

# Lipschitz constant estimation for general neural network architectures using control tools

Patricia Pauli<sup>1</sup>, Dennis Gramlich<sup>2</sup>, and Frank Allgöwer<sup>1</sup>

**Abstract**—This paper is devoted to the estimation of the Lipschitz constant of neural networks using semidefinite programming. For this purpose, we interpret neural networks as time-varying dynamical systems, where the  $k$ -th layer corresponds to the dynamics at time  $k$ . A key novelty with respect to prior work is that we use this interpretation to exploit the series interconnection structure of neural networks with a dynamic programming recursion. Nonlinearities, such as activation functions and nonlinear pooling layers, are handled with integral quadratic constraints. If the neural network contains signal processing layers (convolutional or state space model layers), we realize them as 1-D/2-D/N-D systems and exploit this structure as well. We distinguish ourselves from related work on Lipschitz constant estimation by more extensive structure exploitation (scalability) and a generalization to a large class of common neural network architectures. To show the versatility and computational advantages of our method, we apply it to different neural network architectures trained on MNIST and CIFAR-10.

**Index Terms**—Neural networks, Lipschitz constant, semidefinite program.

## I. INTRODUCTION

NEURAL networks (NNs) are successfully applied in many fields, e.g., in data analysis, pattern recognition, image and video processing, natural language processing, and control [1], [2]. Especially in safety critical systems like autonomous driving, it is imperative that NNs are safe and reliable [3]. The Lipschitz constant of the input-output mapping defined by an NN is closely linked to the robustness of the NN [4], and given that the calculation of the Lipschitz constant for NNs is an NP-hard problem [5], [6], there is a high interest to instead find accurate upper bounds on this Lipschitz constant [7]–[11]. Trivial methods like the product of the spectral norms of the weights [4] can cheaply be computed by the power iteration method, but the resulting bounds can be quite loose, especially for deep NNs. In contrast, semidefinite-programming (SDP) based approaches [10], [11] provide tighter bounds at the price of a computational overhead. In this work, we present SDPs that provide significantly lower Lipschitz bounds for NNs than the commonly used spectral

norm bounds, while showing better scalability than other SDP-based approaches, e.g., LipSDP [10]. We further generalize SDP-based methods for Lipschitz constant estimation to a large class of NNs.

SDP-based methods provide the tightest bounds on the  $\ell_2$  Lipschitz constant for NNs in polynomial time [10]. However, their scalability to deep state-of-the-art NNs is an open research problem which is actively investigated. [12], [13] develop more scalable SDP solvers and [14] do so specific to the problem of SDP-based Lipschitz constant estimation, [15] exploit the chordal sparsity pattern of the underlying linear matrix inequality (LMI) constraint for fully connected NNs and [16], [17] exploit the structure of convolutions. This paper builds on [10], [16], [17] to develop a general, accurate, and scalable SDP-based method for Lipschitz constant estimation for a large family of NN architectures. For improved scalability, we exploit (i) the structure of the individual layer types and (ii) the concatenation structure of the feedforward networks. We do the latter by taking on a dynamic programming perspective and interpreting the layers as the dynamics of a system. This view is novel in this context, leading to a recursive formulation of layer-wise constraints, which is computationally favorable.

In contrast to previous works [10], [17], our approach incorporates many popular layer types including convolutional, deconvolutional and state space model layers [18], residual layers [19], fully connected layers, average and maximum pooling layers, and slope-restricted and GroupSort activation function layers [20]. We especially exploit the structure and shift invariance of convolutions as we incorporate 1-D/2-D/N-D convolutions into the SDP-based analysis using a state space representation of the Roesser type [17], [21], [22].

In summary, the main contribution of this work is a SDP-based method for Lipschitz constant estimation for a general class of NNs that outperforms existing methods in terms of scalability and accuracy. To reach our goal, we exploit N-D systems theory and introduce a dynamic programming perspective for the underlying problem. The remainder of the paper is organized as follows. Section II formally states the problem, introduces all layer definitions and state space representations for convolutions. Next, Section III involves our dynamic programming based approach for Lipschitz constant estimation for NNs and Section IV discusses sources of conservatism. Finally, Section V applies our method on multiple neural network architectures to demonstrate the versatility and improved accuracy and scalability of our approach over previous approaches. We provide easy-to-use code for all considered neural network architectures and layer types.

\*This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - 390740016 and under grant 468094890. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Patricia Pauli.

<sup>1</sup>Patricia Pauli and Frank Allgöwer are with the Institute for Systems Theory and Automatic Control, University of Stuttgart, 70550 Stuttgart, Germany (email: {patricia.pauli, frank.allgower}@ist.uni-stuttgart.de)

<sup>2</sup>Dennis Gramlich is with the Chair of Intelligent Control Systems, RWTH Aachen, 52074 Aachen, Germany (e-mail: dennis.gramlich@ic.rwth-aachen.de).

**Notation:** By  $\|\cdot\|_2$  we either mean the Euclidean norm of a vector or the  $\ell_2$  norm of a signal. By  $\langle \cdot, \cdot \rangle_2$  we denote the  $\ell_2$  inner product. By  $\mathbb{R}^n$  ( $\mathbb{R}_+^n$ ), we mean the space of  $n$ -dimensional vectors with real (positive) entries. By  $\mathbb{S}^n$  ( $\mathbb{S}_+^n$ ), we denote (positive definite) symmetric matrices and by  $\mathbb{D}^n$  ( $\mathbb{D}_+^n$ ) we mean (positive definite) diagonal matrices of dimension  $n$ , respectively. Within our paper, we study convolutional neural networks processing image signals. For this purpose, we understand an image as a sequence  $(u[i_1, \dots, i_d])$  with free variables  $i_1, \dots, i_d \in \mathbb{N}_0$ . In this sequence,  $u[i_1, \dots, i_d]$  is an element of  $\mathbb{R}^c$ , where  $c$  is called the channel dimension (e.g.,  $c = 3$  for RGB images). The *signal dimension*  $d$  will usually be  $d = 2$  for images or  $d = 3$  for medical images. The space of such signals/sequences is denoted by  $\ell_{2e}^c(\mathbb{N}_0^d) := \{u : \mathbb{N}_0^d \rightarrow \mathbb{R}^c\}$ . Images should be understood as sequences in  $\ell_{2e}^c(\mathbb{N}_0^d)$  with a finite square as support. For convenience, we will sometimes use multi-index notation for signals, i.e., we denote  $u[i_1, \dots, i_d]$  as  $u[\mathbf{i}]$  for  $\mathbf{i} \in \mathbb{N}_0^d$ . For these multi-indices, we use the notation  $\mathbf{i} + \mathbf{j}$  for  $(i_1 + j_1, \dots, i_d + j_d)$ ,  $\mathbf{i}\mathbf{j} = (i_1 j_1, \dots, i_d j_d)$  and  $\mathbf{i} \leq \mathbf{j}$  for  $i_1 \leq j_1, \dots, i_d \leq j_d$ . We further denote by  $[\mathbf{i}, \mathbf{j}] = \{\mathbf{t} \in \mathbb{N}_0^d \mid \mathbf{i} \leq \mathbf{t} \leq \mathbf{j}\}$  the *interval* of all multi-indices between  $\mathbf{i}, \mathbf{j} \in \mathbb{N}_0^d$  and by  $|\mathbf{i}, \mathbf{j}|$  the number of elements in this interval. Finally, we define the interval  $[\mathbf{i}, \mathbf{j}] := [\mathbf{i}, \mathbf{j} - 1]$ .

## II. PROBLEM STATEMENT AND DEEP NEURAL NETWORKS

In this work, we understand deep neural networks as a concatenation of simple functions, i.e., as a composition

$$\text{NN}_\theta = \ell_l \circ \ell_{l-1} \circ \dots \circ \ell_2 \circ \ell_1 \quad (1)$$

of layers  $\ell_k, k = 1, \dots, l$  where  $k$  is the layer index and  $\ell \in \{\mathcal{L}, \mathcal{C}, \mathcal{S}, \sigma, \mathcal{P}, \mathcal{F}\}$  is either a linear layer  $\mathcal{L}$ , a convolutional layer  $\mathcal{C}$ , a state space model layer  $\mathcal{S}$ , an activation function layer  $\sigma$ , a pooling layer  $\mathcal{P}$ , or a flattening layer  $\mathcal{F}$ . The parameter  $\theta$  of the neural network  $\text{NN}_\theta$  refers to the collection of parameters (weights and biases)  $\theta_k$  of all the individual layers. We can also write down the NN recursively as the map from  $u_1$  to  $y_l$  defined by

$$y_k = \ell_k(u_k) \quad u_{k+1} = y_k \quad k = 1, \dots, l, \quad (2)$$

where  $u_k \in \mathcal{D}_{k-1}$  and  $y_k \in \mathcal{D}_k$  denote the input and the output of each layer and the real vector spaces  $\mathcal{D}_{k-1}$  and  $\mathcal{D}_k$  are the input and output domain of the layer  $\ell_k$ . We assume here that the layers are always chosen in such a way that the image space of  $\ell_k$  and the domain space of  $\ell_{k+1}$  coincide. Consequently, our Lipschitz constant analysis applies to any finite concatenation of layers  $\ell \in \{\mathcal{L}, \mathcal{C}, \mathcal{S}, \sigma, \mathcal{P}, \mathcal{F}\}$ . In deep learning, the definition of a layer may sometimes refer to a composition of multiple elements of  $\{\mathcal{L}, \mathcal{C}, \mathcal{S}, \sigma, \mathcal{P}, \mathcal{F}\}$ . For example, a linear map is grouped with a diagonally repeated activation function or a convolutional layer, an activation function, and a pooling layer are grouped together as a layer. Our approach can handle such concatenated layer definitions, meaning that we additionally allow  $\ell \in \{\sigma \circ \mathcal{L}, \sigma \circ \mathcal{C}, \mathcal{P} \circ \sigma \circ \mathcal{C}\}$  or a concatenation of even more layers, cmp. Section III-G.

Regardless of the layer definition, our examples usually study convolutional neural networks  $\text{CNN}_\theta$  with the structure

$$\begin{aligned} \mathcal{L}_l \circ \sigma_{l-1} \circ \dots \circ \sigma_{p+2} \circ \mathcal{L}_{p+1} \circ \mathcal{F}_p \circ \dots \\ \dots \circ \mathcal{P}_{p-1} \circ \sigma_{p-2} \circ \mathcal{C}_{p-3} \circ \dots \circ \mathcal{P}_3 \circ \sigma_2 \circ \mathcal{C}_1, \end{aligned}$$

typically found in image classification. These convolutional neural networks (CNNs) are composed of fully connected layers  $\mathcal{L}_k$ , activation function layers  $\sigma_k$ , a flattening operation  $\mathcal{F}_p$ , convolutional layers  $\mathcal{C}_k$ , and (optional) pooling layers  $\mathcal{P}_k$  in the order shown above.

The goal of this work is to provide an accurate and scalable method that determines an upper bound on the Lipschitz constant of a neural network (1) (2).

**Problem 1.** For a given neural network  $\text{NN}_\theta$  with parameters  $\theta$ , find an upper bound on the Lipschitz constant, i.e., find a value  $\gamma \geq 0$  such that

$$\|\text{NN}_\theta(u^1) - \text{NN}_\theta(u^2)\|_2 \leq \gamma \|u^1 - u^2\|_2$$

for all  $u^1, u^2 \in \mathcal{D}_0$ .

We notice that the definition of an NN (1), (2) resembles a dynamical system  $u_{k+1} = \ell_k(u_k)$ . The interpretation of an NN as a dynamical system with time-varying dynamics  $\ell_k$  and state  $u_k$  is very powerful because it enables us to use tools from control and systems theory to analyze properties of NNs. However, we stress that this interpretation should be taken with caution, since the inputs  $u_k$  and  $u_j$  for  $k \neq j$  usually live in different spaces  $\mathcal{D}_{k-1}$  and  $\mathcal{D}_{j-1}$ . As we will see in Section II-A, we allow signal spaces  $\mathcal{D}_k = \ell_{2e}^{c_k}(\mathbb{N}_0^{d_k})$  of  $d_k$ -dimensional signals as well as vector spaces  $\mathcal{D}_k = \mathbb{R}^{c_k}$ . Also the vector (= channel) dimension  $c_k$  may differ from one layer to another. In addition, the nature of the mappings  $\ell_k$  and  $\ell_j$  for  $k \neq j$  can be completely different, including linear and nonlinear mappings. In some less heterogeneous examples, e.g., fully connected networks  $\mathcal{L}_l \circ \sigma_{l-1} \circ \dots \circ \sigma_2 \circ \mathcal{L}_1$ , or fully-convolutional networks and subnetworks  $\mathcal{C}_l \circ \sigma_{l-1} \circ \dots \circ \sigma_2 \circ \mathcal{C}_1$ , this interpretation as dynamical systems is, however, more natural and it is common practice to design a deep backbone of NNs of the same layer type [23].

It is the defining selling point of our work that we *exploit the structure of each layer*, as well as the *composition structure of the NN itself* using different perspectives and methods from control. We have described the NN clearly and yet in sufficient detail as a concatenation of its layers, which is now followed by a definition of each individual layer in Section II-A. Subsequently, in Section II-B, we introduce state space representations for convolutional layers.

### A. Layer definitions

**Convolutional layer:** A convolutional layer  $\mathcal{C}_k$  is a mapping from  $\mathcal{D}_{k-1} = \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}})$  to  $\mathcal{D}_k = \ell_{2e}^{c_k}(\mathbb{N}_0^{d_k})$  which is defined by a convolution kernel  $K_k[\mathbf{t}] \in \mathbb{R}^{c_k \times c_{k-1}}$  for  $0 \leq \mathbf{t} \leq \mathbf{r}_k$  and a bias  $b_k \in \mathbb{R}^{c_k}$ . We write

$$y_k[\mathbf{i}] = b_k + \sum_{0 \leq \mathbf{t} \leq \mathbf{r}_k} K_k[\mathbf{t}] u_k[\mathbf{i} - \mathbf{t}], \quad (3)$$

where  $u_k[\mathbf{i} - \mathbf{t}]$  is set to zero if  $\mathbf{i} - \mathbf{t}$  is not in the domain of  $u_k[\cdot]$ , which accounts for possible zero-padding. A convolutional layer retains the dimension  $d_{k-1} = d_k = d$  but it may change in channel size from  $c_{k-1}$  to  $c_k$ . The multi-index  $\mathbf{r}_k \in \mathbb{N}_0^d$  defines the size of the kernel  $K_k[\cdot]$ .

Our compact description of a convolution (3) includes N-D convolutions ( $d = N$ ). For instance, a 1-D convolution ( $d = 1$ )

$$y_k[i] = b_k + \sum_{t=0}^{r_k} K_k[t] u_k[i - t] \quad (4)$$

operates on a 1-D signal, e.g., a time signal, a 2-D convolution ( $d = 2$ )

$$y_k[i_1, i_2] = b_k + \sum_{t_1=0}^{r_{k1}} \sum_{t_2=0}^{r_{k2}} K_k[t_1, t_2] u_k[i_1 - t_1, i_2 - t_2] \quad (5)$$

operates on signals with two propagation dimensions, e.g., images, and an N-D convolution considers inputs with even more input dimensions, e.g., 3-D convolutions may be used for videos or 3-D medical images.

