Robustness Against Adversarial Attacks in Neural Networks Using Incremental Dissipativity

Bernardo Aquino, *Student Member, IEEE*, Arash Rahnama, *Member, IEEE*, Peter Seiler, *Fellow, IEEE*, Lizhen Lin, and Vijay Gupta, *Fellow, IEEE*

Abstract—**Adversarial examples can easily degrade the classification performance in neural networks. Empirical methods for promoting robustness to such examples have been proposed, but often lack both analytical insights and formal guarantees. Recently, some robustness certificates have appeared in the literature based on system theoretic notions. This letter proposes an incremental dissipativity-based robustness certificate for neural networks in the form of a linear matrix inequality for each layer. We also propose a sufficient spectral norm bound for this certificate which is scalable to neural networks with multiple layers. We demonstrate the improved performance against adversarial attacks on a feed-forward neural network trained on MNIST and an Alexnet trained using CIFAR-10.**

Index Terms—**Adversarial Attacks, Deep Neural Networks, Robust Design, Passivity Theory, Spectral Regularization.**

I. Introduction

NEURAL networks are powerful structures that can represent any non-linear function through appropriate training for classification and regression tasks. However, neural networks still lack formal performance guarantees, limiting their application in safety-critical systems [1]. Furthermore, many studies have shown the susceptibility of neural networks to small perturbations through carefully crafted adversarial attacks, which, in the case of imaging systems, may be imperceptible to the human eye [1], [2], [3], [4].

Different types of defenses have emerged trying to address this shortcoming, with perhaps the most successful of them being adversarial training [5], [6], [7], [8] and defensive distillation [9], [10]. However, even though neural networks with these defenses empirically show superior performance against adversarial attacks than those without such approaches, these methods do not broadly provide either design insights or formal guarantees on robustness.

In the search for guarantees, other studies have proposed systems-theoretic robustness certificates during training. For instance, Lyapunov stability certificates with incremental quadratic constraints were proposed in [11] and imitation learning in [12]. Lipschitz constant-based certificates have been a particularly fruitful approach, of which we highlight [13], which uses a convex optimization approach, [14] with sparse polynomial optimization, [15], which uses averaged activation operators, and [16], which proposes global Lipschitz training using the Alternate Direction Method of Multipliers (ADMM) method.

In this letter, we take an incremental dissipativity-based approach and derive a robustness certificate for neural network architectures in the form of a Linear Matrix Inequality (LMI). This LMI utilizes Incremental Quadratic Constraints (iQCs) to describe non-linearities in each neuron as proposed in [17]. Recently, [18] used iQCs to define a new class of Recurrent (Equilibrium) Neural Networks for applications such as system identification and robust feedback control. We derive a sufficient condition from this LMI as a bound on the spectral norm of the weight matrix in each layer that is easily implementable. There are three advantages of this approach. One, this approach generalizes the Lipschitz condition that has been imposed on neural networks in previous works to a condition guaranteeing that the input output response of the neural network is sector bounded. The Lipschitz condition defines a particular sector that can be expressed as a special case in our framework. Providing more degrees of freedom to the designer in choosing the sector can lead to a better point in the performance-robustness tradeoff. Two, this approach can scale with the number of layers of the neural network. Consideration of deep neural structures and convolutional layers is a known issue with the existing optimization based paradigms [19] and has been explicitly pointed out as a limitation in works such as [16]. Three, our condition provides an insight on the high empirical effectiveness of the usage of spectral norm regularization for improved generalizability in deep network

structures [20] and training stability of Generative Adversarial Networks (GANs) [21].

Passivity analysis for neural networks has been considered for systems with time delay [22], [23], but no training methodology was offered, and the resulting LMIs can quickly become computationally burdensome. Reference [24] provided a passivity approach for robust neural networks, but no certificate was presented. Further, our theoretical development is for the more general notion of incremental sector bounded neural network layers, which allows for negative passivity indices.

II. BACKGROUND

In this letter, we propose an incremental dissipativity-based approach to quantify and engineer the robustness of neural networks against adversarial attacks. We focus on image classification networks, since the design of adversarial perturbations is perhaps the most developed in such systems [25], [26]. Further, the presence of convolutional layers in most of such architectures poses challenges to existing approaches for guaranteeing robustness. In such applications, our main goal is to reduce the classification error rate on adversarial image test sets. One should notice that the proposed approach can be applied to other applications as well.

Robustness Against Adversarial Attacks: Adversarial attacks on neural networks seek to produce a significant change in the output when the input is perturbed slightly. Thus, designing a network that limits the change in the output as a function of the change in the input can mitigate the effects of adversarial attacks. We claim that enforcing neural network systems to be incrementally dissipative (specifically, sector bounded) can limit the variation on the output, given some perturbation in the input.

