Lipschitz Continuity in Deep Learning: A Systematic Review of Theoretical Foundations, Estimation Methods, Regularization Approaches and Certifiable Robustness

Anonymous authors
Paper under double-blind review

Abstract

Lipschitz continuity is a fundamental property of neural networks that characterizes their sensitivity to input perturbations. It plays a pivotal role in deep learning, governing **robustness**, **generalization** and **optimization dynamics**. Despite its importance, research on Lipschitz continuity is scattered across various domains, lacking a unified perspective. This paper addresses this gap by providing a systematic review of Lipschitz continuity in deep learning. We explore its **theoretical foundations**, **estimation methods**, **regularization approaches**, and **certifiable robustness**. By reviewing existing research through the lens of Lipschitz continuity, this survey serves as a comprehensive reference for researchers and practitioners seeking a deeper understanding of Lipschitz continuity and its implications in deep learning. Code: `https://anonymous.4open.science/r/lipschitz_survey-DECE/`

1 Introduction

Deep learning has achieved remarkable success across various domains, including computer vision, natural language processing, and graph-based learning. Advances in state-of-the-art architectures, such as convolutional neural networks (CNNs) (Krizhevsky et al., 2012), transformers (Vaswani et al., 2017; Brown et al., 2020), and graph neural networks (GNNs) (Kipf & Welling, 2017), as well as their large-scale applications, including large language models (LLMs) (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023; OpenAI et al., 2024; DeepSeek-AI et al., 2025) and large vision-language models (LVLMs) (Radford et al., 2021; Li et al., 2022; Team et al., 2025), have significantly expanded the frontiers of deep learning, enabling powerful capabilities across a wide range of tasks.

Beyond their remarkable success, ensuring safety *e.g.*, (Amodei et al., 2016), robustness (Madry et al., 2018) and generalization, remains a critical challenge. These properties are intrinsically linked to the sensitivity of neural networks to input perturbations (Madry et al., 2018; Tsipras et al., 2019; Hendrycks & Dietterich, 2019; Amerehi & Healy, 2025), which can significantly impact their reliability in real-world applications. For instance, adversarial vulnerabilities (Goodfellow et al., 2015; Carlini & Wagner, 2017), and out-of-distribution (OOD) generalization (Sokolić et al., 2017; Neyshabur et al., 2017; Bartlett et al., 2017; Hendrycks et al., 2021) are major concerns in developing reliable deep learning models. A fundamental mathematical concept that characterizes network sensitivity and governs these properties is **Lipschitz continuity**, which measures the upper bound of the outputs of neural networks to inputs.

Although numerous studies have explored Lipschitz continuity from both theoretical (Bartlett et al., 2017; Luo et al., 2025a) and applied (Arjovsky et al., 2017; Gulrajani et al., 2017; Miyato et al., 2018) perspectives, a unified survey that consolidates these insights is still lacking. Existing works primarily focus on isolated aspects, such as adversarial robustness (Szegedy et al., 2014; Madry et al., 2018; Weng et al., 2018b), regularization (Arjovsky et al., 2017; Gulrajani et al., 2017), or theoretical bounds without practical considerations (Bartlett et al., 2017). The absence of a comprehensive survey integrating these perspectives leaves a critical gap in the literature. To address this, we provide a systematic review of Lipschitz continuity in deep learning, covering its **theoretical foundations**, **estimation methods**, **regularization approaches in optimization and architecture**, and **certifiable robustness**. By synthesizing key findings across these domains, this survey serves as a valuable resource for researchers and practitioners seeking a deeper understanding of Lipschitz continuity and its role in trustworthy learning systems. Beyond conducting a survey, we also critically distill the existing knowledge, present it in a more accessible manner, and provide the results or proofs **missing** or **incorrect** in the literature.