An extension of the convolutional layer (3) is a strided convolutional layer  $\mathcal{C}_{\mathbf{s}_k}$  with stride  $\mathbf{s}_k \in \mathbb{N}^d$ . For convolutions with stride  $\mathbf{s}_k = (s_{k1}, \dots, s_{kd})$ , the output is not given by (3), but by

$$y_k[\mathbf{i}] = b_k + \sum_{0 \leq \mathbf{t} \leq \mathbf{r}_k} K_k[\mathbf{t}] u_k[\mathbf{s}_k \mathbf{i} - \mathbf{t}]. \quad (6)$$

This means that we always shift the kernel by  $s_{k1}, \dots, s_{kd}$  along the respective signal dimension  $1, \dots, d$ .

**Activation function layer:** An activation function layer  $\sigma_k$  can be applied to any of our domain spaces  $\mathcal{D}_{k-1} = \mathbb{R}^{c_{k-1}}$  or  $\mathcal{D}_{k-1} = \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}})$ , but it requires  $\mathcal{D}_k \cong \mathcal{D}_{k-1}$ . We consider activation functions that are defined by scalar activation functions  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  that are applied element-wise if applied to a vector  $u_k \in \mathbb{R}^{c_k}$ . To this end, for finite dimensional vector spaces,  $\sigma$  is identified with the function

$$\sigma : \mathbb{R}^{c_k} \rightarrow \mathbb{R}^{c_k}, u_k \mapsto y_k = [\sigma(u_{k1}) \quad \dots \quad \sigma(u_{kc_k})]^\top.$$

We further lift the scalar activation function to signal spaces by defining the activation function layer on  $\ell_{2e}^{c_k}(\mathbb{N}_0^{d_k})$  as the function  $\sigma : \ell_{2e}^{c_k}(\mathbb{N}_0^{d_k}) \rightarrow \ell_{2e}^{c_k}(\mathbb{N}_0^{d_k})$ ,

$$(u_k[\mathbf{i}])_{\mathbf{i} \in \mathbb{N}_0^{d_k}} \mapsto (y_k[\mathbf{i}])_{\mathbf{i} \in \mathbb{N}_0^{d_k}} = (\sigma(u_k[\mathbf{i}]))_{\mathbf{i} \in \mathbb{N}_0^{d_k}}.$$

**Fully connected layer:** In the case of a fully connected layer  $\mathcal{L}_k$  the domain and image spaces are  $\mathcal{D}_{k-1} = \mathbb{R}^{c_{k-1}}$  and  $\mathcal{D}_k = \mathbb{R}^{c_k}$ , i.e., there are only the channel dimensions  $c_{k-1}, c_k$  (= number of neurons of the input and output layer) and no signal dimensions  $d_{k-1}, d_k$ , i.e.,  $d_{k-1} = d_k = 0$ . We define a fully connected layer as an affine function

$$\mathcal{L}_k : \mathbb{R}^{c_{k-1}} \rightarrow \mathbb{R}^{c_k}, u_k \mapsto y_k = b_k + W_k u_k. \quad (7)$$

The vector  $b_k \in \mathbb{R}^{c_k}$  is called the bias and  $W_k \in \mathbb{R}^{c_k \times c_{k-1}}$  is called the weight matrix.

**Remark 1.** Note that the fully connected layer is a special case of a convolutional layer for  $d_{k-1} = d_k = 0$ . Indeed,  $\mathbb{R}^{c_{k-1}} \cong \ell_{2e}^{c_{k-1}}(\{0\}) = \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^0)$  and  $\mathbb{R}^{c_k} \cong \ell_{2e}^{c_k}(\mathbb{N}_0^0)$ . Furthermore, we can understand  $W_k$  as the convolution kernel

which, in the case  $d_k = 0$ , is given by  $K_k[0] := W_k$ . Consequently, all results presented in this work for convolutional layers automatically hold for fully connected layers.

**Pooling layer:** Pooling layers are down-sampling operations from  $\mathcal{D}_{k-1} = \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}})$  to  $\mathcal{D}_k = \ell_{2e}^{c_k}(\mathbb{N}_0^{d_k})$  with  $d_{k-1} = d_k = d$  and  $c_{k-1} = c_k$  that take a batch of input signal entries  $(u_k[\mathbf{s}_k \mathbf{i} + \mathbf{t}] \mid \mathbf{t} \in [0, \mathbf{r}_k])$  and map them channel-wise into one single output signal entry  $y[\mathbf{i}]$ . The two most common pooling layers are average pooling  $\mathcal{P}_k^{\text{av}} : \ell_{2e}^{c_k}(\mathbb{N}_0^d) \rightarrow \ell_{2e}^{c_k}(\mathbb{N}_0^d)$ ,

$$y_k[\mathbf{i}] := \text{mean}(u_k[\mathbf{s}_k \mathbf{i} - \mathbf{t}] \mid \mathbf{t} \in [0, \mathbf{r}_k]) \\ = \frac{1}{|[0, \mathbf{r}_k]|} \sum_{0 \leq \mathbf{t} \leq \mathbf{r}_k} u_k[\mathbf{s}_k \mathbf{i} - \mathbf{t}]$$

and maximum pooling  $\mathcal{P}_k^{\text{max}} : \ell_{2e}^{c_k}(\mathbb{N}_0^d) \rightarrow \ell_{2e}^{c_k}(\mathbb{N}_0^d)$ ,

$$y_k[\mathbf{i}] := \max(u_k[\mathbf{s}_k \mathbf{i} - \mathbf{t}] \mid \mathbf{t} \in [0, \mathbf{r}_k]),$$

where the maximum is applied channel-wise. For most pooling layers the kernel size and the stride coincide ( $\mathbf{r}_k = \mathbf{s}_k$ ), yet sometimes, e.g., in AlexNet [24],  $\mathbf{r}_k > \mathbf{s}_k$  is chosen.

**Flattening operator:** Flattening is a pure reshaping operation, which merges the signal dimensions into the channel dimension. Note that the mapping is not injective, i.e., a square batch  $(u_k[\mathbf{i}] \mid 0 \leq \mathbf{i} < \mathbf{N}_k)$ , for example a finite-dimensional image, is reshaped into the channel dimension and the remaining entries (mostly zeros) are discarded. The typical flattening operation is a vectorization given by

$$\mathcal{F}_k : \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}}) \rightarrow \mathbb{R}^{[0, \mathbf{N}_k] \cdot c_{k-1}}, (u_k[\mathbf{i}])_{\mathbf{i} \in \mathbb{N}_0^{d_{k-1}}} \mapsto y_k,$$

where  $y_k$  is a stacked vector of  $u_k[\mathbf{i}]$ ,  $0 \leq \mathbf{i} < \mathbf{N}_k$ , i.e.,  $y_k^\top = [u_k[0, \dots, 0]^\top \quad \dots \quad u_k[\mathbf{N}_{k1}, \dots, \mathbf{N}_{kd}]^\top]$ . We could also define flattening operators  $\ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}}) \rightarrow \ell_{2e}^{c_k}(\mathbb{N}_0^{d_k})$  with  $1 \leq d_k < d_{k-1}$  contracting only some of the signal dimensions and not all at once. For example, we can flatten 2-D signals into 1-D signals ( $d_k = 1$ ) or into vectors with  $d_k = 0$ .

**State space model layer:** We define a state space model layer as an affine time-invariant system  $\mathcal{S}_k : \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^1) \rightarrow \ell_{2e}^{c_k}(\mathbb{N}_0^1)$

$$\begin{bmatrix} x[i+1] \\ y[i] \end{bmatrix} = \begin{bmatrix} f_k & A_k & B_k \\ g_k & C_k & D_k \end{bmatrix} \begin{bmatrix} 1 \\ x[i] \\ u[i] \end{bmatrix},$$

where  $x[i] \in \mathbb{R}^n$  denotes the state. The state space model is characterized by some matrices  $(A_k, B_k, C_k, D_k, f_k, g_k)$  of appropriate dimensions.

## B. State space representations for convolutions

In the machine learning literature, convolutional layers are usually represented as in (3) using a convolution kernel [25]. However, state space realizations have proven to be more amenable to analysis using tools from robust control than such kernel (impulse response) representations [17]. In the control engineering literature, mappings from  $\ell_{2e}^{c_{k-1}}(\mathbb{N}_0^d)$  to  $\ell_{2e}^{c_k}(\mathbb{N}_0^d)$  are known as N-D systems and, as it is shown in [26], N-D systems with rational transfer functions admit a state

space realization as a so-called Roesser model [21], defined as follows.

**Definition 1** (Roesser model). *An affine  $N$ -D system  $\ell_{2e}^{c_k-1}(\mathbb{N}_0^d) \rightarrow \ell_{2e}^{c_k}(\mathbb{N}_0^d), (u[\mathbf{i}]) \mapsto (y[\mathbf{i}])$  is described by a Roesser model as*

$$\begin{bmatrix} x_1[\mathbf{i} + e_1] \\ \vdots \\ x_d[\mathbf{i} + e_d] \\ y[\mathbf{i}] \end{bmatrix} = \begin{bmatrix} f_1 & A_{11} & \cdots & A_{1d} & B_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_d & A_{d1} & \cdots & A_{dd} & B_d \\ g & C_1 & \cdots & C_d & D \end{bmatrix} \begin{bmatrix} 1 \\ x_1[\mathbf{i}] \\ \vdots \\ x_d[\mathbf{i}] \\ u[\mathbf{i}] \end{bmatrix}, \quad (8)$$

where  $e_i$  denotes the unit vector with 1 in the  $i$ -th position. Here, the collection of matrices  $f_1, f_2, \dots, C_d, D$  is called state space representation of the system,  $x_1[\mathbf{i}] \in \mathbb{R}^{n_1}, \dots, x_d[\mathbf{i}] \in \mathbb{R}^{n_d}$  are the states,  $u[\mathbf{i}] \in \mathbb{R}^{c_k-1}$  is the input and  $y[\mathbf{i}] \in \mathbb{R}^{c_k}$  is the output of the system. We call (8) a linear time-invariant  $N$ -D system ( $N = d$ ) if  $f_1 = f_2 = \dots = f_d = 0$  and  $g = 0$ . Otherwise, we call the system affine time-invariant.

Throughout this section, we drop the layer index  $k$  to improve readability and we further define

$$\left[ \begin{array}{c|cc|c} f & \mathbf{A} & \mathbf{B} \\ \hline g & \mathbf{C} & \mathbf{D} \end{array} \right] := \begin{bmatrix} f_1 & A_{11} & \cdots & A_{1d} & B_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_d & A_{d1} & \cdots & A_{dd} & B_d \\ \hline g & C_1 & \cdots & C_d & D \end{bmatrix}.$$

Realizing 1-D convolutions in state space is straightforward [16]. For the important layer type of 2-D convolutions ( $d = 2$ ), i.e., the 2-D system

$$\begin{bmatrix} x_1[i_1 + 1, i_2] \\ x_2[i_1, i_2 + 1] \\ y[i_1, i_2] \end{bmatrix} = \begin{bmatrix} f_1 & A_{11} & A_{12} & B_1 \\ f_2 & A_{21} & A_{22} & B_2 \\ g & C_1 & C_2 & D \end{bmatrix} \begin{bmatrix} 1 \\ x_1[i_1, i_2] \\ x_2[i_1, i_2] \\ u[i_1, i_2] \end{bmatrix}, \quad (9)$$

we use the construction presented in [22] as stated in Lemma 1.

**Lemma 1** (Realization of 2-D convolutions [22]). *Consider a convolutional layer  $C : \ell_{2e}^{c_k-1}(\mathbb{N}_0^2) \rightarrow \ell_{2e}^{c_k}(\mathbb{N}_0^2)$  with representation (5) characterized by the convolution kernel  $K$  and the bias  $b$ . This layer is realized in state space by the matrices*

$$\left[ \begin{array}{c|c} A_{12} & B_1 \\ \hline C_2 & D \end{array} \right] = \begin{bmatrix} K[r_1, r_2] & \cdots & K[r_1, 1] & K[r_1, 0] \\ \vdots & \ddots & \vdots & \vdots \\ K[1, r_2] & \cdots & K[1, 1] & K[1, 0] \\ \hline K[0, r_2] & \cdots & K[0, 1] & K[0, 0] \end{bmatrix},$$

$$\left[ \begin{array}{c} A_{11} \\ \hline C_1 \end{array} \right] = \begin{bmatrix} 0 & 0 \\ I & 0 \\ 0 & I \end{bmatrix}, \quad \left[ \begin{array}{c|c} A_{22} & B_2 \end{array} \right] = \begin{bmatrix} 0 & I & 0 \\ 0 & 0 & I \end{bmatrix},$$

$$A_{21} = 0, \quad f_1 = 0, \quad f_2 = 0, \quad g = b,$$

where  $K[i_1, i_2] \in \mathbb{R}^{c_k \times c_k-1}$ ,  $i_1 \in [0, r_1]$ ,  $i_2 \in [0, r_2]$ . The state signals  $(x_1[i_1, i_2])_{i_1, i_2 \in \mathbb{N}_0}$  with  $x_1[i_1, i_2] \in \mathbb{R}^{n_1}$ ,  $n_1 = c_k r_1$ , and  $(x_2[i_1, i_2])_{i_1, i_2 \in \mathbb{N}_0}$  with  $x_2[i_1, i_2] \in \mathbb{R}^{n_2}$ ,  $n_2 = c_k-1 r_2$  are given inductively by (5) with  $x_1[i_1, i_2] = 0$ ,  $x_2[i_1, i_2] = 0$  for  $[i_1, i_2] \in (\{0\} \times \mathbb{N}_0) \cup (\mathbb{N}_0 \times \{0\})$ .

*Proof.* See [22] for a proof.  $\square$

To represent strided convolutions in state space, we require a reshaping operator as a strided convolution is only shift invariant with respect to a shift by the stride  $s_k$  along  $i$ . This reshaping operator  $\mathcal{R}_{s_k}$  is given by

$$\ell_{2e}^{c_k-1}(\mathbb{N}_0^{d_k-1}) \rightarrow \ell_{2e}^{c_k-1}([1, s_k]) (\mathbb{N}_0^{d_k-1}),$$

$$(u_k[\mathbf{i}]) \mapsto (\text{vec}(u_k[s_k \mathbf{i} + \mathbf{t}] \mid \mathbf{t} \in [0, s_k])),$$

where  $\text{vec}(u_k[s_k \mathbf{i} + \mathbf{t}] \mid \mathbf{t} \in [0, s_k])$  denotes the stacked vector of the signal entries  $u_k[s_k \mathbf{i} + \mathbf{t}]$ ,  $\mathbf{t} \in [0, s_k]$ . The resulting state space representation for a strided convolution then takes this stacked vector  $\text{vec}(u_k[s_k \mathbf{i} + \mathbf{t}] \mid \mathbf{t} \in [0, s_k])$  as its input. Details on the construction of the Roesser model for strided convolutions and multiple examples can be found in [22].

**Remark 2.** Finding a mapping from  $K$  to  $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$  for  $N$ -D convolutions and dilated convolutions is also possible, see [22].

**Remark 3.** Representing a convolution in state space requires the choice of a propagation direction for both dimensions. Usually, for image inputs we pick the upper left corner as the origin with  $i_1 = i_2 = 0$ . However, any other corner and corresponding propagation directions can also be chosen to represent the convolution equivalently. For state space model layers the propagation dimension, i.e., time is predefined, and cannot be changed.

