|_\infty < K\}$, and hence $h|_B$ is $\sqrt{2}(K+1)$-Lipschitz. Therefore we can set $K = \max\{K_\mathbf{u}, \sup_{W\in\mathcal{W}}\|W\|_{\infty,\infty}\cdot K_\mathbf{u}\}$.

We are ready to apply Theorem 16, together with the GLU definition and its $\sqrt{2}(K+1)$-Lipschitzness, we have

$$2\mathbb{E}_\sigma\left[\sup_{W\in\mathcal{W}}\sup_{\{\mathbf{z}_i\}_{i=1}^N\in Z\cup\{\mathbf{O}\}}\sup_{1\leq k\leq T}\sup_{1\leq j\leq n_u}\frac{1}{N}\sum_{i=1}^{N}\sigma_i GLU_W^{(j)}(\mathbf{z}_i)[k]\right]$$

$$= 2\mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i h(f(\mathbf{x}_i)) \right] \leq 4(K+1) \underbrace{\mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i GELU(\mathbf{z}_i[k])^{(j)} \right]}_{A}$$

$$+ 4(K+1) \underbrace{\mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i W(GELU(\mathbf{z}_i))^{(j)}[k] \right]}_{B}$$

Due to the definition of GELU, its 2-Lipschitzness Qi et al. (2023) and (Ledoux and Talagrand, 1991, Theorem 4.12) we have

$$A = \mathbb{E}_\sigma \left[\sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z \cup \{\mathbf{O}\}} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i GELU(\mathbf{z}_i) \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right] =$$

$$\leq 4\mathbb{E}_\sigma \left[\sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z \cup \{\mathbf{O}\}} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{z}_i \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right] = 4\mathbb{E}_\sigma \left[\sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{z}_i \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right]$$

and

$$B = \mathbb{E}_\sigma \left[\sup_{W\in\mathcal{W}} \sup_{\{\mathbf{z}_i\}_{i=1}^N \in \{\mathbf{O}\}} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i W(GELU(\mathbf{z}_i)) \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right]$$

$$\leq \sup_{W\in\mathcal{W}} \|W\|_{\infty,\infty} \mathbb{E}_\sigma \left[\sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z\{\mathbf{O}\}} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i GELU(\mathbf{z}_i) \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right]$$

$$\leq 4 \sup_{W\in\mathcal{W}} \|W\|_{\infty,\infty} \mathbb{E}_\sigma \left[\sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{z}_i \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right]$$

Here we used the linearity of W and the exact same calculation as in the proof of Lemma 12.

By combining the inequalities above, it follows that

$$\mathbb{E}_\sigma \left[\sup_{W\in\mathcal{W}} \sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \sup_{1\leq k\leq T} \sup_{1\leq j\leq n_u} \left| \frac{1}{N} \sum_{i=1}^{N} \sigma_i GLU_W^{(j)}(\mathbf{z}_i)[k] \right| \right] \leq$$

$$16(K+1) \left(\sup_{W\in\mathcal{W}} \|W\|_{\infty,\infty} + 1 \right) \mathbb{E}_\sigma \left[\sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{z}_i \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right]$$

Substituting the value of K gives the result. ■

D.2. Proof of compostion Lemma 2

Proof [Proof of Lemma 2] Let $Z \subseteq X_1^N$ and $\tilde{Z} = \{\{\varphi_1(\mathbf{u}_i)\}_{i=1}^N \mid \varphi_1 \in \Phi_1\}$. We have

$$\mathbb{E}_\sigma \left[\sup_{\varphi_2\in\Phi_2} \sup_{\varphi_1\in\Phi_1} \sup_{\{\mathbf{u}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \varphi_2(\varphi_1(\mathbf{u}_i)) \right\|_{X_3} \right]$$

$$
\begin{aligned}
&= \mathbb{E}_\sigma \left[\sup_{\varphi_2 \in \Phi_2} \sup_{\{\mathbf{v}_i\}_{i=1}^N \in \tilde{Z}} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i \varphi_2(\mathbf{v}_i) \right\|_{X_3} \right] \\
&\leq \mu_2 \mathbb{E}_\sigma \left[\sup_{\varphi_1 \in \Phi_1} \sup_{\{\mathbf{u}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i \varphi_1(\mathbf{u}_i) \right\|_{X_2} \right] + \frac{c_2}{\sqrt{N}} \\
&\leq \mu_2 \mu_1 \mathbb{E}_\sigma \left[\sup_{\{\mathbf{u}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i \mathbf{u}_i \right\|_{X_1} \right] + \mu_2 \frac{c_1}{\sqrt{N}} + \frac{c_2}{\sqrt{N}}
\end{aligned}
$$

■

The upcoming corollary is straightforward by induction along with the fact that the pooling layer f^{Pool} is $(1,0)$-RC (see Lemma 14).