Consider a neural network $f_{NN} : \mathcal{R}^n \to \mathcal{R}^m$. In the image classification example, the input may be the vectorized version of the image and the output may be a vector of probabilities corresponding to the various classes. We are interested in the deviation between $f_{NN}(x)$ and $f_{NN}(x + \delta_x)$ where x and $x + \delta_x$ are the actual and adversarial inputs, respectively. Recently, a number of works have pointed out that by enforcing the condition

$$\|f_{NN}(x + \delta_x) - f_{NN}(x)\|_2 \le \gamma \|\delta_x\|_2 \qquad (1)$$

for some $\gamma > 0$, we can guarantee that a norm-bounded adversarial perturbation can shift the output of the neural network only by a bounded amount. Intuitively, this leads to a certain robustness in the classification performance of the neural network. While the Lipschitz constant based approach enforces a symmetric sector in the sense of equation 1) above, the incremental sector boundedness constraint that we consider in this letter generalizes it by allowing the sector slopes to be independent of each other and have arbitrary signs. Further, we can consider convolutional layers as well as compose metrics for each layer to obtain a robustness certificate for the entire network. Finally, we can obtain a computationally efficient training approach to ensure that this certificate is met using a spectral norm based condition. In this sense, our work also sheds light on the empirically observed effectiveness of spectral norm based regularization to promote robustness in neural networks. However, this approach cannot guarantee robustness

for all attacks and data sets. As stated in [27], depending on the specific data set, the margin between decision boundaries can be very narrow, and even a tiny perturbation on output can change the classification label. Therefore, our method implicitly assumes there is some margin between classes.

Sector Boundedness: Dissipativity has been widely used in control systems theory due to its close relation to stability analysis and compositional properties [28], [29], [30].

Definition 1 [31]: A discrete-time system with input $\mathbf{x}(k)$ and corresponding output $\mathbf{y}(k)$ at time k is (Q, S, R) dissipative if $\forall k$, the condition $0 \le s(\mathbf{x}(k), \mathbf{y}(k))$ holds for all admissible initial conditions of the system and all admissible sequences of the inputs $\{\mathbf{x}(j)\}_{j=0}^k$, where $s(\mathbf{x}(k), \mathbf{y}(k))$, is the supply rate given by

$$s(\mathbf{x}, \mathbf{y}) = \mathbf{y}^T Q \mathbf{y} + \mathbf{x}^T S \mathbf{y} + \mathbf{y}^T S^T \mathbf{x} + \mathbf{x}^T R \mathbf{x}, \qquad (2)$$

for matrices Q, S, and R of appropriate dimensions.

Depending on the matrices Q, S, and R, the system will exhibit different dynamical properties. We are particularly interested in the following three cases:

Definition 2 [29], [32]: A (Q, S, R) dissipative discrete-time system is:

1) passive, if $Q = 0$, $S = \frac{1}{2}I$, $R = 0$.
2) strictly passive, if $Q = -\delta I$, $S = \frac{1}{2}I$, $R = -\nu I$, for some $\delta > 0$ and $\nu > 0$.
3) sector bounded with slopes $\frac{1 - \sqrt{1 - 4\delta\nu}}{2\delta}$ and $\frac{1 + \sqrt{1 - 4\delta\nu}}{2\delta}$, if $Q = -\delta I$, $S = \frac{1}{2}I$, $R = -\nu I$, for some δ and ν.

For sector bounded systems, the constants δ and ν are called passivity indexes. We will call a sector bounded system output strictly passive (OSP) when $\delta > 0$ and input strictly passive (ISP) when $\nu > 0$. Finally, all the above definitions can be extended to the case of incrementally QSR-dissipative systems, incrementally passive systems, incrementally strictly passive systems, and incrementally sector bounded systems by defining the supply rate as

$$\begin{aligned} s(\Delta_{\mathbf{x}}, \Delta_{\mathbf{y}}) &= \Delta_{\mathbf{y}}^T Q \Delta_{\mathbf{y}} \\ &\quad + \Delta_{\mathbf{x}}^T S \Delta_{\mathbf{y}} + \Delta_{\mathbf{y}}^T S^T \Delta_{\mathbf{x}} + \Delta_{\mathbf{x}}^T R \Delta_{\mathbf{x}}, \end{aligned} \qquad (3)$$

where $\Delta_{\mathbf{x}} = \mathbf{x_1} - \mathbf{x_2}$ and $\Delta_{\mathbf{y}} = \mathbf{y_1} - \mathbf{y_2}$ for two inputs $\mathbf{x_1}$ and $\mathbf{x_2}$, and the corresponding outputs $\mathbf{y_1}$ and $\mathbf{y_2}$.

III. ENFORCING INCREMENTAL SECTOR BOUNDEDNESS

A. Incremental QSR-Dissipativity for a Single Layer

Consider a neural network layer that receives the vector $\mathbf{x}$ as an input and yields the output $\mathbf{y}$. For a non-convolutional layer, $\mathbf{y} = \phi(\mathbf{Wx} + \mathbf{b})$, where $\phi(\cdot)$ is an element-wise nonlinear activation function. In this letter, we assume that the function is incrementally sector bounded by $[\alpha, \beta]$ for some α and β. This assumption is satisfied by most commonly used functions including tanh, ReLU, and leaky ReLU functions, and is commonly made in works such as [11]. We also assume that the same activation function $\phi(.)$ is used in every element of the layer, which is the case for most neural network layer architectures. We can then state the following result.