### III. LIPSCHITZ CONSTANT ESTIMATION

To address Problem 1 of estimating the Lipschitz constant of an NN, we make use of our interpretation (2) of NNs as dynamical systems  $u_{k+1} = \ell_k(u_k)$ . Namely, we can pose the problem of estimating the Lipschitz constant of the neural network  $\text{NN}_\theta$  as the dynamic optimization problem

$$\begin{aligned} \min_{\gamma \in \mathbb{R}} \quad & \gamma \\ \text{s.t.} \quad & \|y_l^1 - y_l^2\|_2 \leq \gamma \|u_1^1 - u_1^2\|_2, \quad \forall u_1^1, u_1^2 \in \mathcal{D}_0, \\ & y_k^1 = \ell_k(u_k^1), \quad y_k^2 = \ell_k(u_k^2), \quad k = 1, \dots, l, \\ & u_{k+1}^1 = y_k^1, \quad u_{k+1}^2 = y_k^2, \quad k = 1, \dots, l-1. \end{aligned} \quad (10)$$

The advantage of the recursive formulation (10) is that it can be solved using a dynamic programming approach. Namely, by making use of the principle of optimality, we can recursively define the incremental value functions

$$V_l(y_l^1, y_l^2) = \|y_l^1 - y_l^2\|_2, \quad y_l^1, y_l^2 \in \mathcal{D}_l$$

$$V_{k-1}(u_k^1, u_k^2) = V_k(\ell_k(u_k^1), \ell_k(u_k^2)), \quad u_k^1, u_k^2 \in \mathcal{D}_{k-1}, \quad (11)$$

for  $k = 1, \dots, l$ , starting from the  $l$ -th layer, and obtain that (10) is equivalent to finding the smallest  $\gamma \in \mathbb{R}_+$  such that  $V_0(u_1^1, u_1^2) \leq \gamma^2 \|u_1^1 - u_1^2\|_2^2$  for all  $u_1^1, u_1^2 \in \mathcal{D}_0$ . We can therefore pose (10) as the optimization problem

$$\min_{\gamma, V_1, \dots, V_l} \quad \gamma^2 \quad (12a)$$

$$\text{s.t.} \quad V_l(y_l^1, y_l^2) \geq \|y_l^1 - y_l^2\|_2^2 \quad (12b)$$

$$V_{k-1}(u_k^1, u_k^2) \geq V_k(\ell_k(u_k^1), \ell_k(u_k^2)), \quad k = l, \dots, 2 \quad (12c)$$

$$\gamma^2 \|u_1^1 - u_1^2\|_2^2 \geq V_1(\ell_1(u_1^1), \ell_1(u_1^2)) \quad (12d)$$

over the incremental value functions, where (12b) to (12d) must hold for all  $u_1^1, u_1^2 \in \mathcal{D}_0$  and  $u_k^1, u_k^2, k = 1, \dots, l$  are obtained by (2). This problem is equivalent to finding the minimal  $\gamma \in \mathbb{R}_+$  such that  $\|\text{NN}_\theta(u_1^1) - \text{NN}_\theta(u_1^2)\|_2 \leq \gamma \|u_1^1 - u_1^2\|_2$  holds for all  $u_1^1, u_1^2 \in \mathcal{D}_0$ . It involves  $l$  constraints of the form (12c) and (12d). Note that this layer-wise splitting is computationally favorable over using one large and sparse constraint for the whole NN.

Still, at the present state, (12) is an intractable problem due to the optimization over the infinite-dimensional objects (functions)  $V_k$  and the infinitely many constraints (12c) which must hold for all  $u_k^1, u_k^2 \in \mathcal{D}_{k-1}$ . For this reason, we refer to a very common relaxation from the control literature, namely, quadratic incremental value functions. To this end, we constrain the functions  $V_k$  to be of the form

$$V_k(y_k^1, y_k^2) = V_{X_k}(y_k^1, y_k^2) := \langle y_k^1 - y_k^2, X_k(y_k^1 - y_k^2) \rangle_2$$

for linear self-adjoint operators  $X_k$  on  $\mathcal{D}_k$ . In the case  $\mathcal{D}_k = \mathbb{R}^{c_k}$  we may simply assume that the operators  $X_k$  are in matrix representation and obtain

$$V_{X_k}(y_k^1, y_k^2) = (y_k^1 - y_k^2)^\top X_k (y_k^1 - y_k^2).$$

In the case  $\mathcal{D}_k = \ell_{2e}^{c_k}(\mathbb{N}_0^{d_k})$ , we can represent  $X_k$  in terms of a sequence of matrices  $(\tilde{X}_k[\mathbf{i}, \mathbf{j}])_{\mathbf{i}, \mathbf{j} \in \mathbb{N}_0^{d_k}}, \tilde{X}_k[\mathbf{i}, \mathbf{j}] \in \mathbb{R}^{c_k \times c_k}$  by

$$V_{X_k}(y_k^1, y_k^2) = \sum_{\mathbf{i}, \mathbf{j} \in \mathbb{N}_0^{d_k}} (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}])^\top \tilde{X}_k[\mathbf{i}, \mathbf{j}] (y_k^1[\mathbf{j}] - y_k^2[\mathbf{j}]).$$

W.l.o.g.  $\tilde{X}_k$  can be assumed to be symmetric, i.e.,  $\tilde{X}_k[\mathbf{i}, \mathbf{j}] = \tilde{X}_k[\mathbf{j}, \mathbf{i}]^\top = \tilde{X}_k[\mathbf{j}, \mathbf{i}]$ . This relaxation is a first step towards rendering the optimization tractable. Particularly, in the following, we derive sufficient LMI conditions for (12c) for every layer  $\ell_k \in \{\mathcal{L}, \mathcal{C}, \mathcal{S}, \sigma, \mathcal{P}, \mathcal{F}\}$ . That means, we formulate LMIs which imply

$$V_{X_{k-1}}(u_k^1, u_k^2) \geq V_{X_k}(\ell_k(u_k^1), \ell_k(u_k^2)), \quad (13)$$

under the assumption  $V_k = V_{X_k}$ . The latter is a quadratic relaxation of (12c) at the  $k$ -th layer. For some layers, we require further restrictions on  $X_k$  to state tractable LMIs, as we will discuss for all individual layer types in the following.

#### A. The convolutional layer

If  $\ell_k = \mathcal{C}_k$  is a convolutional layer, then  $\mathcal{D}_{k-1} = \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_0})$  and  $\mathcal{D}_k = \ell_{2e}^{c_k}(\mathbb{N}_0^{d_k})$ . Convolutional layers described by (3) are shift-invariant mappings and a similar property will be required from the operators  $X_k$ . Particularly, we require that  $X_{k-1}$  and  $X_k$  are of the form

$$\tilde{X}_{k-1}[\mathbf{i}, \mathbf{j}] = \begin{cases} \tilde{X}_{k-1} & \mathbf{i} = \mathbf{j} \\ 0 & \mathbf{i} \neq \mathbf{j}, \end{cases} \quad \tilde{X}_k[\mathbf{i}, \mathbf{j}] = \begin{cases} \tilde{X}_k & \mathbf{i} = \mathbf{j} \\ 0 & \mathbf{i} \neq \mathbf{j}, \end{cases} \quad (14)$$

i.e., these operators are parametrized by matrices  $\tilde{X}_k \in \mathbb{S}^{c_k}$  and  $\tilde{X}_{k-1} \in \mathbb{S}^{c_{k-1}}$  in a *block-diagonally repeated* fashion. This restriction might seem confining at first sight, but it renders the problem computationally tractable and leverages the structure of convolutional layers, i.e., the shift invariance,

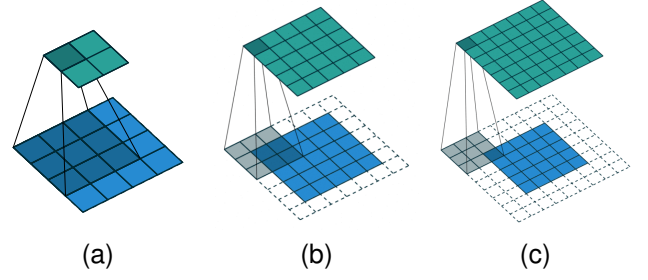


Fig. 1. (a) No padding, (b) same padding, (c) full padding for a  $3 \times 3$  kernel [27].

such that (12c) can be relaxed as an LMI as follows. We denote the convolution-specific restriction in (14) by  $X_k \in \mathcal{H}_C^y$  and  $X_{k-1} \in \mathcal{H}_C^u$ .

**Lemma 2.** Consider the  $k$ -th layer to be a convolutional layer  $\ell_k = \mathcal{C}_k$ . For some operators  $X_k \in \mathcal{H}_C^y$  and  $X_{k-1} \in \mathcal{H}_C^u$ , the convolutional layer (3) represented by a Roesser model (8) satisfies (13) if there exist symmetric matrices  $P_m^k \in \mathbb{S}_+^{n_m}$ ,  $\mathbf{P}_k = \text{blkdiag}(P_1^k, \dots, P_d^k)$  such that

$$\begin{bmatrix} \mathbf{P}_k & 0 \\ 0 & \tilde{X}_{k-1} \end{bmatrix} - \begin{bmatrix} \mathbf{A}_k & \mathbf{B}_k \\ \mathbf{C}_k & \mathbf{D}_k \end{bmatrix}^\top \begin{bmatrix} \mathbf{P}_k & 0 \\ 0 & \tilde{X}_k \end{bmatrix} \begin{bmatrix} \mathbf{A}_k & \mathbf{B}_k \\ \mathbf{C}_k & \mathbf{D}_k \end{bmatrix} \succeq 0. \quad (15)$$

*Proof.* A proof of Lemma 2 is given in [17, Theorem 4] for 2-D systems and in Appendix A-A for N-D systems.  $\square$

We denote the inequality (15) by  $\mathcal{G}_k(X_{k-1}, X_k, \nu_k) \succeq 0$ , where  $\nu_k = \mathbf{P}_k$  contains the slack variables in (15). We further note that Lemma 2 is lossless in the cases  $d = 0, 1$ , i.e., (13) holds for the  $k$ -th layer if and only if (15) is satisfied. For  $d \geq 2$ , this is no longer the case and the conservatism of (15) might depend on the choice of the realization [17] and the blockdiagonal structure of  $\mathbf{P}_k$ .

Note the special structure of (15). If we understand  $\tilde{X}_k/\tilde{X}_{k-1}$  as another block of  $\mathbf{P}_k$ , then this matrix inequality is a Lyapunov inequality for a  $(d+1)$ -D system, where the  $(\mathbf{A}_k, \mathbf{B}_k, \mathbf{C}_k, \mathbf{D}_k)$  block plays the role of the  $A$ -matrix. This system is time-varying along the  $k$ -axis, which should be viewed as the time-axis, and time-invariant along all other axes, which should be viewed as space-axes.

What is more, Lemma 2 can be applied to strided convolutions, but it is not straightforward to do so. In the case of a strided convolution we pick an intermediate metric defined through an operator  $Y_{k-1}$  such that

$$\begin{aligned} & \langle u_k^1 - u_k^2, X_{k-1}(u_k^1 - u_k^2) \rangle_2 \\ &= \langle \mathcal{R}_{s_k}(u_k^1 - u_k^2), Y_{k-1} \mathcal{R}_{s_k}(u_k^1 - u_k^2) \rangle_2 \end{aligned}$$

holds, where we restrict  $Y_{k-1}$  in such a way that

$$\tilde{Y}_{k-1}[\mathbf{i}, \mathbf{j}] = \begin{cases} \tilde{Y}_{k-1} & \mathbf{i} = \mathbf{j} \\ 0 & \mathbf{i} \neq \mathbf{j}, \end{cases}$$

always holds for some matrix  $\tilde{Y}_{k-1} \in \mathbb{S}^{c_{k-1} \lfloor 0, s_k \rfloor}$ . With that, we can guarantee that inequality (13) is satisfied for the layer  $\ell_k = \mathcal{C}_{k, s_k}$  if (15) holds with  $\tilde{X}_{k-1}$  replaced by  $\tilde{Y}_{k-1}$ .

Another design choice for convolutional layers is the kind of *zero-padding* that is used. There are different kinds of padding as shown in Fig. III-A [27]. We distinguish between full padding which increases the output dimension, same padding which preserves it and no/valid padding which decreases it. The proof of Lemma 2 in Appendix A-A relies on full padding, which over-approximates the other cases as we argue in the following. The type of padding decides which finite excerpt  $[N_1, N_2]$  of the infinite signal on  $\mathbb{N}_0^d$  is passed on to the next layer.

Let  $[N_1, N_2]$  define the excerpt that is used with same or no zero-padding. In case of full padding, the chosen excerpt involves all non-zero entries of  $y_k^1[\cdot]$  and  $y_k^2[\cdot]$  such that its value function is  $V_{X_k}(y_k^1, y_k^2)$ . Due to the quadratic nature of  $V_{X_k}$ , its evaluation on a finite excerpt of the same signal yields the first inequality in

$$\sum_{i=N_1}^{N_2} y_k[i]^\top \tilde{X}_k y_k[i] \leq V_{X_k}(y_k^1, y_k^2) \leq V_{X_{k-1}}(u_k^1, u_k^2),$$

and the second inequality, i.e., (13), is implied by (15). This shows that (15) implies (12c) for unpadded and same-padded signals.

#### B. The state space model layer

The state space model layer  $\ell_k = \mathcal{S}_k$  is a generalization of the 1-D convolutional layer, that has recently gained popularity in the machine learning community [18]. The proof of Lemma 2 is independent of the structure of  $A_k$ ,  $B_k$ ,  $C_k$ , and  $D_k$ , which mark the difference between 1-D convolutional and state space model layers. Accordingly, we use (15) as a constraint  $\mathcal{G}_k(X_{k-1}, X_k, \nu_k) \succeq 0$ .

#### C. The fully connected layer

If  $\ell_k = \mathcal{L}_k$  is a fully connected layer, then  $\mathcal{D}_{k-1} = \mathbb{R}^{c_{k-1}} \cong \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^0)$  and  $\mathcal{D}_k = \mathbb{R}^{c_k} \cong \ell_{2e}^{c_k}(\mathbb{N}_0^0)$ . In this case, we can understand  $X_{k-1}$  and  $X_k$  as matrices with  $V_{X_{k-1}}(u_k^1, u_k^2) = (u_k^1 - u_k^2)^\top X_{k-1} (u_k^1 - u_k^2)$  and  $V_{X_k}(y_k^1, y_k^2) = (y_k^1 - y_k^2)^\top X_k (y_k^1 - y_k^2)$ , as mentioned before. We do not impose any further restrictions on  $X_k$ ,  $X_{k-1}$ , i.e.,  $\mathcal{H}_{\mathcal{L}}^y = \mathbb{R}^{c_k \times c_k}$  and  $\mathcal{H}_{\mathcal{L}}^u = \mathbb{R}^{c_{k-1} \times c_{k-1}}$ . The following lemma describes (12c) as an LMI in a lossless manner.