1.1 Scope

This survey aims to consolidate existing knowledge across these domains, providing a comprehensive resource for researchers and practitioners seeking to understand and leverage Lipschitz continuity in deep learning. To ensure a high-quality review, we focus primarily on the literature published in prestigious and other well-regarded sources. The scope of this survey is organized as below:

1. Section 2: Theoretical Foundations presents the formal definitions and fundamental properties of Lipschitz continuity, followed by key theoretical developments in deep learning. It covers the mathematical foundations, Lipschitz properties of activation functions and neural networks, Lipschitz analysis of multi-head self-attention, complexity-theoretic generalization bounds, and advances in training dynamics.
2. Section 3: Estimation Methods surveys the principal methods for estimating the Lipschitz constants of neural networks, including derivative bound propagation, spectral alignment, convex optimization relaxations, and relaxed integer programming approaches.
3. Section 4: Regularization Approaches reviews approaches for enforcing Lipschitz continuity in neural networks through both optimization-based and architectural approaches, covering weight regularization, gradient regularization, activation regularization, and class-margin regularization.
4. Section 5: Certifiable Robustness examines methods for achieving certifiable robustness through Lipschitz bounds, including the theoretical foundations of Lipschitz-margin robustness radii, certification via global Lipschitz bounds, certification via local Lipschitz bounds and certificated robustness in LLMs.

1.2 Contributions

Our key contributions are as follows:

1. **A Unified Systematic Review**. We close a critical gap in the literature by conducting a comprehensive survey of Lipschitz continuity in deep learning. Our review unites disparate research threads — spanning theoretical foundations, estimation methods, regularization approaches, architectural designs, and certified robustness — into a coherent framework that highlights their interconnections from the lens of Lipschitz continuity.
2. **Comprehensive Analysis**. We provide in-depth examinations of existing techniques organized into four categories:
   - **Theoretical Foundations**. We review the formal definitions and properties of Lipschitz continuity, analyze Lipschitz behavior of activations and network architectures, and synthesize key results on generalization bounds, optimization convergence, and training-induced Lipschitz dynamics.
   - **Estimation Methods**. From power-iteration spectral-norm approximations and extreme-value theory to randomized and interval-analysis approaches, we compare each method's tightness, computational overhead, and scalability to modern architectures.
   - **Regularization Approaches**. We critically assess gradient-based penalties, spectral normalization, orthogonal and Parseval constraints, and novel Lipschitz-bounded activations, examining their theoretical guarantees, implementation trade-offs, and empirical effects on robustness, generalization, and training dynamics.

- **Certifiable Robustness**. We survey Lipschitz-based certification frameworks — covering global and local Lipschitz bounds, path-based and convex-relaxation certificates, randomized smoothing, and Lipschitz-constrained architectures — analyzing their provable guarantees, computational requirements, and applicability to real-world deep models.

3. **Results and Proofs**. We contribute new and complementary results, including corrected Lipschitz constants for common activation functions (validated by our numerical experiments) and a Lipschitz continuity bound for arbitrary neural networks represented as directed acyclic graphs (DAGs). Furthermore, we provide theoretical results **missing or incorrect** from prior work, such as rigorous proofs of the closed-form expressions for the Lipschitz constants of common activation functions and a general perturbation bound for certifiable robustness based on Hölder's conjugacy relation.

2 Theoretical Foundations

Lipschitz continuity is a foundational concept in deep learning theory, offering a rigorous framework for quantifying a neural network's sensitivity to input perturbations. It plays a central role in understanding robustness, generalization, and optimization dynamics. Throughout, we use operator $\mathrm{Lip}_p[\bullet]$ to denote the p-norm Lipschitz constant for $\bullet$, and use operator $\mathrm{Lip}[\bullet]$ to denote the 2-norm Lipschitz constant for $\bullet$.

2.1 Preliminaries

Definition 2.1 (p-Norm in Finite-Dimensional Vector Spaces)**.** Let $V \subseteq \mathbb{R}^d$ be a finite-dimensional vector space. The p-norm (*i.e.* ℓ_p-norm) in V is defined as:

$$\|z\|_p = \begin{cases} \left(\sum_{i=1}^{d} |z_i|^p\right)^{1/p}, & \text{if } 1 \le p < \infty, \\ \max_{i=1,\ldots,d} |z_i|, & \text{if } p = \infty, \end{cases} \tag{1}$$

where $z = (z_1, z_2, \ldots, z_d) \in V$ (Rudin, 1987, Definition 3.6, § 3).