Theorem 1: Consider a non-convolutional neural network layer defined by $\mathbf{y} = \phi(\mathbf{Wx} + \mathbf{b})$ where $\phi(.)$ is incrementally sector bounded by $[\alpha, \beta]$. Define $m = \frac{\alpha + \beta}{2}$ and $p = \alpha\beta$. The

layer is incrementally QSR dissipative if the LMI

$$\mathbf{M} = \begin{bmatrix} \mathbf{Q} & \mathbf{S} \\ \mathbf{S}^{\top} & \mathbf{R} \end{bmatrix} + \begin{bmatrix} \mathbf{\Lambda} & -m\mathbf{\Lambda}\mathbf{W} \\ -m\mathbf{W}^{\top}\mathbf{\Lambda} & p\mathbf{W}^{\top}\mathbf{\Lambda}\mathbf{W} \end{bmatrix} \succcurlyeq 0 \qquad (4)$$

is feasible for some $\mathbf{\Lambda}$ defined as

$$\mathbf{\Lambda} = \sum_{1 \leq i \leq n} \lambda_{ii} \mathbf{e}_i \mathbf{e}_i^{\top} \geq 0, \qquad (5)$$

where $\mathbf{e}_i$ is the i-th standard basis vector.

Proof: The proof follows by using an S-procedure following the arguments in [17]. First, for notational ease, for any input $\mathbf{x_i}$ of the neural network layer, denote the output by $\mathbf{y_i}$ and define $\mathbf{v_i} = \mathbf{W}\mathbf{x_i} + \mathbf{b}$. With this notation, we can write:

$$\begin{bmatrix} \mathbf{0} & \mathbf{W} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y_1} - \mathbf{y_2} \\ \mathbf{x_1} - \mathbf{x_2} \end{bmatrix} = \begin{bmatrix} \mathbf{v_1} - \mathbf{v_2} \\ \mathbf{y_1} - \mathbf{y_2} \end{bmatrix}, \qquad (6)$$

for any inputs $\mathbf{x_1}$ and $\mathbf{x_2}$.

The first quadratic form for the S-procedure can be obtained using the iQC approach outlined in [17]. Since the nonlinear function $\phi(\cdot)$ is an element wise function that is incrementally sector bounded by $[\alpha, \beta]$, we have

$$\alpha \leq \frac{\mathbf{y}_1^j - \mathbf{y}_2^j}{\mathbf{v}_1^j - \mathbf{v}_2^j} \leq \beta, \qquad (7)$$

for all $j = 1, \ldots, n$, where n is the number of neurons at the layer under consideration, $\mathbf{v}_1^j$ denotes the input to the nonlinear activation function for the j-th neuron, and $\mathbf{y}_1^j$ denotes its output. The collection of these equations can be written as

$$\begin{bmatrix} \mathbf{v_1} - \mathbf{v_2} \\ \mathbf{y_1} - \mathbf{y_2} \end{bmatrix}^T \begin{bmatrix} p\mathbf{\Gamma} & -m\mathbf{\Gamma} \\ -m\mathbf{\Gamma} & \mathbf{\Gamma} \end{bmatrix} \begin{bmatrix} \mathbf{v_1} - \mathbf{v_2} \\ \mathbf{y_1} - \mathbf{y_2} \end{bmatrix} \leq 0, \qquad (8)$$

for a matrix $\mathbf{\Gamma}$ of Lagrange multipliers with the structure defined in (5) and where $m = \frac{\alpha+\beta}{2}$ and $p = \alpha\beta$ [13]. Finally, (6) yields:

$$\begin{bmatrix} \mathbf{y_1} - \mathbf{y_2} \\ \mathbf{x_1} - \mathbf{x_2} \end{bmatrix}^T \begin{bmatrix} \mathbf{\Gamma} & -m\mathbf{\Gamma}\mathbf{W} \\ -m\mathbf{W}^T\mathbf{\Gamma} & p\mathbf{W}^T\mathbf{\Gamma}\mathbf{W} \end{bmatrix} \begin{bmatrix} \mathbf{y_1} - \mathbf{y_2} \\ \mathbf{x_1} - \mathbf{x_2} \end{bmatrix} \leq 0. \qquad (9)$$

The second quadratic form is obtained from the definition of incremental (QSR) dissipativity as

$$\begin{bmatrix} \mathbf{y_1} - \mathbf{y_2} \\ \mathbf{x_1} - \mathbf{x_2} \end{bmatrix}^T \begin{bmatrix} -\mathbf{Q} & -\mathbf{S} \\ -\mathbf{S}^T & -\mathbf{R} \end{bmatrix} \begin{bmatrix} \mathbf{y_1} - \mathbf{y_2} \\ \mathbf{x_1} - \mathbf{x_2} \end{bmatrix} \leq 0. \qquad (10)$$

Thus, using an S-Procedure on (9) to enforce (10) we obtain:

$$\lambda \begin{bmatrix} \mathbf{\Gamma} & -m\mathbf{\Gamma}\mathbf{W} \\ -m\mathbf{W}^T\mathbf{\Gamma} & p\mathbf{W}^T\mathbf{\Gamma}\mathbf{W} \end{bmatrix} - \begin{bmatrix} -\mathbf{Q} & -\mathbf{S} \\ -\mathbf{S}^T & -\mathbf{R} \end{bmatrix} \succcurlyeq 0, \qquad (11)$$

where $\lambda \geq 0$. The result now follows by defining $\mathbf{\Lambda} = \lambda\mathbf{\Gamma}$. ∎