**Lemma 3.** *Consider the  $k$ -th layer to be a fully connected layer  $\ell_k = \mathcal{L}_k$ . With operators  $X_k \in \mathcal{H}_{\mathcal{L}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{L}}^u$ , a fully connected layer (7) satisfies (12c) if and only if*

$$X_{k-1} - W_k^\top X_k W_k \succeq 0. \quad (16)$$

*Proof.* The proof follows trivially by right/left multiplication with  $u_k^1 - u_k^2$  and its transpose, respectively, for any  $u_k^1, u_k^2 \in \mathbb{R}^{c_{k-1}}$ .  $\square$

We denote the inequality (16) by  $\mathcal{G}_k(X_{k-1}, X_k, \nu_k) \succeq 0$ , where  $\nu_k = []$  (the empty matrix). We mention that (16) is a special case of the Lyapunov equation (15) for  $d = 0$ , cmp. Remark 1. In this case,  $A_k, B_k, C_k, P_k$  are empty matrices and  $D_k$  corresponds to  $W_k$ . This observation corresponds to Remark 1.

#### D. The activation function layer

If  $\ell_k = \sigma_k$  is an activation function layer, then  $\mathcal{D}_k = \mathcal{D}_{k-1} = \mathbb{R}^{c_k}$  and  $\mathcal{D}_k = \mathcal{D}_{k-1} = \ell_{2e}^{c_k}(\mathbb{N}_0^{d_k})$  are both possible (recall  $c_{k-1} = c_k$  and  $d_{k-1} = d_k$  in this case). In case  $d_k > 0$ , we choose the operators  $X_k$  and  $X_{k-1}$  to be block-diagonal and time-invariant, i.e., they satisfy (14). The restriction of time-invariance is not needed for this layer type, i.e., block-diagonal multipliers with varying blocks  $\tilde{X}_k[\mathbf{z}]$  can also be used. However, we use the restriction (14) for simplicity and computational tractability reasons.

The most common activation functions such as ReLU, tanh, and sigmoid are slope-restricted, i.e., they satisfy the quadratic constraint (17) of the following lemma.

**Lemma 4** (Slope-restriction [28], [29]). *Consider an activation function  $\sigma : \mathbb{R}^c \rightarrow \mathbb{R}^c$  that is slope-restricted on  $[0, 1]$ . For any  $\Lambda \in \mathbb{D}_+^c$ ,  $\sigma$  satisfies*

$$\begin{bmatrix} x - y \\ \sigma(x) - \sigma(y) \end{bmatrix}^\top \begin{bmatrix} 0 & \Lambda \\ \Lambda & -2\Lambda \end{bmatrix} \begin{bmatrix} x - y \\ \sigma(x) - \sigma(y) \end{bmatrix} \geq 0, \quad \forall x, y \in \mathbb{R}^c. \quad (17)$$

Note that the published version of [28], i.e., [10], falsely used full matrix multipliers instead of diagonal  $\Lambda_k$  which was later corrected by [29]. For slope-restricted activation functions, (12c) can be relaxed by an LMI as follows.

**Lemma 5.** *Consider the  $k$ -th layer to be an activation function layer  $\ell_k = \sigma_k$  that is slope-restricted on  $[0, 1]$ . For some operators  $X_k \in \mathcal{H}_{\mathcal{L}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{L}}^u$ , this activation function layer satisfies (13) if there exist  $\Lambda_k \in \mathbb{D}_+^{c_k}$  such that*

$$\begin{bmatrix} \tilde{X}_{k-1} & -\Lambda_k \\ -\Lambda_k & 2\Lambda_k - \tilde{X}_k \end{bmatrix} \succeq 0. \quad (18)$$

*Proof.* For two arbitrary inputs  $(u_k^1[\mathbf{z}], u_k^2[\mathbf{z}]) \in \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}})$  with corresponding outputs  $(y_k^1[\mathbf{z}], y_k^2[\mathbf{z}])$ , we left and right multiply (18) with  $[(y_k^1[\mathbf{z}] - y_k^2[\mathbf{z}])^\top \quad (u_k^1[\mathbf{z}] - u_k^2[\mathbf{z}])^\top]$  and its transpose, respectively, which yields

$$\begin{aligned} & (u_k^1[\mathbf{z}] - u_k^2[\mathbf{z}])^\top \tilde{X}_{k-1} (u_k^1[\mathbf{z}] - u_k^2[\mathbf{z}]) \\ & - (y_k^1[\mathbf{z}] - y_k^2[\mathbf{z}])^\top \tilde{X}_k (y_k^1[\mathbf{z}] - y_k^2[\mathbf{z}]) \\ & \geq [\bullet]^\top \begin{bmatrix} 0 & \Lambda_k \\ \Lambda_k & -2\Lambda_k \end{bmatrix} \begin{bmatrix} u_k^1[\mathbf{z}] - u_k^2[\mathbf{z}] \\ y_k^1[\mathbf{z}] - y_k^2[\mathbf{z}] \end{bmatrix}, \end{aligned}$$

where  $[\bullet]^\top$  is inferred by symmetry. Subsequently, we sum over  $\mathbf{z} \in \mathbb{N}_0^d$  and obtain

$$\begin{aligned} & V_{X_{k-1}}(u_k^1, u_k^2) - V_{X_k}(y_k^1, y_k^2) \\ & \geq \sum_{\mathbf{z} \in \mathbb{N}_0^d} [\bullet]^\top \begin{bmatrix} 0 & \Lambda_k \\ \Lambda_k & -2\Lambda_k \end{bmatrix} \begin{bmatrix} u_k^1[\mathbf{z}] - u_k^2[\mathbf{z}] \\ y_k^1[\mathbf{z}] - y_k^2[\mathbf{z}] \end{bmatrix} \geq 0, \end{aligned}$$

wherein the last inequality follows from Lemma 4.  $\square$

Note that Lemma 5 also includes activation functions applied subsequent to fully connected layers, where technically we need to infer  $X_k \in \mathcal{H}_{\mathcal{L}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{L}}^u$  instead of  $\mathcal{H}_{\mathcal{L}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{L}}^u$ . In fact,  $X_k \in \mathcal{H}_{\mathcal{L}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{L}}^u$  are special cases of  $X_k \in \mathcal{H}_{\mathcal{L}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{L}}^u$  for  $d = 0$ , cmp.

Remark 1. Consequently, we denote the constraint (18) by  $\mathcal{G}_k(X_{k-1}, X_k, \nu_k) \succeq 0$  where  $\nu_k = \Lambda_k$ .

Beside slope-restricted activations, another class of activation functions that has recently gained popularity are gradient norm preserving activations such as GroupSort and MaxMin [20]. These activations are not applied element-wise but to a vector input  $u[i] \in \mathbb{R}^c$  consisting of all preactivations at  $i$ . GroupSort separates the  $c$  preactivations into  $N$  groups each of size  $n_g$ , i.e.,  $c = Nn_g$ , and then sorts these groups in ascending order. With the restriction

$$\tilde{X}_k[i, j] = \begin{cases} \tilde{X}_k \in \mathcal{T}_{n_g}^c & i = j \\ 0 & i \neq j \end{cases}$$

and an equivalent definition for  $\tilde{X}_{k-1}[i, j]$ , where

$$\mathcal{T}_{n_g}^c = \{T \in \mathbb{S}^c \mid T = \text{diag}(\lambda) \otimes I_{n_g} + \text{diag}(\gamma) \otimes \mathbf{1}_{n_g} \mathbf{1}_{n_g}^\top, \\ \lambda \in \mathbb{R}_+^{c/n_g}, \gamma \in \mathbb{R}^{c/n_g}\},$$

we can handle GroupSort activation functions using the following lemma [30]. We denote these structural constraints by  $X_k \in \mathcal{H}_{\sigma_{GS}}^y$  and  $X_{k-1} \in \mathcal{H}_{\sigma_{GS}}^u$ .

**Lemma 6.** Consider the  $k$ -th layer to be a GroupSort activation function  $\ell_k = \sigma_k^{\text{GS}}$ . For some operators  $X_k \in \mathcal{H}_{\sigma_{GS}}^y$  and  $X_{k-1} \in \mathcal{H}_{\sigma_{GS}}^u$ , the GroupSort activation function satisfies the metric bound (13) if the matrices  $\tilde{X}_k$  and  $\tilde{X}_{k-1}$  satisfy  $0 \preceq \tilde{X}_k \preceq \tilde{X}_{k-1}$ .

*Proof.* The proof is deferred to Appendix A-B.  $\square$

### E. The pooling layer

If  $\ell_k = \mathcal{P}_k$  is a pooling layer, then  $\mathcal{D}_{k-1} = \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^d)$  and  $\mathcal{D}_k = \ell_{2e}^{c_k}(\mathbb{N}_0^d)$  for  $d = d_k = d_{k-1}$  and  $c_k = c_{k-1}$ . The handling of pooling layers is very similar to the handling of activation function layers. As discussed in [16], for both layer types there exist quadratic constraints, based on which we find LMI constraints for the respective layers.

Since pooling layers, i.e., subsampling layers, only make sense on the signal spaces  $\ell_{2e}^{c_k}(\mathbb{N}_0^d)$ , we consider these signal spaces as domain and image spaces and restriction (14) on the operators  $X_k$  and  $X_{k-1}$ . Accordingly, we consider again static shift invariant value functions

$$V_{X_{k-1}}(u_k^1, u_k^2) = \sum_{i \in \mathbb{N}_0^d} (u_k^1[i] - u_k^2[i])^\top \tilde{X}_{k-1} (u_k^1[i] - u_k^2[i]), \\ V_{X_k}(y_k^1, y_k^2) = \sum_{i \in \mathbb{N}_0^d} (y_k^1[i] - y_k^2[i])^\top \tilde{X}_k (y_k^1[i] - y_k^2[i]).$$

Note that theoretically, we could study this problem in the non-static, non-shift-invariant case. However, pooling layers will be concatenated with convolutional layers, which is why the operators  $X_k$  and  $X_{k-1}$  must be shift-invariant and static in the end. The following lemma shows how we handle the metric bound (13) with average pooling layers.

**Lemma 7.** For the average pooling layer, the metric bound (13) is satisfied if the matrices  $\tilde{X}_k$  and  $\tilde{X}_{k-1}$  satisfy the simple matrix inequality  $0 \preceq \mu_k^2 \tilde{X}_k \preceq \tilde{X}_{k-1}$ ,  $\mu_k$  being the Lipschitz constant of the average pooling layer.

Normally, we will set  $\tilde{X}_{k-1} = \mu_k^2 \tilde{X}_k$ . For maximum pooling layers, we require an additional restriction, namely  $\tilde{X}_k = \text{diag}(\lambda_k)$ ,  $\lambda_k \in \mathbb{R}^{c_k}$ ,  $\tilde{X}_{k-1} = \text{diag}(\lambda_{k-1})$ ,  $\lambda_{k-1} \in \mathbb{R}^{c_{k-1}}$ , yielding the next lemma.

**Lemma 8.** For the maximum pooling layer, the metric bound (12c) is satisfied if the matrices  $\tilde{X}_k$  and  $\tilde{X}_{k-1}$ , that are parametrized as  $\tilde{X}_k = \text{diag}(\lambda_k)$ ,  $\tilde{X}_{k-1} = \text{diag}(\lambda_{k-1})$ , satisfy  $0 \leq \mu_k^2 \lambda_k^i \leq \lambda_{k-1}^i$  for  $i = 1, \dots, c_k$  with Lipschitz constant  $\mu_k$  of the maximum pooling layer.

Again, we will normally require  $\mu_k^2 \lambda_k^i = \lambda_{k-1}^i$ . It is common to choose the kernel size  $r_k$  and the stride  $s_k$  to be the same. In that case the Lipschitz constant of a maximum pooling layer is 1. We denote the restriction of the operators for maximum pooling layers, including the diagonality constraints, by  $X_k \in \mathcal{H}_{\mathcal{P}^{\max}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{P}^{\max}}^u$ . Furthermore, we denote by  $\mathcal{G}_k(X_{k-1}, X_k, \nu_k) \succeq 0$  the respective constraint for these layers with  $\nu_k = [\ ]$ .

### F. Flattening operations

In our setup, flattening operations have the role of reshaping tensor outputs from  $\ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}})$  as vectors. In particular, they rearrange the output of a convolutional layer with  $d_{k-1} > 0$  as a vector before it can serve as an input for a fully connected layer. We have mentioned that, theoretically, a flattening operation could also map/project an element from  $\ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}})$  to  $\ell_{2e}^{c_k}(\mathbb{N}_0^{d_k})$ , where  $d_{k-1} > d_k$ . However, we consider only the most relevant case of  $d_k = 0$  in this section.

In this case,  $\mathcal{F}_k$  maps a patch  $(u_k[i] \mid N_{k1} \leq i \leq N_{k2})$  to the stacked vector of  $u_k[i]$ , i.e.,  $y_k = [u_k[N_{k1}]^\top \ \dots \ u_k[N_{k2}]^\top]^\top$ . Thus, we obtain the following lemma.

**Lemma 9.** Consider a flattening operation  $\mathcal{F}_{k+1} : \ell_{2e}^{c_k}(\mathbb{N}_0^{d_k}) \rightarrow \mathbb{R}^{c_{k+1}}$ , with support  $[N_{k1}, N_{k2}]$  and  $c_{k+1} = c_k | [N_{k1}, N_{k2}] |$ . The incremental value function  $V_{X_k}$  can be denoted as

$$\sum_{i, j \in \mathbb{N}_0^{d_k}} (u^1[i] - u^2[i])^\top \tilde{X}_k[i, j] (u^1[j] - u^2[j]).$$

Then the dynamic programming inequality (13) is satisfied if and only if

$$\begin{bmatrix} \tilde{X}_k[N_{k1}, N_{k1}] & \dots & \tilde{X}_k[N_{k1}, N_{k2}] \\ \vdots & \ddots & \vdots \\ \tilde{X}_k[N_{k2}, N_{k1}] & \dots & \tilde{X}_k[N_{k2}, N_{k2}] \end{bmatrix} \succeq X_{k+1}. \quad (19)$$

The matrix inequality (19) is an instance of  $\mathcal{G}_k(X_k, X_{k+1}, \nu_k) \succeq 0$  with  $\nu = [\ ]$  and  $\mathcal{H}_{\mathcal{F}}^y$  and  $\mathcal{H}_{\mathcal{F}}^u$  technically impose no additional restrictions on  $X_k$  and  $X_{k+1}$  of the flattening operation. However, usually, the value function  $V_{X_k}$  will be both static and time-invariant due to output restrictions on  $X_k$  of the previous layer, e.g.,  $X_k \in \mathcal{H}_{\mathcal{L}}^y$ , i.e.,  $\tilde{X}_k[i, j] = \tilde{X}_k$  for  $i = j$  and zero otherwise. In addition, we can require equality in (19), in which case  $X_{k+1} = I_{|[N_{k1}, N_{k2}]|} \otimes \tilde{X}_k$  is a block diagonal matrix with  $|[N_{k1}, N_{k2}]|$  copies of  $\tilde{X}_k$  on its diagonal.

### G. Subnetworks

Up to now, we considered all building blocks of (1) as individual entities and require individual constraints (13) for all these layers. However, for the implementation of (12) and computational reasons, it is convenient to combine multiple layers as a subnetwork. We then include a constraint of type (13) for the subnetwork. A typical concatenation is the combination of linear layers with the succeeding nonlinear activation functions, i. e.,  $\sigma \circ \mathcal{C}$  for convolutional layers or  $\sigma \circ \mathcal{L}$  for fully connected layers.