Corollary 17 *Let $\mathcal{F}$ be a set of deep SSM models, i.e. according to Definition of deep SSM, for $f \in \mathcal{F}$ we have $f = f^{\text{Dec}} \circ f^{\text{Pool}} \circ f^{\text{B}_L} \circ \ldots \circ f^{\text{B}_1} \circ f^{\text{Enc}}$, such that f^{Enc}, f^{Dec} and each f^{B_i} are from (μ_0, c_0)-RC, (μ_{L+1}, c_{L+1})-RC and (μ_i, c_i)-RC sets respectively for all i. Then $\mathcal{F}$ is $\left(\prod_{i=0}^{L+1} \mu_i, \sum_{j=1}^{L} \left[\prod_{i=j+1}^{L+1} \mu_i \right] c_j \right)$-RC.*

We can see that the SSM layer can only increase the input's complexity by the factor $\|\Sigma\|_p$, $p = 1, 2$, a quantity that gets smaller as the system gets more stable. This gets even more crucial when dealing with long range sequences, because the Neural Network layers are constant in time.

Remark 18 *The results of of Lemma 14 hold for unbounded input spaces as well. The reason for restricting the input space to a ball of radius r is the composition with MLPs or GLU layers, as discussed in the interpretation of Definition 1.*

Appendix E. Statement and proof of the main theorem

So far we showed in D that each component of a deep SSM model satisfies Definition 1. We also proved that the composition of such components also satisfies the definition. The main theorem summarizes these results and exploits the fact that the Rademacher complexity of a (μ, c)-RC set of models is upper bounded by terms containing μ and c.

Before we state the formal theorem let us discuss balls in Banach spaces regarding the contraction lemma.

Remark 19 *The results of of Lemma 14 hold for unbounded input spaces as well. The reason for restricting the input space to a ball of radius r is the composition with GLU layers, as discussed in the interpretation of Definition 1. As a result of Lemma 14, the restriction of a deep SSM model to a ball of an arbitrary radius r has its image contain in a ball with a radius $\hat{r}$ depending on r and the possible parameter set of each layer in the composite model. The exact value of $\hat{r}$ can be calculated using Lemma 14. Namely, consider a residual SSM block, defined as $f^{\text{B}}(\mathbf{z})[k] = g(\mathcal{S}_{\Sigma_i}(\mathbf{z})[k]) + \alpha \mathbf{z}[k]$ for all $k \in [T]$. Let $R_g(r)$ and $R_\Sigma(r)$ be*

the radius of the ball that is an image of the ball of radius r in the domain, which can be obtained from Lemma 14. Then the radius $\hat{r} = R_f(r) = (R_g(r)R_\Sigma(r) + \alpha)r$. We can apply this formula recursively to get the radius belonging to a deep SSM model containing several blocks.

Theorem 20 *Let $\mathcal{F}$ be a set of deep SSM models, namely let $f \in \mathcal{F}$ has the form $f = f^{\mathrm{Dec}} \circ f^{\mathrm{Pool}} \circ f^{\mathrm{B}_L} \circ \ldots \circ f^{\mathrm{B}_1} \circ f^{\mathrm{Enc}}$ with layer parameter sets $\mathcal{W}_{\mathrm{Dec}}$, $\mathcal{W}_{\mathrm{B}_L}, \ldots, \mathcal{W}_{\mathrm{B}_1}$ and $\mathcal{W}_{\mathrm{Enc}}$ respectively, where f^{B_i} is an SSM block for all i, i.e. $f^{\mathrm{B}_i}(\mathbf{z})[k] = g_i(\mathcal{S}_{\Sigma_i}(\mathbf{z})[k]) + \alpha_i \mathbf{z}[k]$ for all $k \in [T]$, and $\mathcal{W}_{\mathrm{B}_i} = \mathcal{E}_i \times \mathcal{W}_{g_i}$. Let Assumption 11 hold and let us assume that each $f \in \mathcal{F}$ maps from $B_{(\mathbb{R}^{n_u}, \|\cdot\|_2)}(K_{\mathbf{u}})$, and each set of nonlinearities g_i is (μ_{g_i}, c_{g_i})-RC.*