For a convolutional layer, we can derive a similar result by using the result from [33] that the convolution operation with a filter is equivalent to a matrix multiplication of a block circulant matrix composed by the filter coefficients. Specifically, if the input image $\mathbf{X}$ is convolved with a filter with impulse response $\mathbf{F} \in \mathcal{R}^{n \times n}$, then define the doubly block circulant matrix $\mathbf{C}$ as the matrix

$$\begin{bmatrix} \text{circ}(\mathbf{F}(\mathbf{0}, :)) & \text{circ}(\mathbf{F}(\mathbf{1}, :)) & \ldots & \text{circ}(\mathbf{F}(\mathbf{n-1}, :)) \\ \text{circ}(\mathbf{F}(\mathbf{n-1}, :)) & \text{circ}(\mathbf{F}(\mathbf{0}, :)) & \ldots & \text{circ}(\mathbf{F}(\mathbf{n-2}, :)) \\ \vdots & \vdots & \ddots & \vdots \\ \text{circ}(\mathbf{F}(\mathbf{1}, :)) & \text{circ}(\mathbf{F}(\mathbf{2}, :)) & \ldots & \text{circ}(\mathbf{F}(\mathbf{0}, :)) \end{bmatrix}, \qquad (12)$$

where $\mathbf{F}(\mathbf{p}, :)$ denotes the p-th row of the matrix $\mathbf{F}$ and $\text{circ}(\mathbf{v})$ with a vector $\mathbf{v} \in \mathcal{R}^{n \times 1}$ produces an $n \times n$ circulant matrix with the first row as the vector $\mathbf{v}$. Then, [24], [33] showed that the output $\mathbf{O}$ of the convolution can be expressed as

$$\text{vec}(\mathbf{O}) = \mathbf{C}\text{vec}(\mathbf{X}), \qquad (13)$$

where $\text{vec}(.)$ is the standard vectorization operation. Thus, we can extend Theorem 1 to a convolutional layer as follows.

Corollary 1: Consider the setting of Theorem 1 but with a convolutional neural network layer in which the input is convolved with the filter $\mathbf{F}$, vectorized, and transmitted through an element-wise non-linear activation function $\phi(.)$ that is sector bounded by $[\alpha, \beta]$. Define $m = \frac{\alpha+\beta}{2}$ and $p = \alpha\beta$. The layer is incrementally QSR dissipative if the LMI

$$\mathbf{M} = \begin{bmatrix} \mathbf{Q} & \mathbf{S} \\ \mathbf{S}^{\top} & \mathbf{R} \end{bmatrix} + \begin{bmatrix} \mathbf{\Lambda} & -m\mathbf{\Lambda}\mathbf{C} \\ -m\mathbf{C}^{\top}\mathbf{\Lambda} & p\mathbf{C}^{\top}\mathbf{\Lambda}\mathbf{C} \end{bmatrix} \succcurlyeq 0 \qquad (14)$$

is feasible for some $\mathbf{\Lambda}$ defined as

$$\mathbf{\Lambda} = \sum_{1 \leq i \leq n} \lambda_{ii} \mathbf{e}_i \mathbf{e}_i \geq 0, \qquad (15)$$

where $\mathbf{e}_i$ is the i-th standard basis vector and $\mathbf{C}$ is defined as in (12).

We note that by construction, the matrix $\mathbf{\Lambda}$ with the structure in (5) is positive definite.

B. Extension for a Multi Layered Neural Network

The argument given above can be extended to consider the entire neural network instead of one layer. However, as was noted in [16] even for the LMIs resulting from the simpler Lipschitz constraint, this approach quickly becomes computationally cumbersome. Instead, we can utilize the compositional property of (QSR)-dissipativity to ensure that a multi-layered neural network is sector bounded by imposing constraints on each layer separately. We have the following result.

Theorem 2: Consider a neural network with n layers, where each layer i is incrementally QSR dissipative with $\mathbf{Q} = -\delta_i\mathbf{I}$, $\mathbf{S} = 0.5\mathbf{I}$, and $\mathbf{R} = -\nu_i\mathbf{I}$. Then, the neural network is incrementally sector bounded with parameters $\mathbf{Q} = -\delta\mathbf{I}$, $\mathbf{S} = 0.5\mathbf{I}$, and $\mathbf{R} = -\nu\mathbf{I}$ if the matrix $\mathbf{A} \preceq 0$, where

$$\mathbf{A} \triangleq \begin{bmatrix} -\nu_1 + \nu & \frac{1}{2} & 0 & \cdots & -\frac{1}{2} \\ \frac{1}{2} & -\nu_2 + \delta_1 & \frac{1}{2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{2} & -\nu_n + \delta_{n-1} & \frac{1}{2} \\ -\frac{1}{2} & 0 & \cdots & \frac{1}{2} & -\delta_n + \delta \end{bmatrix}.$$

Proof: Proof follows directly be viewing the neural network as a cascade of layers and applying [34, Th. 5]. ∎

Theorem 2 thus provides one way of ensuring that the neural network is sector bounded, and hence, robust. Specifically, we can impose the constraint (4) during the training of the network. However, this requires solving an SDP problem on each gradient descent step or, at best, after a certain number of epochs, as described in [16]. While ensuring $\mathbf{M}_i \succcurlyeq 0$, with $i = 1, \ldots, l$ for each layer separately is computationally more tractable than optimizing the entire neural network through a large SDP, it nonetheless makes the training much slower. This makes it desirable to further reduce the computational complexity as discussed next.