**Lemma 10.** *Consider the  $k$ -th layer to be the concatenation of a convolutional layer and an activation function layer, that is slope-restricted on  $[0, 1]$ ,  $\ell_k = (\sigma \circ \mathcal{C})_k$ . For some  $X_k \in \mathcal{H}_{\mathcal{C}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{C}}^u$ , the concatenation  $(\sigma \circ \mathcal{C})_k$  satisfies (13) if there exist symmetric matrices  $P_m \in \mathbb{S}_{+}^{n_m}$ ,  $P = \text{blkdiag}(P_1, \dots, P_d)$  and a diagonal matrix  $\Lambda_k \in \mathbb{D}_{+}^{c_k}$  such that*

$$\begin{bmatrix} P_k - A_k^\top P_k A_k & -A_k^\top P_k B_k & -C_k^\top \Lambda_k \\ -B_k^\top P_k A_k & \tilde{X}_{k-1} - B_k^\top P_k B_k & -D_k^\top \Lambda_k \\ -\Lambda_k C_k & -\Lambda_k D_k & 2\Lambda_k - \tilde{X}_k \end{bmatrix} \succeq 0. \quad (20)$$

*Proof.* The proof follows along the lines of the proof of Lemma 15, additionally using typical arguments from robust control [31]. It can be found in Appendix A-C.  $\square$

The condition (20) is treated as an instance of  $\mathcal{G}(X_k, X_{k-1}, \nu_k)$  with  $\nu_k = (P_k, \Lambda_k)$ . For an additional pooling layer, i. e.  $\mathcal{P} \circ \sigma \circ \mathcal{C}$ , we can extend Lemma 10 easily by replacing  $\tilde{X}_k$  with  $\mu_k^2 \tilde{X}_k$  and considering the output restriction  $X_k \in \mathcal{H}_{\mathcal{P}_{\max}}^y$  in case a maximum pooling layer is added.

**Lemma 11.** *Consider the  $k$ -th layer to be the concatenation of a fully connected layer and an activation function layer, that is slope-restricted on  $[0, 1]$ ,  $\ell_k = (\sigma \circ \mathcal{L})_k$ . For some  $X_k \in \mathcal{H}_{\mathcal{L}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{L}}^u$ , the concatenation  $(\sigma \circ \mathcal{L})_k$  satisfies (13) if there exists a diagonal matrix  $\Lambda_k \in \mathbb{D}_{+}^{n_{y_k}}$  such that*

$$\begin{bmatrix} X_{k-1} & -W_k^\top \Lambda_k \\ -\Lambda_k W_k & 2\Lambda_k - X_k \end{bmatrix} \succeq 0. \quad (21)$$

*Proof.* We can view condition (21) as a special case of (20), cmp. Remark 1, and therefore, we refer to the proof of Lemma 10 in Appendix A-C.  $\square$

Note that we can also combine more layers, yielding larger and sparser LMIs but renouncing the decision variables  $X_k$  at the transition between the layers. Extensions of Lemmas 10 and Lemma 11 of this kind can be found in Appendix B-A. If we combine all layers of a fully connected neural network, we obtain the LMI originally proposed in [10].

**Remark 4.** *Throughout this subsection, we consider slope-restricted activations. However, all LMIs can also be formulated for GroupSort activations based on Lemma 6 [30].*

### H. Residual layers and skip connections

In deep learning, neural network structures that include skip connections, called residual NNs or ResNets, have proven to avoid vanishing and exploding gradients [19]. We define such

residual layers as a combination of linear layers (convolutional or fully connected) and nonlinear activation functions

$$y_k = \sigma(u_k + \mathcal{M}(u_k)), \quad (22)$$

where  $\mathcal{M}(u_k)$  is a feedforward NN (1) of arbitrary length and  $u_k \in \mathcal{D}_{k-1}$  and  $y_k \in \mathcal{D}_k$ . We in addition require  $\mathcal{D}_k = \mathcal{D}_{k-1}$  as well as  $\mathcal{M} : \mathcal{D}_{k-1} \rightarrow \mathcal{D}_{k-1}$ . For example, a ResNet layer that skips a fully connected network with one hidden layer reads

$$y_k = \sigma(u_k + W_2 \sigma(W_1 u_k + b_k)). \quad (23)$$

with  $W_1 \in \mathbb{R}^{n_{v_k} \times c_{k-1}}$ ,  $W_2 \in \mathbb{R}^{c_k \times n_{v_k}}$ ,  $n_{v_k}$  being the dimension of  $v_k := \sigma(W_1 u_k + b_k)$ .

In the following lemma, we describe how the simple skip connection (23) leads to an LMI relaxation for (13). More general skip connections can be treated with the same arguments; see Appendix B-B.

**Lemma 12.** *Consider the  $k$ -th layer to be a residual layer (23) with activation functions that are slope-restricted in  $[0, 1]$ . For some  $X_k \in \mathcal{H}_{\mathcal{L}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{L}}^u$ , the ResNet layer (23) satisfies (13) if there exist  $\Lambda_1 \in \mathbb{D}_{+}^{n_{v_k}}$ ,  $\Lambda_2 \in \mathbb{D}_{+}^{n_{y_k}}$  such that*

$$\begin{bmatrix} X_{k-1} & -W_1^\top \Lambda_1 & -\Lambda_2 \\ -\Lambda_1 W_1 & 2\Lambda_1 & -W_2^\top \Lambda_2 \\ -\Lambda_2 & -\Lambda_2 W_2 & 2\Lambda_2 - X_k \end{bmatrix} \succeq 0. \quad (24)$$

*Proof.* A proof can be found in Appendix A-D  $\square$

### I. Semidefinite relaxation for Lipschitz constant estimation

In the previous sections, we discussed relaxations for the dynamic programming inequality (12c) for incremental value functions defined by self-adjoint operators  $X_k$ , i.e., for every single type of layer or subnetwork  $\ell \in \{\mathcal{L}, \mathcal{C}, \sigma, \mathcal{P}, \mathcal{R}, \mathcal{F}\} \cup \{\sigma \circ \mathcal{L}, \sigma \circ \mathcal{C}, \mathcal{P} \circ \sigma \circ \mathcal{C}\} \cup \{\text{residual layer}\}$ , we formulated a semi-definite constraint relaxing (12c). As a result, we can pose the optimization problem

$$\begin{aligned} \min_{X_0, \dots, X_l, \nu_1, \dots, \nu_l, \gamma^2} \quad & \gamma^2 \\ \text{subject to} \quad & X_0 = \gamma^2 I, \\ & \mathcal{G}_k(X_k, X_{k-1}, \nu_k) \succeq 0, \quad k = 1, \dots, l, \\ & X_k \in \mathcal{H}_{\ell_k}^y \cap \mathcal{H}_{\ell_{k+1}}^u, \quad k = 1, \dots, l-1, \\ & X_l = I. \end{aligned} \quad (25)$$

as a tractable relaxation of (12). The resulting optimization problem (25) is an SDP in the variables involved with at most one SDP constraint per layer (flattening and pooling layers do not produce SDP constraints). Here, the index  $k = 1, \dots, l$  counts through all layers/subnetworks that we consider.

Notice that each operator  $X_k$  has to satisfy the restrictions of two layers by  $X_k \in \mathcal{H}_{\ell_k}^y \cap \mathcal{H}_{\ell_{k+1}}^u$ . An example is the concatenation of a convolution and a strided convolution that invokes  $\tilde{Y}_k = I_{[1, s]} \otimes \tilde{X}_k$  or the flattening layer invoking  $X_k = I_{[N_{1k}, N_{2k}]} \otimes \tilde{X}_k$  onto  $X_k$ .



#### IV. ANALYSIS OF THE CONSERVATISM

With the exact dynamic programming recursion (11) it is (theoretically) possible to compute the exact Lipschitz constant of an NN using (12). For reasons of computational tractability, however, we propose the relaxation (25). For the derivation of this SDP, several relaxation steps were made resulting in the following sources of conservatism.

- 1) Quadratic incremental value functions  $V_{X_k}(y_k^1, y_k^2) = \langle y_k^1 - y_k^2, X_k(y_k^1 - y_k^2) \rangle_2$ .
- 2) Layer specific restrictions  $X_{k-1} \in \mathcal{H}_{\ell_k}^u$  and  $X_k \in \mathcal{H}_{\ell_k}^y$ .
- 3) Cut-off errors caused by handling convolutional layers as mappings on infinite-dimensional sequence spaces, whereas in reality only finitely supported image signals are processed.

Note that 3) can be viewed as a special case of 2), since considering general operators  $X_k$  instead of space-shift invariant operators would resolve this issue.

Our approach leverages the concatenation and individual layer structures of the NN, resulting in computational advantages and superior scalability compared to [10], [17]. Having discussed the sources of conservatism in our methodology, we will now justify why our proposed method for estimating the Lipschitz constant is *not* more restrictive than [10], [17]. To this end, we analyze a fully connected NN of the form

$$\text{FNN}_\theta = \mathcal{L}_l \circ \sigma_{l-1} \circ \dots \circ \sigma_2 \circ \mathcal{L}_1,$$

also considered in LipSDP [10]. As suggested in [16], we can use the semi-definite constraint

$$\begin{bmatrix} Q_l - I & S_l & & & \\ S_l^\top & R_l + Q_{l-1} & S_{l-1} & & \\ & S_{l-1}^\top & R_{l-1} + Q_{l-2} & \ddots & \\ & & \ddots & \ddots & S_1 \\ & & & S_1^\top & R_1 + \gamma^2 I \end{bmatrix} \preceq 0, \quad (26)$$

where

$$\begin{bmatrix} -2\Lambda_k & \Lambda_k W_k \\ W_k^\top \Lambda_k & 0 \end{bmatrix} \preceq \begin{bmatrix} Q_k & S_k \\ S_k^\top & R_k \end{bmatrix}, \quad k = 1, \dots, l \quad (27)$$

is used instead of (21) as layer-wise LMI constraints for fully connected layers for Lipschitz constant estimation. To recover LipSDP [10] simply replace the conic inequality in (27) with an equality.

The following theorem implies that it poses no restriction to parameterize the dissipativity blocks as  $\begin{bmatrix} Q_k & S_k \\ S_k^\top & R_k \end{bmatrix} = \begin{bmatrix} X_k & 0 \\ 0 & -X_{k-1} \end{bmatrix}$ , where  $X_l = I$  and  $X_0 = \gamma^2 I$ , as done in this work.

**Theorem 1.** Assume that the matrix inequality (26) is satisfied. Then there exists a sequence of matrices  $X_0, \dots, X_l$  such that  $X_0 = \gamma^2 I$ ,  $X_l = I$  and

$$\begin{bmatrix} Q_k & S_k \\ S_k^\top & R_k \end{bmatrix} \preceq \begin{bmatrix} X_k & 0 \\ 0 & -X_{k-1} \end{bmatrix}, \quad k = 1, \dots, l. \quad (28)$$

*Proof.* See Appendix A-E.  $\square$

It follows from Theorem 1 that the LMIs (26), (27) are equivalent to (21),  $k = 1, \dots, l$ ,  $X_0 = \gamma^2 I$ ,  $X_l = I$ ,  $X_k = I$ .

Consequently, the parameterization of  $Q_k, S_k, R_k$  via  $X_k$  does not introduce conservatism into the problem, i.e, the optimal value of  $\gamma$  found solving

$$\min_{\gamma^2, \Lambda, Q, S, R} \gamma^2 \quad \text{s. t.} \quad (26), (27), \quad (29)$$

where  $\Lambda = (\Lambda_1, \dots, \Lambda_l)$ ,  $Q = (Q_1, \dots, Q_l)$ ,  $S = (S_1, \dots, S_l)$  and  $R = (R_1, \dots, R_l)$ , and the optimal value of  $\gamma$  from

$$\min_{\gamma^2, \Lambda, X} \gamma^2 \quad \text{s. t.} \quad (21), \quad k = 1, \dots, l, \quad X_0 = \gamma^2 I, X_l = I, \Lambda_l = I \quad (30)$$

where  $\Lambda = (\Lambda_1, \dots, \Lambda_l)$ ,  $X = (X_1, \dots, X_{l-1})$ , are equivalent. We note that (30) is an instance of (25) for a fully connected neural network.

Another consequence of the result in Theorem 1 is that our approach of choosing the sequence of matrices  $(X_k)$  such that they satisfy (28) as our incremental value functions in (25) is not more conservative than LipSDP. The relation of (25) to [17], that includes convolutional layers, can be shown in a similar fashion.

**Remark 5.** For the special case of a fully connected NN our proposed layer-wise LMI constraints (21) correspond to the decomposition of the LMI in LipSDP by chordal sparsity [15], also yielding a set of LMI constraints that are equivalent to LipSDP [10].

**Remark 6.** The result of Theorem 1 can be interpreted as the statement that for a series interconnection of QSR-dissipative mappings, it suffices to consider supply rates  $s(u_k^1 - u_k^2, y_k^1 - y_k^2)$  of the form

$$\left\langle \begin{bmatrix} u_k^1 - u_k^2 \\ y_k^1 - y_k^2 \end{bmatrix}, \begin{bmatrix} X_k & 0 \\ 0 & -X_{k-1} \end{bmatrix} \begin{bmatrix} u_k^1 - u_k^2 \\ y_k^1 - y_k^2 \end{bmatrix} \right\rangle_2.$$

To summarize, the SDP (25) exploits the structure of NNs in two ways. Firstly, it exploits the concatenation structure of neural networks to generate  $l$  small SDP constraints instead of one large and sparse constraint, and, secondly, it utilizes the fact that convolutional layers and state space model layers are dynamical systems. This gives (25) one advantage over [17], where only the dynamical system nature of convolutional layers is exploited and two advantages over LipSDP [10] in terms of scalability.

**Remark 7.** Convolutional layers can be recast as fully connected layers and the experiments in [17] show that this recasting can reduce conservatism in comparison to (25) as it relaxes the layer specific restrictions in comparison to  $X_{k-1} \in \mathcal{H}_C^u$  and  $X_k \in \mathcal{H}_C^y$ . However, as it also becomes apparent in [17], this relaxation has a high computational cost.

#### V. SIMULATION RESULTS

In this section, we illustrate the computational advantages of our method and show its versatility by using it for Lipschitz constant estimation for multiple popular neural network architectures. We provide our code at <https://github.com/ppauli/GLipSDP>. This code is written in a modular fashion such that it can be applied easily to any neural network architecture

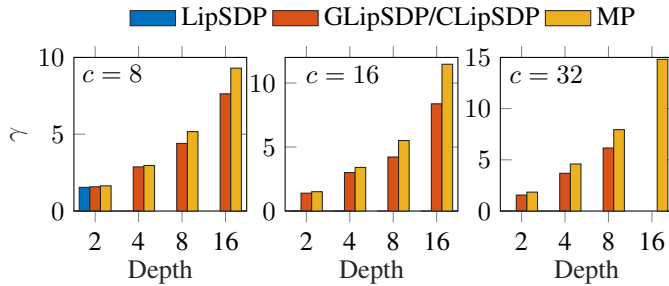


Fig. 2. Lipschitz bounds  $\gamma$  using LipSDP, GLipSDP/CLipSDP, and the matrix product bound (MP) on fully convolutional neural networks with depths  $d = \{2, 4, 8, 16\}$  and channel sizes  $c = \{8, 16, 32\}$ .