Under these assumptions, there exists $\hat{r} < \infty$ such that the image of each $f \in \mathcal{F}$ is contained in the ball $B_{(\mathbb{R}, |\cdot|)}(\hat{r})$ and the following holds with probability at least $1 - \delta$.

$$\mathbb{P}_{S \sim \mathcal{D}^N} \left[\forall f \in \mathcal{F} \quad \mathcal{L}(f) - \mathcal{L}^S_{emp}(f) \leq \frac{\mu K_{\mathbf{u}} L_l + c L_l}{\sqrt{N}} + K_l \sqrt{\frac{2\log(4/\delta)}{N}} \right], \tag{4}$$

$\mu \leq K_{\mathrm{Enc}} K_{\mathrm{Dec}} (\mu_{g_1} K_2 + \alpha_1) \prod_{i=2}^{L} (\mu_{g_i} K_1 + \alpha_i)$, $c \leq K_{\mathrm{Dec}} \sum_{j=1}^{L} \left[\prod_{i=j+1}^{L} (\mu_{g_i} K_1 + \alpha_i) \right] c_{g_j}$ and $K_l > 0$ such that $|l(\cdot, \cdot)| \leq K_l$. In particular, we obtain $K_l \leq 2 L_l \max\{K_{\mathrm{Dec}} \hat{r}, K_y\}$.

Proof [Proof of Theorem 20]

First, let us recite the definition of Rademacher complexity.

Definition 21 *(Shalev-Shwartz and Ben-David, 2014, Def. 26.1) The Rademacher complexity of a bounded set $\mathcal{A} \subset \mathbb{R}^m$ is defined as*

$$R(\mathcal{A}) = \mathbb{E}_\sigma \left[\sup_{a \in \mathcal{A}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i a_i \right],$$

where the random variables σ_i are i.i.d such that $\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma = -1] = 0.5$. The Rademacher complexity of a set of functions $\mathcal{F}$ over a set of samples $S = \{s_1 \ldots s_m\}$ is defined as $R_S(\mathcal{F}) = R(\{(f(s_1), \ldots, f(s_m)) \mid f \in \mathcal{F}\})$.

The following is a standard theorem we use in the proof.

Theorem 22 *(Shalev-Shwartz and Ben-David, 2014, Thm. 26.5) Let L_0 denote the set of functions of the form $(\mathbf{u}, y) \mapsto l(f(\mathbf{u}), y)$ for $f \in \mathcal{F}$. Let K_l be such that the functions from L_0 all take values from the interval $[0, K_l]$. Then for any $\delta \in (0, 1)$ we have*

$$\mathbb{P}_{\mathbf{S} \sim \mathcal{D}^N} \left(\forall f \in \mathcal{F} : \mathcal{L}(f) - \mathcal{L}^S_{emp}(f) \leq 2 R_S(L_0) + K_l \sqrt{\frac{2\log(4/\delta)}{N}} \right) \geq 1 - \delta.$$

We wish to apply the Theorem 22 to the set of deep SSM models $\mathcal{F}$. Let us fix a random sample $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_N\} \subset \left(\ell^{2,2}_T(\mathbb{R}^{n_{\mathrm{in}}}) \right)^N$. As the loss function is Lipschitz according to Assumption 11, we have that for any $f \in \mathcal{F}$

$$|l(f(\mathbf{u}), y)| \leq 2 L_l \max\{f(\mathbf{u}), y\} \leq 2 L_l \max\{K_{\mathrm{Dec}} \hat{r}, K_y\},$$

thus $K_l \leq 2L_l \max\{K_{\text{Dec}} \hat{r}\, K_y\}$. The constant $\hat{r}$ exists as a corollary of Lemma 14, see Remark 19.

Again by the Lipschitzness of the loss and the Contraction lemma (Shalev-Shwartz and Ben-David, 2014, Lemma 26.9) we have

$$R_S(L_0) \leq L_l \cdot R_S(\mathcal{F}).$$

It is enough to bound the Rademacher complexity of $\mathcal{F}$ to conclude the proof. Let us consider the deep SSM models as a composite of mappings as

$$B_{\ell_T^{2,2}(\mathbb{R}^{n_{\text{in}}})}(K_{\mathbf{u}}) \xrightarrow{\text{Encoder}} B_{\ell_T^{2,2}(\mathbb{R}^{n_u})}(K_{\mathbf{u}} K_{\text{Enc}}) \xrightarrow{\text{B}_1} B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r_1) \xrightarrow{\text{B}_2} \dots \xrightarrow{\text{B}_L}$$
$$B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r_{L+1}) \xrightarrow{\text{Pooling}} (\mathbb{R}^{n_u}, \|\cdot\|_\infty) \xrightarrow{\text{Decoder}} (\mathbb{R}, |\cdot|),$$