IV. Training Robust Neural Networks

In this section we derive a sufficient condition on the spectral norm of the weight matrix that provides a feasible solution for the LMI (4) for sector bounded systems. We focus once again on a single neural network layer in the setting of Theorem 1 for ease of notation. Denote the spectral norm of matrix $\mathbf{A}$ by $\|\mathbf{A}\|_2$. Further, define the infimum seminorm $\|\mathbf{A}\|_i$ as the square root of the minimum eigenvalue of $\mathbf{A}^\top\mathbf{A}$. While the spectral norm is sub-multiplicative, the infimum seminorm is super-multiplicative. Further, the two norms are related for invertible matrices through the relation $\|\mathbf{A}^{-1}\|_2 = \|\mathbf{A}\|_i^{-1}$ [35, eq. (2.5)]. Finally, for a normal matrix $\mathbf{\Gamma}$, we have [36, Th. 2.5] $\|\mathbf{\Gamma} - I\|_i = \|\mathbf{\Gamma}\|_i - 1$.

Note that for a sector bound on a neural network to lead to bounded perturbation of the output given a perturbation in the input, it is natural to assume that the slope of the upper bound on the sector given by δ is positive. The slope of the lower bound ν can have arbitrary sign, but should be bounded. We make these assumptions in presenting the following result.

Theorem 3: Consider the setting of Theorem 1 with $\mathbf{Q} = -\delta\mathbf{I}$, $\mathbf{S} = 0.5\mathbf{I}$, and $\mathbf{R} = -\nu\mathbf{I}$ with $\delta > 0$. If the following inequalities hold

$$\|\mathbf{W}\|_2 \leq \frac{1}{|m|}\left(1 - \frac{(\delta + 0.5)(1 - |p|\|\mathbf{W}\|_i^2)}{\delta - \nu}\right) \quad (16)$$

$$|p|\|\mathbf{W}\|_i^2 \leq 1, \quad (17)$$

then (4) is feasible with a matrix $\mathbf{\Lambda}$ of the form (5) with

$$\|\mathbf{\Lambda}\|_2 = \|\mathbf{\Lambda}\|_i \geq \frac{\delta - \nu}{1 - |p|\|\mathbf{W}\|_i^2}. \quad (18)$$

Proof: A sufficient condition for the feasibility of the LMI (4) is that the matrix $\mathbf{M}$ be block diagonally dominant with diagonal blocks positive semi-definite [35]. In other words, if there exists a matrix $\mathbf{\Lambda}$ of the form (5) that ensures that the following four conditions are satisfied, then the LMI (4) is feasible for that $\mathbf{\Lambda}$:

$$\mathbf{\Lambda} + \mathbf{Q} \succ 0 \quad (19)$$

$$\mathbf{W}^\top p\mathbf{\Lambda W} + \mathbf{R} \succ 0 \quad (20)$$

$$\left\|(\mathbf{Q} + \mathbf{\Lambda})^{-1}(\mathbf{S} - m\mathbf{\Lambda W})\right\|_2 \leq 1 \quad (21)$$

$$\left\|(\mathbf{W}^\top p\mathbf{\Lambda W} + \mathbf{R})^{-1}(\mathbf{S}^\top - m\mathbf{W}^\top\mathbf{\Lambda})\right\|_2 \leq 1. \quad (22)$$

We now consider these conditions one by one.

1) *Claim 1:* Any $\mathbf{\Lambda}$ that satisfies equation

$$\|\mathbf{\Lambda}\|_i \geq \delta, \quad (23)$$

will ensure that (19) is satisfied. This claim follows by noting that $Q = -\delta\mathbf{I}$, $\delta > 0$, and $\mathbf{\Lambda} \succcurlyeq 0$.

2) *Claim 2:* Any $\mathbf{\Lambda}$ that satisfies the condition

$$|p|\|\mathbf{W}\|_i^2\|\mathbf{\Lambda}\|_i > \nu \quad (24)$$

will ensure that (20) is satisfied. This is because

$$(20) \iff |p|\|\mathbf{W}^\top\mathbf{\Lambda W}\|_i > \nu$$
$$\stackrel{(a)}{\impliedby} |p|\|\mathbf{W}^\top\|_i\|\mathbf{\Lambda}\|_i\|\mathbf{W}\|_i > \nu$$
$$\iff |p|\|\mathbf{W}\|_i^2\|\mathbf{\Lambda}\|_i > \nu,$$

where (a) follows from the supermultiplicativity of the infimum norm.

3) *Claim 3:* Any $\mathbf{\Lambda}$ that satisfies the condition

$$\delta + 0.5 + |m|\|\mathbf{\Lambda}\|_2\|\mathbf{W}\|_2 \leq \|\mathbf{\Lambda}\|_i \quad (25)$$

will ensure that equation (21) is satisfied. This claim follows by noting that

$$(21) \stackrel{(a)}{\impliedby} \left\|(\mathbf{Q} + \mathbf{\Lambda})^{-1}\right\|_2\|\mathbf{S} - m\mathbf{\Lambda W}\|_2 \leq 1$$
$$\stackrel{(b)}{\iff} \|\mathbf{S} - m\mathbf{\Lambda W}\|_2 \leq \|\mathbf{Q} + \mathbf{\Lambda}\|_i$$
$$\iff \|0.5\mathbf{I} - m\mathbf{\Lambda W}\|_2 \leq \|\mathbf{\Lambda} - \delta\mathbf{I}\|_i$$
$$\stackrel{(c)}{\impliedby} 0.5 + |m|\|\mathbf{\Lambda}\|_2\|\mathbf{W}\|_2 \leq \|\mathbf{\Lambda} - \delta\mathbf{I}\|_i$$
$$\iff 0.5 + |m|\|\mathbf{\Lambda}\|_2\|\mathbf{W}\|_2 \leq \|\mathbf{\Lambda}\|_i - \delta$$
$$\iff \delta + 0.5 + |m|\|\mathbf{\Lambda}\|_2\|\mathbf{W}\|_2 \leq \|\mathbf{\Lambda}\|_i,$$

where (a) and (c) follow from submultiplicativity and subadditivity, respectively, of the spectral norm, while (b) follows from the relation $\|\mathbf{A}^{-1}\|_2 = \|\mathbf{A}\|_i^{-1}$.