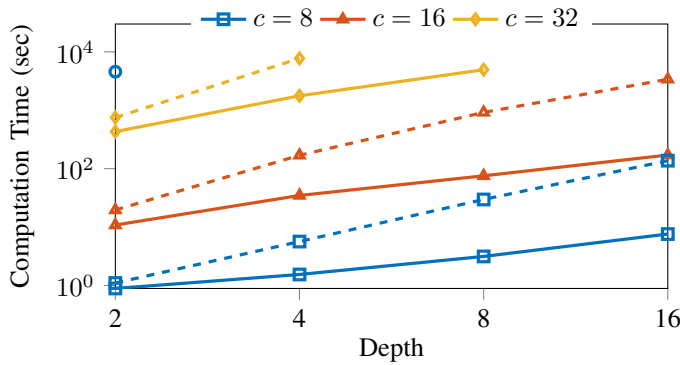


Fig. 3. Computation times for fully convolutional networks with depths  $d = \{2, 4, 8, 16\}$  and channel sizes  $c = \{8, 16, 32\}$  for GLipSDP (—) and CLipSDP (---).  $\bullet$  indicates the computation time using LipSDP. For larger networks, LipSDP runs into memory issues.

involving layers considered in this paper. All computations are carried out on a standard i7 note book using Yalmip [32] with the solver Mosek [33] in Matlab.

In the following subsections, we denote our method by GLipSDP (general LipSDP) based on SDP (25) and compare it to LipSDP [34] and CLipSDP (convolutional LipSDP) [17]. In addition, we compute a trivial matrix norm product bound (MP), the product of the spectral norms of the weights [4].

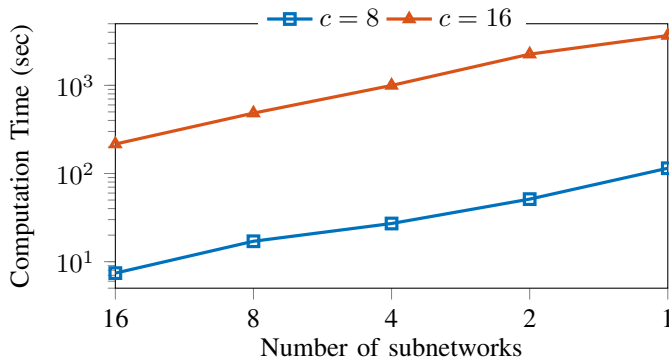


Fig. 4. Computation times for GLipSDP using 16, 8, 4, 2, 1 subnetworks for a 16-layer fully-convolutional network with 8 and 16 channels. The resulting Lipschitz bound is the same for all computations.

#### A. Scalability improvements on fully convolutional networks

First, we show the computational advantage of our structure exploiting approach in fully convolutional neural networks. To do so, we train CNNs with backbones of depths  $d = \{2, 4, 8, 16\}$  and channel sizes  $c = \{8, 16, 32\}$  on the MNIST dataset [35]. We then analyze the fully convolutional backbones of these neural networks, i.e., a subnetwork  $\sigma \circ \mathcal{C} \cdots \sigma \circ \mathcal{C}$  which only consists of convolutional layers. The input size to the backbone is  $14 \times 14$  and is kept constant throughout all backbone layers.

Fig. 2 shows the Lipschitz bounds obtained using LipSDP, GLipSDP (ours), CLipSDP, and MP. We first note that GLipSDP and CLipSDP produce the same bounds, both relying on a 2-D systems representation for the convolutional layers. As expected from the discussion on conservatism in Section IV, the bounds obtained using LipSDP are tighter than the ones obtained using GLipSDP/CLipSDP. However, LipSDP relies on a sparse and large Toeplitz matrix description of convolutional layers, cmp. Remark 7, and the underlying SDP runs into memory issues for all networks except the smallest one ( $c = 8, d = 2$ ). Further, we point out that both GLipSDP/CLipSDP and LipSDP are tighter than the trivial matrix product bound MP.

In Fig. 3 we compare the computation times of GLipSDP, CLipSDP, and LipSDP. For  $d = 2, c = 8$ , we can compare all three methods, noting that LipSDP takes more than 5000 times longer than GLipSDP, while CLipSDP is only slightly slower than GLipSDP. For all other networks, we compare GLipSDP and CLipSDP, noting that the deeper the network the larger the computational advantage of using GLipSDP. What distinguishes our method GLipSDP from CLipSDP is that we consider layer-wise LMIs rather than one large and sparse LMI constraint.

Next, we only consider the 16-layer fully-convolutional networks with channel sizes 8 and 16 ( $d = 16, c = \{8, 16\}$ ) and we apply GLipSDP but vary the number of layers combined in subnetworks, cmp. Section III-G. More specifically, we compute an upper Lipschitz bound using layer-wise LMI constraints, i.e., 16 LMI constraints, and then combine 2, 4, 8, 16 layers to form subnetworks, then applying GLipSDP with 8, 4, 2, and 1 LMI constraints instead of 16. GLipSDP yields the same Lipschitz bounds for all subnetwork configurations, yet it requires different computation times that are shown in Fig. 4. In this experiment, we clearly see that it is computationally advantageous to use multiple smaller LMI constraints, i.e., exploit the layer-by-layer structure of the network.

#### B. Convolutional neural networks for image classification

Next, we compute upper bounds on the Lipschitz constant for typical CNN architectures, including LeNet-5 [36], the NNs used in [37], and 18-layer residual neural networks, on the MNIST [35] and CIFAR-10 datasets. Details on the architectures are deferred to Appendix C. In Table I, we compare our method (GLipSDP) to the trivial matrix product bound (MP) and two variations of GLipSDP and GLipSDP:

- **S-LipSDP:** As LipSDP runs into memory issues for the chosen architectures, we apply LipSDP on possibly large

TABLE I  
LIPSCHITZ BOUNDS (COMPUTATION TIMES) FOR DIFFERENT MODELS WITH STATED ACCURACIES.

Dataset	Model	Acc.	GLipSDP	S-GLipSDP	S-LipSDP	MP
MNIST	LeNet-5	99.0%	<b>201.5</b> (126)	237.8 (117)	292.4 (17)	423.7
	2C2F	96.0%	<b>7.042</b> (909)	10.46 (329)	9.629 (7344)	16.06
	4C3F	99.2%	–	<b>5.564E+5</b> (2659)	1.319E+7 (638)	1.624E+7
	FC-R18	97.4%	<b>7.265E+3</b> (59)	5.904E+4 (4)	1.278E+5 (1)	4.435E+5
	C-R18	98.7%	3.616E+8 (98)	6.056E+8 (36)	<b>9.78E+7</b> (77471)	7.818E+8
CIFAR-10	LeNet-5	61.2%	3182.6 (152)	3500.6 (97)	<b>3134.8</b> (22)	4875.7
	6C2F	68.4%	–	<b>1.213E+7</b> (8729)	–	1.688E+7

subnetworks that are analyzed separately. The product of the Lipschitz estimates for the subnetworks yields an upper bound for the network.

- **S-GLipSDP:** We also compute S-GLipSDP (Split GLipSDP), splitting the neural network into subnetworks that are convenient to handle and apply GLipSDP to the subnetworks. Again, the product of the bounds of the subnetworks gives an upper bound for the entire NN.

All splits into subnetworks for S-GLipSDP and S-LipSDP are listed in Table III and IV, respectively, in Appendix C.

In Table I, we summarize the resulting Lipschitz bounds and computation times of the different neural networks. We observe that for LeNet-5 trained on MNIST, 2C2F, and FC-R18 GLipSDP produces the best bounds and shows reasonably short computation times. For 4C3F and 6C2F, we point out that GLipSDP runs into memory issues. However, here, S-GLipSDP generates the lowest Lipschitz bounds. C-R18 is an example demonstrating that the introduced conservatism in the handling of convolutional layers as 2-D systems can lead to larger bounds than using S-LipSDP. However, comparing the computation times, we recognize that GLipSDP is almost 800 times faster while also generating a much better bound than the matrix product bound. In LeNet-5 for CIFAR-10, S-LipSDP is slightly better than GLipSDP. We suspect that handling maximum pooling layers via quadratic constraints is not ideal and, as discussed before, using a 2-D systems representation for convolutions introduces conservatism.

## VI. CONCLUSION

We presented a versatile and scalable approach for Lipschitz constant estimation for a large class of neural network architectures. Our approach views the neural network as a time-varying dynamical system, where we interpret the layer indices as time indices. This view allows us to exploit the layer-wise composition structure of neural networks. In addition, we leverage the structure of the individual layers, especially of convolutional layers that we represent as N-D systems of the Roesser type. We wrote our code in a modular fashion such that it can easily be used for all neural networks that include the layer types considered in this paper. Future work includes the synthesis of Lipschitz bounded neural networks based on the LMIs presented in this paper.

## ACKNOWLEDGMENTS

The authors thank Andrea Iannelli for helpful comments on the manuscript.

## APPENDIX A ADDITIONAL PROOFS

### A. Proof of Lemma 2

We assume that the convolutional layer  $\ell_k = \mathcal{C}_k$  is realized as a Roesser system (8). This means that for any  $(u[\mathbf{i}]) \in \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}})$ , there exists a uniquely defined  $(x[\mathbf{i}]) \in \ell_{2e}^n(\mathbb{N}_0^{d_{k-1}})$  with  $x[\mathbf{i}] = [x_1[\mathbf{i}]^\top \cdots x_{d_{k-1}}[\mathbf{i}]^\top]^\top$  and

$$x_j[\mathbf{i}] = 0 \quad \forall \mathbf{i} \in \mathbb{N}_0^{d_{k-1}}, i_j = 0, \quad (31)$$

such that  $(u_k[\mathbf{i}], x[\mathbf{i}], y_k[\mathbf{i}])$  with  $y_k = \ell_k(u_k)$  satisfies (8), where  $i_j$  denotes the  $j$ -th index in  $\mathbf{i}$ .

Hence, let two arbitrary inputs  $(u_k^1[\mathbf{i}]), (u_k^2[\mathbf{i}]) \in \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}})$  be given and let  $(y_k^1[\mathbf{i}]), (y_k^2[\mathbf{i}]), (x^1[\mathbf{i}]), (x^2[\mathbf{i}])$  denote the corresponding state and output response of the layer  $\ell_k$ . We deliberately exclude a layer index for  $x[\mathbf{i}]$ , and use the subscript in  $x_j[\mathbf{i}]$  to count through  $j = 1, \dots, d$ . Multiplying the matrix inequality (15) from the left by the vector  $[(x^1[\mathbf{i}] - x^2[\mathbf{i}])^\top \ (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}])^\top]^\top$  and from the right by its transpose yields the inequality

$$\begin{aligned} & \sum_{j=1}^d (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}])^\top P_j (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}]) \\ & + (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}])^\top \tilde{X}_{k-1} (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}]) \\ & \leq \sum_{j=1}^d (x_j^1[\mathbf{i} + e_j] - x_j^2[\mathbf{i} + e_j])^\top P_j (x_j^1[\mathbf{i} + e_j] - x_j^2[\mathbf{i} + e_j]) \\ & + (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}])^\top \tilde{X}_k (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}]). \end{aligned}$$

Note that the bias terms do not need to be considered, since they cancel out when computing the differences  $x_j^1[\mathbf{i} + e_j] - x_j^2[\mathbf{i} + e_j]$ . Summing this inequality over all  $\mathbf{i} \in \mathbb{N}_0^{d_{k-1}}$  yields

$$\begin{aligned} & \sum_{\mathbf{i} \in \mathbb{N}_0^d} \sum_{j=1}^d (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}])^\top P_j (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}]) \\ & + \sum_{\mathbf{i} \in \mathbb{N}_0^d} (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}])^\top \tilde{X}_{k-1} (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}]) \\ & \leq \sum_{j=1}^d \sum_{\mathbf{i} \in \mathbb{N}_0^d, i_j \geq 1} (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}])^\top P_j (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}]) \\ & + \sum_{\mathbf{i} \in \mathbb{N}_0^d} (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}])^\top \tilde{X}_k (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}]). \end{aligned}$$

These sums all converge since all signals are in  $\ell_{2e}(\mathbb{N}_0^{d_{k-1}})$ , as the convolutional layer is a finite impulse response filter and, therefore, it is stable. Canceling terms on both sides yields

$$\begin{aligned} & \sum_{\mathbf{i} \in \mathbb{N}_0^d} (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}])^\top \tilde{X}_{k-1} (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}]) \\ & \leq \sum_{j=1}^d \sum_{\mathbf{i} \in \mathbb{N}_0^d, i_j=0} (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}])^\top P_j (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}]) \\ & + \sum_{\mathbf{i} \in \mathbb{N}_0^d} (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}])^\top \tilde{X}_k (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}]). \end{aligned}$$

Here, the sum over the boundary terms  $\sum_{j=1}^d \sum_{\mathbf{i} \in \mathbb{N}_0^d, i_j=0} (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}])^\top P_j (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}])$  is zero, cmp. (31), such that this inequality is exactly what we had to show.

### B. Proof of Lemma 6

To prove Lemma 6, we require the following lemma, which is a simplification of [30, Lemma 1] that directly follows for  $P = S = 0$  in [30, Lemma 1].

**Lemma 13.** Consider a GroupSort activation  $\sigma^{\text{GS}} : \mathbb{R}^c \rightarrow \mathbb{R}^c$  with group size  $n_g$ . For any  $T \in \mathcal{T}_{n_g}^c$ ,  $\sigma^{\text{GS}}$  satisfies

$$\begin{bmatrix} x - y \\ \sigma(x) - \sigma(y) \end{bmatrix}^\top \begin{bmatrix} T & 0 \\ 0 & -T \end{bmatrix} \begin{bmatrix} x - y \\ \sigma(x) - \sigma(y) \end{bmatrix} \geq 0, \quad \forall x, y \in \mathbb{R}^c.$$

If  $\tilde{X}_{k-1} \in \mathcal{T}_{n_g}^c$  and  $\tilde{X}_k \in \mathcal{T}_{n_g}^c$  satisfy  $0 \preceq \tilde{X}_k \preceq \tilde{X}_{k-1}$ , there exists a multiplier  $T \in \mathcal{T}_{n_g}^c$  that satisfies  $0 \preceq \tilde{X}_k \preceq T \preceq \tilde{X}_{k-1}$ , for which we equivalently write

$$\begin{bmatrix} \tilde{X}_{k-1} - T & 0 \\ 0 & -\tilde{X}_k + T \end{bmatrix} \succeq 0. \quad (32)$$

Let  $(u_k^1[\mathbf{i}], (u_k^2[\mathbf{i}]) \in \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}})$  be two arbitrary inputs with corresponding outputs  $(y_k^1[\mathbf{i}], (y_k^2[\mathbf{i}])$  pf the GroupSort activation layer. We multiply (32) with  $[(u_k^1 - u_k^2)^\top (y_k^1 - y_k^2)^\top]$  from the left and its transpose from the right and further sum over  $\mathbf{i} \in \mathbb{N}_0^d$ , to obtain

$$\begin{aligned} & V_{X_{k-1}}(u_k^1, u_k^2) - V_{X_k}(y_k^1, y_k^2) \\ & \geq \sum_{\mathbf{i} \in \mathbb{N}_0^d} \begin{bmatrix} u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}] \\ y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}] \end{bmatrix}^\top \begin{bmatrix} T & 0 \\ 0 & -T \end{bmatrix} \begin{bmatrix} u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}] \\ y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}] \end{bmatrix} \geq 0, \end{aligned}$$

where the last inequality follows from Lemma 13.