Definition 2.2 (Globally K-Lipschitz Continuous)**.** Let $f : X \mapsto Y$ be a function, where $X \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}^c$. The function f is said to be *globally K-Lipschitz continuous* (with respect to the 2-norm) if there exists a constant scalar $K > 0$ such that:

$$\|f(u) - f(v)\|_2 \le K\|u - v\|_2, \quad \forall\, u, v \in X, \tag{2}$$

(Rudin, 1976, Theorem 9.19, § 9). Throughout, unless specified otherwise, the *Lipschitz norm* is assumed to be the 2-norm.

Definition 2.3 (Globally $K_{p\to q}$-Lipschitz Continuous)**.** Let:

$$1 \le p, q \le \infty \tag{3}$$

be two scalars. A function $f : X \to Y$, with $X \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}^c$, is said to be *$K_{p\to q}$-Lipschitz continuous* if there exists a constant $K_{p\to q} > 0$ such that:

$$\|f(u) - f(v)\|_q \;\le\; K_{p\to q}\, \|u - v\|_p, \quad \forall\, u, v \in X, \tag{4}$$

where $\|\cdot\|_p$ and $\|\cdot\|_q$ denote the ℓ_p and ℓ_q norms, respectively. Particularly, $K_{2\to 2}$-Lipschitz continuous reduces to K-Lipschitz continuous with respect to 2-norm.

Remark 2.4. Definition 2.3 (Globally $K_{p\to q}$-Lipschitz Continuous) is non-standard in the mathematical literature; however, it is useful for discussions of certifiable robustness in the context of deep learning.

Lemma 2.5 (Lipschitz Constant Bounds Gradient Norm)**.** *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function. Then f is K-Lipschitz continuous (with respect to the 2-norm) if and only if*

$$K \ge \sup_{x \in \mathrm{dom}(f)} \|\nabla f(x)\|_2, \tag{5}$$

(Rudin, 1976, Theorem 9.19, § 9). In the case $0 < K < 1$, the function f is called a contraction map. In particular, for the case that $\mathrm{dom}(f)$ *is a convex set — any points connected through the* Mean Value Theorem *remain within* $\mathrm{dom}(f)$*, thus the tight bound is achieved such that:*

$$K = \sup_{x \in \mathrm{dom}(f)} \|\nabla f(x)\|_2. \tag{6}$$

Remark 2.6. In Lemma 2.5 (Lipschitz Constant Bounds Gradient Norm), of the deep learning sense, the domain of a neural network $f : \mathbb{R}^m \to \mathbb{R}$ can be assumed to be a convex set on $\mathbb{R}^m$. Consequently, the tight Lipschitz bound is attained.

2.2 1-Lipschitz Orthogonal Matrix

Definition 2.7 (Skew-Hermitian & Skew-Symmetric Matrix)**.** A matrix A is referred to as a *skew-hermitian matrix*, if A satisfies:

$$A = -A^*. \tag{7}$$

Particularly, if A is a real matrix, then A is also referred to as a *skew-symmetric matrix.*

Proposition 2.8 (Lipschitz Constant of Semi-Orthogonal Matrix)**.** *If a matrix $W \in \mathbb{R}^{m \times n}$ with:*

$$W^\top W = I_n \qquad or \qquad WW^\top = I_m, \tag{8}$$

then W is semi-orthogonal with Lipschitz constant 1. If $m \geq n$, W is an isometric map over the entire $\mathbb{R}^n$; if $m < n$, W is a contraction map on $\ker(W)$ *strictly.*