4) *Claim 4:* Any $\mathbf{\Lambda}$ that satisfies the condition

$$0.5 + |m|\|\mathbf{\Lambda}\|_2\|\mathbf{W}\|_2 \leq |p|\|\mathbf{W}\|_i^2\|\mathbf{\Lambda}\|_i - \nu \quad (26)$$

will ensure that (22) is satisfied. This claim follows by noting that

$$(22) \impliedby \left\|\left(p\mathbf{W}^\top\mathbf{\Lambda W} + \mathbf{R}\right)^{-1}\right\|_2\|\mathbf{S} - m\mathbf{\Lambda W}\|_2 \leq 1$$
$$\iff (\|\mathbf{S}\|_2 + |m|\|\mathbf{\Lambda}\|_2\|\mathbf{W}\|_2) \leq \left\|p\mathbf{W}^\top\mathbf{\Lambda W} + \mathbf{R}\right\|_i$$
$$\iff 0.5 + |m|\|\mathbf{\Lambda}\|_2\|\mathbf{W}\|_2 \leq \left\|p\mathbf{W}^\top\mathbf{\Lambda W} - \nu\mathbf{I}\right\|_i$$
$$\iff 0.5 + |m|\|\mathbf{\Lambda}\|_2\|\mathbf{W}\|_2 \leq \left\|p\mathbf{W}^\top\mathbf{\Lambda W}\right\|_i - \nu$$
$$\impliedby 0.5 + |m|\|\mathbf{\Lambda}\|_2\|\mathbf{W}\|_2 \leq |p|\|\mathbf{W}\|_i^2\|\mathbf{\Lambda}\|_i - \nu.$$

Since (25) is a sufficient condition for (23) and (26) is a sufficient condition for (24), we have shown that if there exists a matrix $\mathbf{\Lambda}$ of the form (5) that ensures that the two conditions (25) and (26) are satisfied, then the LMI (4) is feasible for that $\mathbf{\Lambda}$. The proof now follows by noting that if (16), (17), and (18) hold, then (25) and (26) are satisfied. ∎

This result provides a computationally easy to impose condition that guarantees sector boundedness of the mapping defined by each neural network layer in terms of the spectral norm of the weight matrix. The conditions for each layer can be combined to guarantee the sector boundedness of the entire neural network following Theorem 2. While this theorem provides only a sufficient condition, a few observations can be made from how easy it is to satisfy the condition (16).

- If $p = 0$ (as is the case with ReLU for instance), then the bound on the right hand side of this condition is smaller. This implies that considering leaky ReLU (for which $p \neq 0$) may lead to superior robustness performance as compared to ReLU activation functions, as has been observed empirically in the literature [37].
- Similarly, if $\nu > 0$, then the condition becomes harder to satisfy. Thus, imposing strict passivity on the neural network ($\nu > 0$) may be overly conservative and lead to low performance as compared to simply sector bounding it and allowing $\nu < 0$.

- We emphasize that although the result aligns with the empirical observation in the literature that regularizing the loss function with the spectral norm of the weight matrix leads to superior robustness against adversarial attacks, our result provides an analytical justification of such a procedure and further identifies the region in which the spectral norm should be bounded to get such robustness. Furthermore, by not using an SDP approach, we can expand our technique to deep Neural Network structures, which is a shortcoming of SDP methods, as discussed in [38]. However, the trade-off is an increased conservatism in the spectral norm bound.

- Finally, although our motivation in this letter was to ensure robustness against adversarial perturbations, imposing passivity and sector boundedness on neural networks is of independent interest. For instance, this result can be used to guarantee stability of a system where the controller is implemented as a neural network through standard results in passivity based control.

V. EXPERIMENTAL RESULTS

We now present the experimental setup and performance improvement when the results above are utilized while training a neural network. Implementation is provided in Python using Tensorflow 1.15.4 on https://github.com/beaquino/Robust-Design-of-Neural-Networks-using-Dissipativity. A MATLAB script to obtain sector bounds for each layer to satisfy Theorem 2 is also available.

We use two commonly known data sets for image classification, MNIST [39] and CIFAR-10 [40]. For the MNIST data set, we use a 3 layer feed-forward network with leaky ReLU activation function (with $a = 0.1$, and therefore $\alpha = 0.1$ and $\beta = 1$) on the first two layers. For the CIFAR-10 dataset, we use an Alexnet [41], which is composed of 2 convolution layers, each followed by a max pooling layer and 3 fully connected layers. Leaky ReLU activation function (with $a = 0.1$, and therefore $\alpha = 0.1$ and $\beta = 1$) are used on the first four layers. In some implementations, a final Softmax layer may be utilized for conversion into probabilities. Since Softmax is only an exponential average, we do not consider such a layer without loss of generality. The adversarial attacks chosen for testing are the Fast Gradient Sign Method (FGSM) attack [25], and the Projected Gradient Descent Method attack (PGDM) [26] using a range of strength ϵ in the interval $(0.1, 0.5)$. Strength 0 represents no attack.