### C. Proof of Lemma 10

Let two arbitrary inputs  $(u_k^1[\mathbf{i}], (u_k^2[\mathbf{i}]) \in \ell_{2e}^{c_{k-1}}(\mathbb{N}_0^{d_{k-1}})$  be given and let  $(y_k^1[\mathbf{i}], (y_k^2[\mathbf{i}]), (x^1[\mathbf{i}], (x^2[\mathbf{i}])$  denote the corresponding state and output response of the layer  $\ell_k$ , where  $x^m[\mathbf{i}] = [x_1^m[\mathbf{i}]^\top \cdots x_{d_{k-1}}^m[\mathbf{i}]^\top]^\top$ ,  $m = 1, 2$ . We left/right multiply (15) with

$$[(x^1[\mathbf{i}] - x^2[\mathbf{i}])^\top (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}])^\top (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}])^\top]$$

and its transpose, respectively, and obtain

$$\begin{aligned} & \sum_{j=1}^d (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}])^\top P_j (x_j^1[\mathbf{i}] - x_j^2[\mathbf{i}]) \\ & + (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}])^\top \tilde{X}_{k-1} (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}]) \\ & + 2(y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}])^\top \Lambda_k (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}]) - 2(y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}])^\top \\ & \quad \Lambda_k (\mathbf{C}_k(x_k^1[\mathbf{i}] - x_k^2[\mathbf{i}]) + \mathbf{D}_k(u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}])) \\ & \leq \sum_{j=1}^d (x_j^1[\mathbf{i} + e_j] - x_j^2[\mathbf{i} + e_j])^\top P_j (x_j^1[\mathbf{i} + e_j] - x_j^2[\mathbf{i} + e_j]) \\ & + (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}])^\top \tilde{X}_k (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}]). \end{aligned}$$

Subsequent summation over all  $\mathbf{i} \in \mathbb{N}_0^{d_{k-1}}$  then yields

$$\begin{aligned} & \sum_{\mathbf{i} \in \mathbb{N}_0^d} (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}])^\top \tilde{X}_{k-1} (u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}]) \\ & + 2(y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}])^\top \Lambda_k (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}]) - 2(y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}])^\top \\ & \quad \Lambda_k (\mathbf{C}_k(x_k^1[\mathbf{i}] - x_k^2[\mathbf{i}]) + \mathbf{D}_k(u_k^1[\mathbf{i}] - u_k^2[\mathbf{i}])) \\ & \leq \sum_{\mathbf{i} \in \mathbb{N}_0^d} (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}])^\top \tilde{X}_k (y_k^1[\mathbf{i}] - y_k^2[\mathbf{i}]), \end{aligned}$$

again using the arguments laid out in the proof of Lemma 2.

By Lemma 5, we conclude that  $2(y_k^1 - y_k^2)^\top \Lambda_k (y_k^1 - y_k^2) - 2(y_k^1 - y_k^2)^\top \Lambda_k (\mathbf{C}_k(x_k^1 - x_k^2) + \mathbf{D}_k(u_k^1 - u_k^2)) \geq 0$  for all  $\mathbf{i} \in \mathbb{N}_0^d$  such that we obtain (13).

### D. Proof of Lemma 12

For some  $u_k^1, u_k^2 \in \mathbb{R}^{n_{u_k}}$  and the corresponding intermediate outputs  $v_k^1, v_k^2 \in \mathbb{R}^{n_{v_k}}$ , and outputs  $y_k^1, y_k^2 \in \mathbb{R}^{n_{y_k}}$  of the ResNet layer (23), we left/right multiply (24) with  $[(u_k^1 - u_k^2)^\top (v_k^1 - v_k^2)^\top (y_k^1 - y_k^2)^\top]$  and its transpose, respectively. We obtain

$$\begin{aligned} & V_{X_{k-1}}(u_k^1, u_k^2) - V_{X_k}(y_k^1, y_k^2) \geq \\ & - 2(v_k^1 - v_k^2)^\top \Lambda_1 (v_k^1 - v_k^2) + 2(v_k^1 - v_k^2)^\top \Lambda_1 W_1 (u_k^1 - u_k^2) \\ & - 2(y_k^1 - y_k^2)^\top \Lambda_2 (y_k^1 - y_k^2) + \\ & 2(y_k^1 - y_k^2)^\top \Lambda_2 ((W_2 v_k^1 + u_k^1) - (W_2 v_k^2 + u_k^2)). \end{aligned}$$

Given that the activation functions are slope-restricted on  $[0, 1]$ , we use Lemma 5 to conclude that  $-2(v_k^1 - v_k^2)^\top \Lambda_1 (v_k^1 - v_k^2) + 2(v_k^1 - v_k^2)^\top \Lambda_1 W_1 (u_k^1 - u_k^2) \geq 0$  and  $-2(y_k^1 - y_k^2)^\top \Lambda_2 (y_k^1 - y_k^2) + 2(y_k^1 - y_k^2)^\top \Lambda_2 ((W_2 v_k^1 + u_k^1) - (W_2 v_k^2 + u_k^2)) \geq 0$ , respectively. It follows that  $V_{X_{k-1}}(u_k^1, u_k^2) - V_{X_k}(y_k^1, y_k^2) \geq 0$ .

### E. Proof of Theorem 1

We prove Theorem 1 by induction.

**Induction hypothesis:** If for some  $(Q_1, \dots, Q_l)$ ,  $(R_1, \dots, R_l)$ ,  $(S_1, \dots, S_l)$ ,  $Z_l$  and  $\gamma > 0$

$$\begin{bmatrix} Q_l - Z_l & S_l & & & \\ S_l^\top & R_l + Q_{l-1} & S_{l-1} & & \\ & S_{l-1}^\top & R_{l-1} + Q_{l-2} & \ddots & \\ & & \ddots & \ddots & S_1 \\ & & & S_1^\top & R_1 + \gamma^2 I \end{bmatrix} \preceq 0 \quad (33)$$

is satisfied, then there exists a sequence of matrices  $(X_0, \dots, X_l)$  such that  $X_0 = \gamma^2 I$ ,  $X_l = Z_l$  and (28) holds.

**Start of induction:**  $l = 1$ . Assume

$$\begin{bmatrix} Q_1 - Z_1 & S_1 \\ S_1^\top & R_1 + \gamma^2 I \end{bmatrix} \preceq 0$$

holds for some  $Q_1, R_1, S_1, \gamma > 0$ , and  $Z_1 \succeq 0$ . Then obviously (28) is satisfied with  $X_0 = \gamma^2 I$ ,  $X_1 = Z_1$ :

$$\begin{bmatrix} Q_1 & S_1 \\ S_1^\top & R_1 \end{bmatrix} \preceq \begin{bmatrix} Z_1 & 0 \\ 0 & -\gamma^2 I \end{bmatrix}.$$

**Induction step:**  $l \rightarrow l + 1$ . Assume that our induction hypothesis holds for  $l$ . Let for some  $(Q_1, \dots, Q_{l+1})$ ,  $(R_1, \dots, R_{l+1})$ ,  $(S_1, \dots, S_{l+1})$ ,  $Z_{l+1}$ ,  $\gamma > 0$  the inequality

$$\begin{bmatrix} Q_{l+1} - Z_{l+1} & S_{l+1} & & & \\ S_{l+1}^\top & R_{l+1} + Q_l & S_l & & \\ & S_l^\top & R_l + Q_{l-1} & \ddots & \\ & & \ddots & \ddots & S_1 \\ & & & S_1^\top & R_1 + \gamma^2 I \end{bmatrix} \preceq 0 \quad (34)$$

hold, which implies  $Q_{l+1} - Z_{l+1} \preceq 0$ . There exists an orthogonal matrix  $V$  of the eigenvectors of  $Q_{l+1} - Z_{l+1}$  that diagonalizes  $Q_{l+1} - Z_{l+1}$  by a similarity transformation, i.e.,  $V^\top(Q_{l+1} - Z_{l+1})V$  is a diagonal matrix. We construct  $V = [V_1 \ V_2]$  in such a way that  $V^\top(Q_{l+1} - Z_{l+1})V = \text{diag}(0, \dots, 0, v_1, \dots, v_n) = \text{blkdiag}(0, D)$ ,  $V_1^\top(Q_{l+1} - Z_{l+1})V_1 = 0$ ,  $V_2^\top(Q_{l+1} - Z_{l+1})V_2 = D$ , where  $v_1, \dots, v_n < 0$  and  $n$  is the rank of  $Q_{l+1} - Z_{l+1}$ . Next, we left and right multiply (34) with the full-rank matrix  $\text{diag}(V^\top, I)$  and its transpose  $\text{diag}(V, I)$ , respectively, which yields

$$\begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & D \end{bmatrix} & \begin{bmatrix} 0 \\ V_2^\top S_{l+1} \end{bmatrix} \\ \begin{bmatrix} 0 & S_{l+1}^\top V_2 \end{bmatrix} & R_{l+1} + Q_l & S_l & & \\ & \ddots & \ddots & \ddots & S_1 \\ & & S_1^\top & R_1 + \gamma^2 I \end{bmatrix} \preceq 0 \quad (35)$$

and further, we drop the  $c_{l+1} - n$  zero rows and columns of (35), resulting in

$$\begin{bmatrix} D & V_2^\top S_{l+1} & & & \\ S_{l+1}^\top V_2 & R_{l+1} + Q_l & S_l & & \\ & S_l^\top & R_l + Q_{l-1} & \ddots & \\ & & \ddots & \ddots & S_1 \\ & & & S_1^\top & R_1 + \gamma^2 I \end{bmatrix} \preceq 0. \quad (36)$$

We now apply the Schur complement to (36) with respect to  $D$ , which yields that (36) is negative semi-definite if and only if (33) and  $D \prec 0$  hold, where  $Z_l = S_{l+1}^\top V_2 D^{-1} V_2^\top S_{l+1} - R_{l+1}$ . Given that the diagonal matrix  $D$  has only entries  $v_1, \dots, v_n < 0$ ,  $D \prec 0$  is satisfied and  $D$  is invertible. By the induction hypothesis, there exists a sequence of matrices  $X_0, \dots, X_l$  such that  $X_0 = \gamma^2 I$ ,  $X_l = Z_l$ , (28). The equality  $X_l = Z_l$  implies that there exists at least one  $X_l$  that satisfies  $X_l \preceq Z_l$  and  $\begin{bmatrix} Q_l & S_l \\ S_l^\top & R_l \end{bmatrix} \preceq \begin{bmatrix} X_l & 0 \\ 0 & -X_{l-1} \end{bmatrix}$ . Here,  $X_l \preceq Z_l$  reads  $X_l \preceq$

$S_{l+1}^\top V_2 D^{-1} V_2^\top S_{l+1} - R_{l+1}$ . By the Schur complement, we then get

$$\begin{bmatrix} D & V_2^\top S_{l+1} \\ S_{l+1}^\top V_2 & R_{l+1} + X_l \end{bmatrix} \preceq 0,$$

to which we again add the dropped  $c_{l+1} - n$  zero rows and columns, yielding

$$\begin{bmatrix} V^\top(Q_l - Z_l)V & V^\top S_{l+1} \\ S_{l+1}^\top V & R_{l+1} + X_l \end{bmatrix} \preceq 0. \quad (37)$$

Subsequently, we left and right multiply (37) with  $\text{diag}(V, I)$  and its transpose  $\text{diag}(V^\top, I)$ , respectively, yielding

$$\begin{bmatrix} Q_{l+1} - Z_{l+1} & S_{l+1} \\ S_{l+1}^\top & R_{l+1} + X_l \end{bmatrix} \preceq 0,$$

and we further set  $X_{l+1} = Z_{l+1}$ , which concludes the induction step.

The statement of the theorem is a special case of our induction hypothesis for  $Z_l = I$ .

## APPENDIX B FURTHER LMI CONSTRAINTS

### A. Subnetworks

Usually we consider the combination of a linear layer with a nonlinear activation as shown in Section II-B and formulate LMI constraints for this combination. However, combining multiple layers is also possible. While producing larger LMI constraints, we renounce the use of the decision variables at the transition of layers, i.e.,  $X_k$ , which reduces the number of decision variables. The following LMIs state the corresponding constraints.

**Lemma 14.** Consider the  $k$ -th layer to be a fully connected subnetwork  $(\sigma_l \circ \mathcal{L}_{l-1} \circ \dots \circ \sigma_2 \circ \mathcal{L}_1)_k$  with activation functions that are slope-restricted in  $[0, 1]$ . For some  $X_k \in \mathcal{H}_{\mathcal{L}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{L}}^u$ , this subnetwork satisfies (13) if there exist  $\Lambda_j \in \mathbb{D}_+^{n_{y_j}}$ ,  $j = 1, \dots, l$ , such that  $\mathcal{G}_{\mathcal{L}}(X_{k-1}, X_k, \nu_k) :=$

$$\begin{bmatrix} X_{k-1} & -W_1^\top \Lambda_1 & 0 & \dots & 0 \\ -\Lambda_1 W_1 & 2\Lambda_1 & -W_2^\top \Lambda_1 & \ddots & \vdots \\ 0 & -\Lambda_2 W_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 2\Lambda_{l-1} & -W_l^\top \Lambda_l \\ 0 & \dots & 0 & -\Lambda_l W_l & 2\Lambda_l - X_k \end{bmatrix} \succeq 0. \quad (38)$$

**Lemma 15.** Consider the  $k$ -th layer to be a fully convolutional subnetwork  $(\sigma_l \circ \mathcal{C}_{l-1} \circ \dots \circ \sigma_2 \circ \mathcal{C}_1)_k$  with activation functions that are slope-restricted in  $[0, 1]$ . For some  $X_k \in \mathcal{H}_{\mathcal{C}}^y$  and  $X_{k-1} \in \mathcal{H}_{\mathcal{C}}^u$ , this subnetwork satisfies (13) if there exist  $\Lambda_j \in \mathbb{D}_+^{c_j}$ ,  $\mathbf{P}_j = \text{blkdiag}(P_1^j, \dots, P_d^j)$ ,  $P_i^j \in \mathbb{S}_+^{n_i}$ ,  $i = 1, \dots, d$ ,  $j = 1, \dots, l$  such that (39).

### B. ResNets

In Section III-H, we briefly introduced ResNet layers with two fully connected linear layers in the residual path and presented LMI conditions that imply (13) for such layers. We now consider ResNet layers (22) that skip multiple fully

$$\mathcal{G}_C(X_{k-1}, X_k, \nu) :=$$

$$\begin{bmatrix} X_{k-1} - B_1^\top P_1 B_1 & -B_1^\top P_1 A_1 & -D_1^\top \Lambda_1 & & & & & \\ -A_1^\top P_1 B_1 & P_1 - A_1^\top P_1 A_1 & -C_1^\top \Lambda_1 & & & & & \\ -\Lambda_1 D_1 & -\Lambda_1 C_1 & 2\Lambda_1 - B_2^\top P_2 B_2 & -B_2^\top P_2 A_2 & -D_2^\top \Lambda_2 & & & \\ & & -A_2^\top P_2 B_2 & P_2 - A_2^\top P_2 A_2 & -C_2^\top \Lambda_2 & & & \\ & & -\Lambda_2 D_2 & -\Lambda_2 C_2 & 2\Lambda_2 - B_3^\top P_3 B_3 & \ddots & & \\ & & & & \ddots & \ddots & & \\ & & & & & & -B_l^\top P_l A_l & -D_l^\top \Lambda_l \\ & & & & & & -A_l^\top P_l B_l & P_l - A_l^\top P_l A_l & -C_l^\top \Lambda_l \\ & & & & & & -\Lambda_l D_l & -\Lambda_l C_l & 2\Lambda_l - X_k \end{bmatrix} \succeq 0 \quad (39)$$

connected layers, i.e., where  $\mathcal{M} = \mathcal{L}_l \circ \dots \circ \sigma \circ \mathcal{L}_1$ , or multiple convolutional layers, i.e., where  $\mathcal{M} = \mathcal{C}_l \circ \dots \circ \sigma \circ \mathcal{C}_1$ . For such ResNet layers, we can state the following lemmas.