Proof. **Case 1: $m \geq n$ (Isometric Map).** For $\forall\, x, y \in \mathbb{R}^n$:

$$\begin{aligned} \|W(x-y)\|_2^2 &= (W(x-y))^\top W(x-y) & (9)\\ &= (x-y)^\top W^\top W(x-y) & (10)\\ &= (x-y)^\top I_n (x-y) & (11)\\ &= \|x-y\|_2^2, & (12) \end{aligned}$$

then by definition:

$$\mathrm{Lip}\,[W] = \sup_{x \neq y} \frac{\|W(x-y)\|_2}{\|x-y\|_2} = 1. \tag{13}$$

Case 1: $m < n$ (Contraction on Kernel). For $\exists\, x - y \in \ker(W)$, such that:

$$\|W(x-y)\|_2 = 0 < \|x-y\|_2, \tag{14}$$

then W is a contraction map on the kernel $\ker(W)$ strictly.

□

2.2.1 Matrix Orthogonalization

We also survey useful matrix orthogonalization algorithms below.

Lie Group Exponential Map. Let:

$$U(n) = \Big\{ A \in \mathbb{C}^{n \times n} \mid A^* A = I \Big\} \tag{15}$$

be an unitary group and:

$$\mathfrak{u}(n) = \{ B \in \mathbb{R}^{n \times n} \ : \ B = -B^* \} \tag{16}$$

be a *Skew-Hermitian* group. Then the *Lie Exponential Map* connects $\mathfrak{u}(n)$ and $U(n)$:

$$\exp(A) = I + A + \frac{1}{2}A^2 + \cdots , \tag{17}$$

which is surjective. For a matrix W, its corresponding skew-Hermitian matrix can be constructed by taking:

$$B = W - W^* \tag{18}$$

(Lezcano-Casado & Martínez-Rubio, 2019).

Cayley Transform. If $B \in \mathbb{R}^{n\times n}$ is a *skew-symmetric* matrix, the surjective map:

$$\phi(B) = (I + \frac{1}{2}B)(I - \frac{1}{2}B)^{-1} \tag{19}$$

is referred to as the *Cayley map* which transforms B to an orthogonal matrix $\phi(B)$ (Cayley, 1846; Trockman & Kolter, 2021).

***Björck Orthogonalization* Approximation.** Suppose $W_0 \in \mathbb{R}^{n\times n}$ is not an orthogonal matrix. Let W_k be an iterative sequence such that W_k is the closest orthogonal matrix to W_0 as $k \to \infty$. *Björck Orthogonalization* (Björck & Bowie, 1971) is a differentiable method to iteratively solve the closest orthogonal matrix to a given matrix W_0. Chernodub & Nowicki (2017) use *Björck Orthogonalization* in an optimization problem for constraining the orthogonality of parameter matrices (Chernodub & Nowicki, 2017). *Björck Orthogonalization* (Björck & Bowie, 1971) is defined by:

$$W_{k+1} = W_k\Big(I + \frac{1}{2}Q_k + \cdots + (-1)^p \binom{-\frac{1}{2}}{p} Q_k^p\Big), \tag{20}$$

where W_k is the k-th iterative result, p is a chosen hyper-parameter, $\binom{-\frac{1}{2}}{p}$ is binomial coefficient, and Q_k is:

$$Q_k = I - W_k^\top W_k. \tag{21}$$

As $n \to \infty$, $W_k^\top W_k$ converges to I_n. For a rectangular matrix $W_0 \in \mathbb{R}^{m\times n}$, if $\mathrm{rank}(W_0) = n$, the convergence still holds:

$$W_k^\top W_k \to I_n \qquad \text{as} \qquad n \to \infty. \tag{22}$$

2.3 Functional Inequalities

Proposition 2.9 (Lipschitz Constant of Function Composition)**.** *Let $f : X \to Y$ and $g : Y \to Z$ be two bounded functions. Then the Lipschitz constant of their composition $g \circ f : X \to Z$ admits:*

$$\mathrm{Lip}_p\left[g \circ f\right] \leq \mathrm{Lip}_p\left[g\right]\ \mathrm{Lip}_p\left[f\right]. \tag{23}$$