We split training and test sets as 86%-14% of the dataset for MNIST and 80%-20% for CIFAR-10, and we train the model for 200 epochs using Adam optimization. Parameters ν and δ should be chosen before the training procedure. We select the pair $(\nu, \delta) = (-2, 0.4)$ as the indices for the entire network, which restricts the neural network to a sector approximately between $(-0.215, 0.465)$. These values were chosen as the basis of comparison, because they presented the best result among other different choices. We tuned them as hyperparameters and optimized them on the MNIST model. Given these values, we select the individual indexes (ν_i, δ_i) for each layer, using Theorem 2, which are then used to calculate the spectral norm bound for each layer.
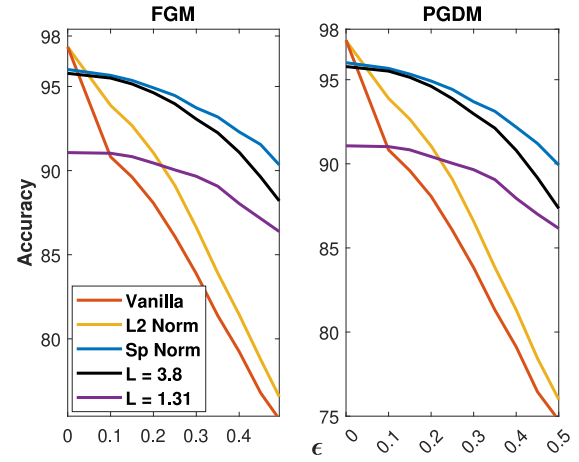

Fig. 1. Accuracy for both Fast Gradient Sign Method attack and Projected Gradient Descent Method attack, compared for a Vanilla model and spectral norm regularized with passivity indexes $(-2, 0.4)$. Network was trained on MNIST dataset.
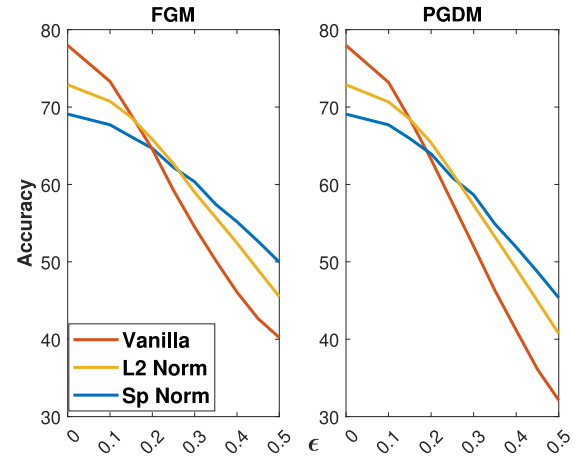

Fig. 2. Accuracy for both Projected Gradient Descent Method attack and Projected Gradient Descent Method attack, compared for a Vanilla model and spectral norm regularized with passivity indexes $(-2, 0.4)$. Network was trained on CIFAR-10 dataset.

Figure 1 presents the results (Sp Norm) for MNIST data, with the accuracy for a test set generated using FGM attack in the left panel and the accuracy for a test set generated using PGDM attack in the right. Figure 2 presents the results (Sp Norm) for CIFAR-10 data, with the left panel presenting the accuracy for a test set generated using FGM attack and the right panel presenting the accuracy for a test set generated using PGDM attack. Also plotted are the accuracy with the commonly used method of L_2-regularization as a comparison point (L2 Norm) and a regularly trained Neural Network (Vanilla). Both figures show an improvement on classification for both cases especially as the attack strength increases, demonstrating the effectiveness of the proposed approach.

A remark can be made about the computational tractability of the proposed approach. Considering the entire neural network at once to impose sector boundedness (or even the simpler condition of Lipschitz constant) is computationally tractable only for shallow networks. Our approach of considering each layer separately, and crucially imposing a spectral norm constraint, is more scalable. Note that while using the spectral norm forces us to calculate the maximum

eigenvalue of a matrix at every gradient descent step, this can be performed efficiently using the power iteration method.

VI. Conclusion

We proposed a robustness certificate for neural networks based on QSR-dissipativity. This method guarantees that a change in the input produces a bounded change in the output, that is, the neural network function is incrementally sector bounded. We first expressed the certificate in the form of a linear matrix inequality. By using the compositional properties of dissipativity, we then decomposed the certificate into one for individual layers of the neural network. We also proposed a sufficient condition based on a spectral norm bound to offer a more computationally tractable problem for deep neural network structures. We presented the results for experiments using a 3 layer feed-forward network and an Alexnet structurure, trained with MNIST and CIFAR-10 respectively. Results showed superior performance when compared to vanilla training and L_2 regularization.

References

[1] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[2] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proc. 10th ACM Workshop Artif. Intell. Security*, 2017, pp. 3–14.

[3] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 274–283.

[4] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.

[5] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, 2016.

[6] A. Shafahi *et al.*, "Adversarial training for free!" 2019, *arXiv:1904.12843*.

[7] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," 2017, *arXiv:1705.07204*.