**Lemma 16.** *Consider the  $k$ -th layer to be a ResNet layer (22) with  $\mathcal{M} = \mathcal{L}_l \circ \dots \circ \sigma \circ \mathcal{L}_1$  and activation functions that are slope-restricted in  $[0, 1]$ . The ResNet layer (22) satisfies (13) if there exist  $\Lambda_j \in \mathbb{D}_+^{n_{y_k}}$ ,  $j = 1, \dots, l$  such that*

$$\mathcal{G}_C(X_{k-1}, X_k, \nu_k) + \begin{bmatrix} 0 & \dots & 0 & -\Lambda_l \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ddots & \vdots & \\ -\Lambda_l & 0 & \dots & 0 \end{bmatrix} \succeq 0,$$

where  $\mathcal{G}_C(X_{k-1}, X_k, \nu_k)$  is defined by (38).

Lemma 16 gives and LMI condition for a ResNet layer with fully-connected layers in  $\mathcal{M}$ , whereas Lemma 17 is concerned with convolutional layers in  $\mathcal{M}$ .

**Lemma 17.** *Consider the  $k$ -th layer to be a ResNet layer (22) with  $\mathcal{M} = \mathcal{C}_l \circ \dots \circ \sigma \circ \mathcal{C}_1$  and activation functions that are slope-restricted in  $[0, 1]$ . The ResNet layer (22) satisfies (13) if there exist  $\Lambda_j \in \mathbb{D}_+^{c_j}$ ,  $\mathbf{P}_j = \text{blkdiag}(P_1^j, \dots, P_d^j)$ ,  $P_i^j \in \mathbb{S}_+^{n_i}$ ,  $i = 1, \dots, d$ ,  $j = 1, \dots, l$  such that*

$$\mathcal{G}_C(X_{k-1}, X_k, \nu_k) + \begin{bmatrix} 0 & \dots & 0 & -\Lambda_l \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ddots & \vdots & \\ -\Lambda_l & 0 & \dots & 0 \end{bmatrix} \succeq 0,$$

where  $\mathcal{G}_C(X_{k-1}, X_k, \nu_k)$  is defined by (39).

## APPENDIX C NEURAL NETWORK ARCHITECTURES

We analyze the well-known LeNet-5 [36] and other typical CNN architectures [37] as well as 18-layer residual neural networks inspired by [19]. To describe the NN architectures, similar to [37], we denote a 2-D convolutional layer by  $c(C, K, S)$ , where  $C$  is the number of output channels,  $K$  the symmetric kernel size and  $S$  the symmetric stride. A dense fully connected layer is denoted by  $d(N)$ , where  $N$  is the

TABLE II  
NEURAL NETWORK ARCHITECTURES.

Model	Specification
LeNet-5:	$c(6, 5, 1).p(\text{av}, 2, 2).c(16, 5, 1).p(\text{av}, 2, 2).d(120).d(84).d(10)$
2C2F:	$c(16, 4, 2).c(32, 4, 2).d(100).d(10)$
4C3F:	$c(32, 3, 1).c(32, 4, 2).c(64, 3, 1).c(64, 4, 2).d(512)^2.d(10)$
FC-R18:	$d(64).\text{res}(64, 2)^8.d(10)$
C-R18:	$c(16, 7, 2).p(\text{max}, 3, 2).\text{res}(16, 3, 1, 2)^8.p(\text{av}, 2, 2).d(10)$
LeNet-5:	$c(6, 5, 1).p(\text{max}, 2, 2).c(16, 5, 1).p(\text{max}, 2, 2).d(120).d(84).d(10)$
6C2F:	$c(32, 3, 1)^2.c(32, 4, 2).c(64, 3, 1)^2.c(64, 4, 2).d(512).d(10)$

TABLE III  
SPLITS INTO SUBNETWORKS FOR S-GLIPSDP

Model	Specification
LeNet-5:	$c(6, 5, 1).p(\text{av}, 2, 2).c(16, 5, 1).p(\text{av}, 2, 2) \mid d(120).d(84).d(10)$
2C2F:	$c(16, 4, 2).c(32, 4, 2) \mid d(100).d(10)$
4C3F:	$c(32, 3, 1).c(32, 4, 2).c(64, 3, 1).c(64, 4, 2) \mid d(512) \mid d(512) \mid d(10)$
FC-R18:	$d(64) \mid \text{res}(64, 2) \mid \dots \mid \text{res}(64, 2) \mid d(10)$
C-R18:	$c(16, 7, 2).p(\text{max}, 3, 2) \mid \text{res}(16, 3, 1, 2) \mid \dots \mid \text{res}(16, 3, 1, 2).p(\text{av}, 2, 2) \mid d(10)$
LeNet-5:	$c(6, 5, 1).p(\text{max}, 2, 2).c(16, 5, 1).p(\text{max}, 2, 2) \mid d(120).d(84).d(10)$
6C2F:	$c(32, 3, 1)^2.c(32, 4, 2) \mid c(64, 3, 1)^2 \mid c(64, 4, 2) \mid d(512) \mid d(10)$

number of output neurons. In addition, by  $p(\text{type}, K, S)$  we mean pooling layers of type either average or maximum, with kernel size  $K$  and stride  $S$ . We denote residual layers with convolutional layers in the residual path by  $\text{res}(C, K, S, L)$  where all convolutions are of the same shape and  $L$  denotes the number of layers in the residual path. In addition, we denote a residual layer containing fully connected layers in the residual path by  $\text{res}(N, L)$ ,  $N$  being the number of neurons and  $L$  the number of skipped layers, considering  $\sigma \circ \mathcal{L}$  as one layer. Using the described nomenclature, we list all utilized architectures in Table II.

For the methods S-LipSDP and S-GlipSDP, we require suitable subnetworks, as specified in Table III and Table IV. S-LipSDP requires a split at every pooling layer as it does not allow to include pooling layers by quadratic constraints, and for 6C3F and 4C3F splits are chosen as large as possible before running into memory issues. For S-LipSDP on C-R18 and FC-R18, we apply LipSDP to the residual paths. The sum of the Lipschitz constants of the parallel paths, i.e.,  $1 + \gamma(\text{residual path})$ , provides an upper bound on the Lipschitz constant for the residual layer.

TABLE IV  
SPLITS INTO SUBNETWORKS FOR S-LIPSDP

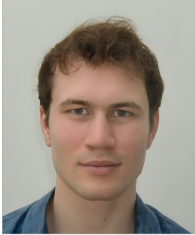
Model	Specification
LeNet-5:	$c(6, 5, 1).p(av, 2, 2) \mid c(16, 5, 1).p(av, 2, 2) \mid d(120).d(84).d(10)$
2C2F:	$c(16, 4, 2).c(32, 4, 2) \mid d(100).d(10)$
4C3F:	$c(32, 3, 1) \mid c(32, 4, 2) \mid c(64, 3, 1) \mid c(64, 4, 2) \mid d(512) \mid d(512).d(10)$
FC-R18:	$d(64) \mid 1 + d(64).d(64) \mid \dots \mid 1 + d(64).d(64) \mid d(10)$
C-R18:	$c(16, 7, 2).p(max, 3, 2) \mid 1 + c(16, 3, 1).c(16, 3, 1) \mid \dots \mid 1 + c(16, 3, 1).c(16, 3, 1) \mid p(av, 2, 2) \mid d(10)$
LeNet-5:	$c(6, 5, 1).p(max, 2, 2) \mid c(16, 5, 1).p(max, 2, 2) \mid d(120).d(84).d(10)$
6C2F:	–

## REFERENCES

- [1] C. M. Bishop, “Neural networks and their applications,” *Review of scientific instruments*, vol. 65, no. 6, pp. 1803–1832, 1994.
- [2] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, “A survey of convolutional neural networks: analysis, applications, and prospects,” *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.
- [3] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, and V. H. C. de Albuquerque, “Deep learning for safe autonomous driving: Current challenges and future directions,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4316–4336, 2020.
- [4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.
- [5] A. Virmaux and K. Scaman, “Lipschitz regularity of deep neural networks: analysis and efficient estimation,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [6] M. Jordan and A. G. Dimakis, “Exactly computing the local Lipschitz constant of ReLU networks,” in *Advances in Neural Information Processing Systems*, 2020, pp. 7344–7353.
- [7] A. Virmaux and K. Scaman, “Lipschitz regularity of deep neural networks: analysis and efficient estimation,” in *Advances in Neural Information Processing Systems*, 2018.
- [8] Y. Tsuzuku, I. Sato, and M. Sugiyama, “Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks,” in *Advances in Neural Information Processing Systems*, 2018.
- [9] P. L. Combettes and J.-C. Pesquet, “Lipschitz certificates for layered network structures driven by averaged activation operators,” *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 2, pp. 529–557, 2020.
- [10] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. Pappas, “Efficient and accurate estimation of lipschitz constants for deep neural networks,” *Advances in neural information processing systems*, vol. 32, 2019.
- [11] F. Latorre, P. Rolland, and V. Cevher, “Lipschitz constant estimation of neural networks via sparse polynomial optimization,” in *International Conference on Learning Representations*, 2020.
- [12] S. Dathathri, K. Dvijotham, A. Kurakin, A. Raghunathan, J. Uesato, R. R. Bunel, S. Shankar, J. Steinhardt, I. Goodfellow, P. S. Liang *et al.*, “Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5318–5331.
- [13] B. Roig-Solvas and M. Sznajder, “A globally convergent lp and socp-based algorithm for semidefinite programming,” *arXiv preprint arXiv:2202.12374*, 2022.
- [14] Z. Wang, A. J. Havens, A. Araujo, Y. Zheng, B. Hu, Y. Chen, and S. Jha, “On the scalability and memory efficiency of semidefinite programs for lipschitz constant estimation of neural networks,” in *International Conference on Learning Representations*, 2024.
- [15] A. Xue, L. Lindemann, A. Robey, H. Hassani, G. J. Pappas, and R. Alur, “Chordal sparsity for Lipschitz constant estimation of deep neural networks,” in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 3389–3396.
- [16] P. Pauli, D. Gramlich, and F. Allgöwer, “Lipschitz constant estimation for 1d convolutional neural networks,” in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 1321–1332.
- [17] D. Gramlich, P. Pauli, C. W. Scherer, F. Allgöwer, and C. Ebenbauer, “Convolutional neural networks as 2-d systems,” *arXiv preprint arXiv:2303.03042*, 2023.
- [18] A. Gu, K. Goel, and C. Ré, “Efficiently modeling long sequences with structured state spaces,” *arXiv preprint arXiv:2111.00396*, 2021.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.
- [20] C. Anil, J. Lucas, and R. Grosse, “Sorting out Lipschitz function approximation,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 291–301.
- [21] R. Roesser, “A discrete state-space model for linear image processing,” *IEEE Transactions on Automatic Control*, vol. 20, no. 1, 1975.
- [22] P. Pauli, D. Gramlich, and F. Allgöwer, “State space representations of the roesser type for convolutional layers,” *arXiv preprint arXiv:2403.11938*, 2024.
- [23] K. Hu, A. Zou, Z. Wang, K. Leino, and M. Fredrikson, “Scaling in depth: Unlocking robustness certification on imagenet,” *arXiv preprint arXiv:2301.12549*, 2023.
- [24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [26] C. J. Bett and M. Lemmon, “On linear fractional representations of multidimensional rational matrix functions,” *ISIS*, vol. 97, p. 008, 1997.
- [27] V. Dumoulin and F. Visin, “A guide to convolution arithmetic for deep learning,” *arXiv preprint arXiv:1603.07285*, 2016.
- [28] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. J. Pappas, “Efficient and accurate estimation of lipschitz constants for deep neural networks,” *arXiv preprint arXiv:1906.04893*, 2023.
- [29] P. Pauli, A. Koch, J. Berberich, P. Kohler, and F. Allgöwer, “Training robust neural networks using Lipschitz bounds,” *IEEE Control Systems Letters*, vol. 6, pp. 121–126, 2021.
- [30] P. Pauli, A. Havens, A. Araujo, S. Garg, F. Khorrami, F. Allgöwer, and B. Hu, “Novel quadratic constraints for extending lipsdp beyond slope-restricted activations,” in *International Conference on Learning Representations*, 2024.
- [31] C. Scherer and S. Weiland, “Linear matrix inequalities in control,” *Lecture Notes, Dutch Institute for Systems and Control, Delft, The Netherlands*, vol. 3, no. 2, 2000.
- [32] J. Lofberg, “Yalmip: A toolbox for modeling and optimization in MATLAB,” in *Proc. of the CACSD Conference*, Taipei, Taiwan, 2004.
- [33] MOSEK ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 9.2.5*, 2020. [Online]. Available: <http://docs.mosek.com/9.2/toolbox/index.html>
- [34] M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado, “Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems,” *SIAM Journal on Optimization*, vol. 28, no. 3, pp. 2654–2689, 2018.
- [35] Y. LeCun and C. Cortes, “MNIST handwritten digit database,” 2010.
- [36] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [37] K. Leino, Z. Wang, and M. Fredrikson, “Globally-robust neural networks,” in *International Conference on Machine Learning*, 2021.



**Patricia Pauli** received the Master's degree in Mechanical Engineering and Computational Engineering from the Technical University of Darmstadt, Germany, in 2019. She has since been a Ph.D. student with the Institute for Systems Theory and Automatic Control under supervision of Prof. Frank Allgöwer and a member of the International Max-Planck Research School for Intelligent Systems (IMPRS-IS). Her research interests are in the area of robust machine learning and learning-based control.



**Dennis Gramlich** received the Master's degree in Engineering Cybernetics and Mathematics from the University of Stuttgart, Germany, in 2020. He was a Ph.D. student with the Institute for Systems Theory and Automatic Control at the University of Stuttgart under the supervision of Prof. Christian Ebenbauer from May 2020 to October 2021 and is now a Ph.D. student with the Institute for Intelligent Control at RWTH Aachen University under the supervision of Prof. Christian Ebenbauer. His research interests are Robust Control and Robust Trajectory Optimization.



**Frank Allgöwer** studied Engineering Cybernetics and Applied Mathematics in Stuttgart and at the University of California, Los Angeles (UCLA), respectively, and received his Ph.D. degree from the University of Stuttgart in Germany. Since 1999 he is the Director of the Institute for Systems Theory and Automatic Control and professor at the University of Stuttgart. His research interests include networked control, cooperative control, predictive control, and nonlinear control with application to a wide range of fields including systems biology. For the years 2017-

2020 Frank served as President of the International Federation of Automatic Control (IFAC) and for the years 2012-2020 as Vice President of the German Research Foundation DFG.