where the constants r_i exist as a corollary of Lemma 14, see Remark 19. Therefore, the SSM layer in the first SSM block is considered as a map $B_{\ell_T^{2,2}(\mathbb{R}^{n_u})}(K_{\text{Enc}} K_{\mathbf{u}}) \to B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r_1)$, while the rest of the SSM layers in the SSM blocks are considered as a map $B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r_i) \to B_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})}(r_{i+1})$. This is needed, because the Encoder is constant in time, therefore the Composition Lemma wouldn't be able to carry the $\ell_T^{2,2}$ norm of the input through the chain of estimation along the entire model. This is one of the key technical points which makes it possible to establish a time independent bound.

By the conditions of the Theorem and the stability assumption in Assumption 11 we have that the Encoder, Decoder, Pooling, SSM and MLP layers are each (μ, c)-RC for some μ and c from Lemma 14. By Lemma 2 we have that the composition of an SSM layer and an MLP is (μ, c)-RC. A residual SSM block is then $(\mu + \alpha, c)$-RC, because

$$\mathbb{E}_\sigma \left[\sup_{g \circ \mathcal{S}_\Sigma} \sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i (g(\mathcal{S}_\Sigma(\mathbf{z}_i)) + \alpha \mathbf{z}_i) \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right] \leq$$
$$\mathbb{E}_\sigma \left[\sup_{g \circ \mathcal{S}_\Sigma} \sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i g(\mathcal{S}_\Sigma(\mathbf{z}_i)) \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right] + \alpha \mathbb{E}_\sigma \left[\sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i \mathbf{z}_i \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right]$$
$$\leq (\mu + \alpha) \mathbb{E}_\sigma \left[\sup_{\{\mathbf{z}_i\}_{i=1}^N \in Z} \left\| \frac{1}{N} \sum_{i=1}^N \sigma_i \mathbf{z}_i \right\|_{\ell_T^{\infty,\infty}(\mathbb{R}^{n_u})} \right] + \frac{c}{\sqrt{N}}$$

Hence, by Corollary 17, the whole deep SSM model is (μ, c)-RC as a map between $X_1 = B_{\ell_T^{2,2}(\mathbb{R}^{n_{\text{in}}})}(K_{\mathbf{u}})$ and $X_2 = (\mathbb{R}, |\cdot|)$. The Theorem is then a direct corollary of the following Lemma.

Lemma 23 *Let $\mathcal{F}$ be a set of functions between $X_1 = B_{\ell_T^{2,2}(\mathbb{R}^{n_{\text{in}}})}(K_{\mathbf{u}})$ and $X_2 = (\mathbb{R}, |\cdot|)$ that is (μ, c)-RC. The Rademacher complexity of $\mathcal{F}$ w.r.t. some dataset S for which Assumption 11 holds, admits the following inequality.*

$$R_S(\mathcal{F}) \leq \frac{\mu K_{\mathbf{u}} + c}{\sqrt{N}}.$$

Proof

$$R_S(\mathcal{F}) = R(\{(f(\mathbf{u}_1), \dots, f(\mathbf{u}_N))^T \mid f \in \mathcal{F}\}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i f(\mathbf{u}_i) \right]$$
$$\leq \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{N} \sigma_i f(\mathbf{u}_i) \right| \right] \leq \mu \mathbb{E}_\sigma \left[\left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{u}_i \right\|_{\ell_T^{2,2}(\mathbb{R}^{n_{\text{in}}})} \right] + \frac{c}{\sqrt{N}}$$

By definition

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{u}_i \right\|_{\ell_T^{2,2}(\mathbb{R}^{n_{\text{in}}})} = \sqrt{\sum_{k=1}^{T} \left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{u}_i[k] \right\|_2^2}$$
$$= \sqrt{\sum_{k=1}^{T} \left\langle \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{u}_i[k], \frac{1}{N} \sum_{j=1}^{N} \sigma_j \mathbf{u}_j[k] \right\rangle_{\mathbb{R}^{n_{\text{in}}}}}$$
$$= \sqrt{\sum_{k=1}^{T} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_i \sigma_j \left\langle \mathbf{u}_i[k], \mathbf{u}_j[k] \right\rangle_{\mathbb{R}^{n_{\text{in}}}}}$$