[8] A. Robey, L. F. O. Chamon, G. J. Pappas, H. Hassani, and A. Ribeiro, "Adversarial robustness with semi-infinite constrained learning," 2021, *arXiv:2110.15767*.

[9] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Security Privacy*, San Jose, CA, USA, 2016, pp. 582–597.

[10] N. Papernot and P. McDaniel, "On the effectiveness of defensive distillation," 2016, *arXiv:1607.05113*.

[11] H. Yin, P. Seiler, and M. Arcak, "Stability analysis using quadratic constraints for systems with neural network controllers," *IEEE Trans. Autom. Control*, early access, Mar. 29, 2021, doi: 10.1109/TAC.2021.3069388.

[12] H. Yin, P. Seiler, M. Jin, and M. Arcak, "Imitation learning with stability and safety guarantees," *IEEE Contr. Syst. Lett.*, vol. 6, pp. 409–414, 2021.

[13] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. J. Pappas, "Efficient and accurate estimation of Lipschitz constants for deep neural networks," 2019, *arXiv:1906.04893*.

[14] F. Latorre, P. Rolland, and V. Cevher, "Lipschitz constant estimation of neural networks via sparse polynomial optimization," 2020, *arXiv:2004.08688*.

[15] P. L. Combettes and J.-C. Pesquet, "Lipschitz certificates for layered network structures driven by averaged activation operators," *SIAM J. Math. Data Sci.*, vol. 2, no. 2, pp. 529–557, 2020.

[16] P. Pauli, A. Koch, J. Berberich, P. Kohler, and F. Allgower, "Training robust neural networks using Lipschitz bounds," *IEEE Contr. Syst. Lett.*, vol. 6, pp. 121–126, 2021.

[17] M. Fazlyab, M. Morari, and G. J. Pappas, "Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming," *IEEE Trans. Autom. Control*, vol. 67, no. 1, pp. 1–15, Jan. 2022, doi: 10.1109/TAC.2020.3046193.

[18] M. Revay, R. Wang, and I. R. Manchester, "Recurrent equilibrium networks: Flexible dynamic models with guaranteed stability and robustness," 2021, *arXiv:2104.05942*.

[19] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Des. Implement. (OSDI)*, 2016, pp. 265–283.

[20] Y. Yoshida and T. Miyato, "Spectral norm regularization for improving the generalizability of deep learning," 2017, *arXiv:1705.10941*.

[21] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.

[22] C. Li and X. Liao, "Passivity analysis of neural networks with time delay," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 52, no. 8, pp. 471–475, Aug. 2005.

[23] S. Xu, W. X. Zheng, and Y. Zou, "Passivity analysis of neural networks with time-varying delays," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 56, no. 4, pp. 325–329, Apr. 2009.

[24] A. Rahnama, A. T. Nguyen, and E. Raff, "Robust design of deep neural networks against adversarial attacks based on Lyapunov theory," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8175–8184.

[25] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[27] K. Nar, O. Ocal, S. S. Sastry, and K. Ramchandran, "Cross-entropy loss and low-rank features have responsibility for adversarial examples," 2019, *arXiv:1901.08360*.

[28] A. Van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*. Berlin, Germany: Springer, 2000.

[29] D. Hill and P. Moylan, "The stability of nonlinear dissipative systems," *IEEE Trans. Autom. Control*, vol. TAC-21, no. 5, pp. 708–711, Oct. 1976.

[30] J. Bao and L. Peter, *Process Control: The Passive Systems Approach*. London, U.K.: Springer-Verlag, Jan. 2007.

[31] C. I. Byrnes and W. Lin, "Losslessness, feedback equivalence, and the global stabilization of discrete-time nonlinear systems," *IEEE Trans. Autom. Control*, vol. 39, no. 1, pp. 83–98, Jan. 1994.

[32] J. C. Willems, "Dissipative dynamical systems part II: Linear systems with quadratic supply rates," *Arch. Rational Mech. Anal.*, vol. 45, no. 5, pp. 352–393, 1972.

[33] H. Sedghi, V. Gupta, and P. M. Long, "The singular values of convolutional layers," 2018, *arXiv:1805.10408*.

[34] H. Yu and P. J. Antsaklis, "A passivity measure of systems in cascade based on passivity indices," in *Proc. 49th IEEE Conf. Decis. Control (CDC)*, Atlanta, GA, USA, 2010, pp. 2186–2191.

[35] D. G. Feingold and R. S. Varga, "Block diagonally dominant matrices and generalizations of the Gerschgorin circle theorem," *Pac. J. Math.*, vol. 12, no. 4, pp. 1241–1250, 1962.

[36] W. H. Bennett, "Block diagonal dominance and the design of decentralized compensation," M.S. thesis, Dept. Electr. Eng., Univ. Maryland, College Park, MD, USA, 1979. [Online]. Available: https://user.eng.umd.edu/~baras/publications/dissertations/1975-1995/79-MS-Bennett.pdf

[37] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.

[38] P. Pauli, N. Funcke, D. Gramlich, M. A. Msalmi, and F. Allgöwer, "Neural network training under semidefinite constraints," 2022, *arXiv:2201.00632*.

[39] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the Web]," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.

[40] A. Krizhevsky, "Learning multiple layers of features from tiny images," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Rep. TR-2009, 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst. Vol. 1*, 2012, pp. 1097–1105.