Proof.

$$\begin{aligned} \mathrm{Lip}_p\left[g \circ f\right] &= \sup_{\|x-y\|_p \neq 0} \frac{\|f(g(x)) - f(g(y))\|_p}{\|x-y\|_p} & (24)\\ &\leq \mathrm{Lip}_p\left[f\right] \sup_{\|x-y\|_p \neq 0} \frac{\|g(x) - g(y)\|_p}{\|x-y\|_p} & (25)\\ &= \mathrm{Lip}_p\left[f\right]\ \mathrm{Lip}_p\left[g\right]. & (26) \end{aligned}$$

□

Proposition 2.10 (Lipschitz Constant of Function Addition)**.** *Let f and g be two bounded functions on* $\mathrm{dom}(g) \cap \mathrm{dom}(f)$*. Then the Lipschitz constant of $g+f$ admits:*

$$\mathrm{Lip}_p\left[g+f\right] \leq \mathrm{Lip}_p\left[g\right] + \mathrm{Lip}_p\left[f\right]. \tag{27}$$

Proof.

$$\begin{aligned}
\mathrm{Lip}_p\left[g+f\right] &= \sup_{\|x-y\|_p \neq 0} \frac{\|f(x)+g(x)-\left[f(y)+g(y)\right]\|_p}{\|x-y\|_p} && (28)\\
&= \sup_{\|x-y\|_p \neq 0} \frac{\|f(x)-f(y)+g(x)-g(y)\|_p}{\|x-y\|_p} && (29)\\
&\leq \sup_{\|x-y\|_p \neq 0} \frac{\|f(x)-f(y)\|_p}{\|x-y\|_p} + \sup_{\|x-y\|_p \neq 0} \frac{\|g(x)-g(y)\|_p}{\|x-y\|_p} && (30)\\
&= \mathrm{Lip}_p\left[g\right] + \mathrm{Lip}_p\left[f\right]. && (31)
\end{aligned}$$

□

Proposition 2.11 (Lipschitz Constant of Function Concatenation)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^m$ and $g : \mathbb{R}^d \to \mathbb{R}^n$ be two Lipschitz continuous functions. Let:*

$$(f \oplus g)(x) := \begin{bmatrix} f(x) \\ g(x) \end{bmatrix} \in \mathbb{R}^{m+n} \tag{32}$$

be a function by concatenating the outputs of f and g. Then, the following inequality holds true:

$$\Big(\mathrm{Lip}_p\left[f \oplus g\right]\Big)^p \leq \Big(\mathrm{Lip}_p\left[f\right]\Big)^p + \Big(\mathrm{Lip}_p\left[g\right]\Big)^p. \tag{33}$$

Proof. For any $x, y \in \mathbb{R}^d$, we have:

$$\begin{aligned}
\big\|(f \oplus g)(x) - (f \oplus g)(y)\big\|_p^p &= \left\| \begin{bmatrix} f(x) \\ g(x) \end{bmatrix} - \begin{bmatrix} f(y) \\ g(y) \end{bmatrix} \right\|_p^p && (34)\\
&= \sum_{i=1}^{m} \Big(|f(x)-f(y)|^p\Big)_i + \sum_{j=1}^{n} \Big(|g(x)-g(y)|^p\Big)_j && (35)\\
&\leq \bigg(\Big(\mathrm{Lip}_p\left[f\right]\Big)^p + \Big(\mathrm{Lip}_p\left[g\right]\Big)^p\bigg) \|x-y\|_p^p. && (36)
\end{aligned}$$

Hence:

$$\Big(\mathrm{Lip}_p\left[f \oplus g\right]\Big)^p = \sup_{x \neq y} \frac{\big\|(f \oplus g)(x) - (f \oplus g)(y)\big\|_p^p}{\|x-y\|_p^p} \leq \Big(\mathrm{Lip}_p\left[f\right]\Big)^p + \Big(\mathrm{Lip}_p\left[g\right]\Big)^p. \tag{37}$$

□

2.4 Sub-Differential Convex Functions

For a function that is not differentiable but is locally Lipschitz continuous — such as the ReLU activation — the Clarke sub-differential provides a generalized notion of gradient that allows one to analyze and bound local Lipschitz constants (Clarke, 1975).