Therefore

$$\mathbb{E}_\sigma \left[\left\| \frac{1}{N} \sum_{i=1}^{N} \sigma_i \mathbf{u}_i \right\|_{\ell_T^{2,2}(\mathbb{R}^{n_{\text{in}}})} \right] = \mathbb{E}_\sigma \left[\sqrt{\sum_{k=1}^{T} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_i \sigma_j \left\langle \mathbf{u}_i[k], \mathbf{u}_j[k] \right\rangle_{\mathbb{R}^{n_{\text{in}}}}} \right]$$
$$\leq \sqrt{\mathbb{E}_\sigma \left[\sum_{k=1}^{T} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_i \sigma_j \left\langle \mathbf{u}_i[k], \mathbf{u}_j[k] \right\rangle_{\mathbb{R}^{n_{\text{in}}}} \right]}$$
$$= \sqrt{\sum_{k=1}^{T} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}_\sigma \left[\sigma_i \sigma_j \right] \left\langle \mathbf{u}_i[k], \mathbf{u}_j[k] \right\rangle_{\mathbb{R}^{n_{\text{in}}}}}$$
$$= \sqrt{\sum_{k=1}^{T} \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}_\sigma \left[\sigma_i^2 \right] \left\langle \mathbf{u}_i[k], \mathbf{u}_i[k] \right\rangle_{\mathbb{R}^{n_{\text{in}}}}}$$
$$= \sqrt{\frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{T} \|\mathbf{u}_i[k]\|_2^2} = \sqrt{\frac{1}{N^2} \sum_{i=1}^{N} \|\mathbf{u}_i\|_{\ell_T^{2,2}(\mathbb{R}^{n_{\text{in}}})}^2} \leq \sqrt{\frac{1}{N^2} N K_{\mathbf{u}}^2} \leq \frac{K_{\mathbf{u}}}{\sqrt{N}}$$

Hence we have

$$R_S(\mathcal{F}) \leq \frac{\mu K_{\mathbf{u}} + c}{\sqrt{N}}$$

■

The constants μ and c are obtained by substituting the results of Lemma 14 into the Corollary 17.

■

References

Nil-Jana Akpinar, Bernhard Kratzwald, and Stefan Feuerriegel. Sample complexity bounds for rnns with application to combinatorial graph problems (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13745–13746, 4 2020. doi: 10.1609/aaai.v34i10.7144. URL https://ojs.aaai.org/index.php/AAAI/article/view/7144.

Athanasios C Antoulas. *Approximation of large-scale dynamical systems*. SIAM, 2005.

VS Chellaboina, WM Haddad, DS Bernstein, and DA Wilson. Induced convolution operator norms for discrete-time linear systems. In *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No. 99CH36304)*, volume 1, pages 487–492. IEEE, 1999.

Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks. In *Proceedings of AISTATS 2020*, volume 108 of *PMLR*, pages 1233–1243, 8 2020.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022.

Bruce Hajek and Maxim Raginsky. Ece 543: Statistical learning theory. *University of Illinois lecture notes*, 2019.

Joshua Hanson and Maxim Raginsky. Rademacher complexity of neural odes via chen-fliess series. *arXiv preprint arXiv:2401.16655*, 2024.

Joshua Hanson, Maxim Raginsky, and Eduardo Sontag. Learning recurrent neural net models of nonlinear systems. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *PMLR*, pages 425–435. PMLR, 6 2021.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Boris Joukovsky, Tanmoy Mukherjee, Huynh Van Luong, and Nikos Deligiannis. Generalization error bounds for deep unfolding rnns. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *PMLR*, pages 1515–1524. PMLR, 7 2021.

Pascal Koiran and Eduardo D. Sontag. Vapnik-chervonenkis dimension of recurrent neural networks. *Discrete Applied Mathematics*, 86(1):63–79, 1998.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. *arXiv preprint arXiv:2303.06349*, 2023.

Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, and Lei Zhang. Lipsformer: Introducing lipschitz continuity to vision transformers. *arXiv preprint arXiv:2304.09856*, 2023.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022.

Eduardo D Sontag. A learning result for continuous-time recurrent neural networks. *Systems & control letters*, 34(3):151–158, 1998.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2024.

Lan V Truong. Generalization error bounds on deep learning with markov datasets. *Advances in Neural Information Processing Systems*, 35:23452–23462, 2022a.

Lan V Truong. On rademacher complexity-based generalization bounds for deep learning. *arXiv preprint arXiv:2208.04284*, 2022b.

Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jiong Zhang, Qi Lei, and Inderjit Dhillon. Stabilizing gradients for deep neural networks via efficient SVD parameterization. In *35th ICML*, volume 80 of *PMLR*, pages 5806–5814. PMLR, 7 2018.