Definition 2.12 (Generalized Gradient)**.** Let $f : \mathbb{R}^m \to \mathbb{R}$ be a function on X. The sub-differential of f at point x is a set, satisfying:

$$\partial_s f(x) = \Big\{ g \mid \forall x, y \in X,\, \exists\, g \in \mathbb{R}^m \text{ s.t. } f(y) \geq f(x) + \langle g, y-x \rangle \Big\}. \tag{38}$$

If f is convex, then:

$$\partial_s f(x) \neq \varnothing \qquad \text{and} \qquad \partial_s f(x) = \{\nabla f(x^*) \mid \forall\, x^* \to x\}. \tag{39}$$

Definition 2.13 (Clarke Sub-Gradient)**.** Let $f : \mathbb{R}^m \to \mathbb{R}$ be a proper convex function on X. If f is locally Lipschitz continuous on X, then by Rademacher's theorem it is differentiable almost everywhere. The Clarke sub-differential f at a point x is defined as the convex hull of all limit points of gradients at differentiable points approaching x. That is, a vector g lies in the Clarke sub-differential at x if there exists a sequence of points $x_k \to x$ ($k \in \mathbb{N}$) at which f is differentiable such that the gradients $\nabla f(x_k)$ converge to g. Then the Clarke sub-differential is a convex hull, satisfying:

$$\partial f(x) = \text{conv}\Big\{g \mid \exists \text{ a sequence } (x_k) \to x \text{ as } k \to \infty, \nabla f(x_k) \to g\Big\} \tag{40}$$

(Clarke, 1975, Definition 1.1). If f is convex, then $\partial f(x) = \partial_s f(x)$. This definition also extends to vector-valued functions (Clarke, 1990).

Lemma 2.14 (Local Lipschitz Constant of Sub-Differential Function)**.** *Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a convex function differentiable almost everywhere on X. Then the $p \to q$ local Lipschitz constant of f on X is given as:*

$$\text{Lip}_{p\to q}\left[f; X\right] = \sup_{G\in\partial f(x)} \sup_{\|v\|_p\neq 0} \frac{\|G^\top v\|_q}{\|v\|_p} = \sup_{G\in\partial f(x)} \sup_{\|v\|_p=1} \|G^\top\|_q, \tag{41}$$

such that:

$$\|f(x) - f(y)\|_q \leq \text{Lip}_{p\to q}\left[f; X\right]\ \|x - y\|_p, \tag{42}$$

for $\forall\, x, y \in X$ in the sense of Clarke sub-gradient (Jordan & Dimakis, 2020; Boyd et al., 2022).

2.5 Lipschitz Continuity of Activation Functions

An activation function:

$$\phi : \mathbb{R}^d \to \mathbb{R}^d \tag{43}$$

is a nonlinear mapping applied after neuron's output at a layer, enabling neural networks to learn complex patterns. For example, if ϕ is piecewise defined as:

$$\phi(x) = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise} \end{cases},$$

then the ϕ is referred to as *ReLU* activation function (Nair & Hinton, 2010).

The Lipschitz constants of commonly used activation functions are provided in Table 1. The constants are derived as the supremum of the 2-norm of the derivatives of activation functions. Proofs are provided in the Appendix A. To numerically validate the results, we have conducted experiments by maximizing the gradient norm using Lemma 2.5 for a further sanity check.

Notes. The Lipschitz constant for the *softmax* function reported in the literature (Gouk et al., 2021) and *Sigmoid* function reported in the literature (Virmaux & Scaman, 2018) as 1. We show in Appendix A.6 that the exact Lipschitz constant of *softmax* is indeed $\frac{1}{2}$, validated by our numerical experiment. This result has been reported a recent concurrent literature (Nair, 2025). We also show in Appendix A.1 that exact Lipschitz constant of *sigmoid* is indeed $\frac{1}{4}$, validated by our numerical experiment. For *Leaky ReLU* and *ELU*, the parameter α is typically chosen with $\alpha < 1$, so their effective Lipschitz constants are usually 1 in practice.

Auto-Differentiation based Numerical Method. Suppose $f : \mathbb{R}^d \to \mathbb{R}$ be a first- and second-order differentiable neural network under domain convex condition (Remark 2.6). Leverage the autograd mechanism of PyTorch, we use gradient descent — *e.g.*, Adam (Kingma & Ba, 2014) — to numerically approximate the supremum:

$$\text{Lip}[f] = \sup_x \|f'(x)\|_2, \tag{44}$$

using Lemma 2.5.

Table 1: Lipschitz constants of activation functions.

Activation	Definition	Lipschitz Constant & Proof	Citations
ReLU	$\max(0, x)$	1	Nair & Hinton (2010, Sec. 2); Virmaux & Scaman (2018, Sec. 5)
Leaky ReLU	$\max(\alpha x, x)$	$\max(1, \alpha)$	Maas et al. (2013, Sec. 2); Virmaux & Scaman (2018, Sec. 5)
Sigmoid	$\frac{1}{1+e^{-x}}$	$\frac{1}{4}$ (Appendix A.1)	Virmaux & Scaman (2018, Sec. 5)
Tanh	$\tanh(x)$	1 (Appendix A.2)	Cybenko (1989); Virmaux & Scaman (2018, Sec. 5)
Softplus	$\log(1 + e^x)$	1 (Appendix A.3)	Dugas et al. (2000, Sec. 2); Virmaux & Scaman (2018, Sec. 5)
ELU	$\begin{cases} x & x > 0 \\ \alpha(e^x - 1) & x \leq 0 \end{cases}$	$\max(1, \alpha)$	Clevert et al. (2016, Sec. 3)
Swish	$x \cdot \sigma(x)$	≈ 1.1 (Appendix A.4)	Ramachandran et al. (2018, Sec. 1)
GELU	$\frac{x}{2}\left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$	≈ 1.1 (Appendix A.5)	Hendrycks & Gimpel (2016, Sec. 2)
Softmax	$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$	$\frac{1}{2}$ (Appendix A.6)	Gouk et al. (2021, Sec. 3.3)

2.6 Lipschitz Continuity of Dot-Product Self-Attention

LLMs (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023; OpenAI et al., 2024; DeepSeek-AI et al., 2025), built on transformer architectures (Vaswani et al., 2017), rely heavily on dot-product self-attention mechanisms, whose Lipschitz continuity governs their sensitivity to input perturbations, such as token-level adversarial attacks. Dot-product self-attention mechanisms form the core of transformer architectures, enabling effective modeling of long-range dependencies in sequences for applications such as natural language processing and vision-language tasks.

Single-Head Dot-Product Self-Attention. For an input sequence matrix $x \in \mathbb{R}^{n\times d}$, where n is the sequence length and d is the input dimension, a self-attention layer computes the output as

$$\text{Attention}(x) = \underbrace{\text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)}_{\text{row-wise softmax}} V \in \mathbb{R}^{n\times m}, \tag{45}$$

(Vaswani et al., 2017, Equation 1) where

$$Q = xW_Q \in \mathbb{R}^{n\times m}, \quad K = xW_K \in \mathbb{R}^{n\times m}, \quad V = xW_V \in \mathbb{R}^{n\times m}, \tag{46}$$

are the *query*, *key*, and *value* matrices, with weight matrices:

$$W_Q, W_K, W_V \in \mathbb{R}^{d\times m}, \tag{47}$$

and m denotes the hidden dimension. The softmax is applied row-wise to normalize attention scores. The scaling factor $\frac{1}{\sqrt{d}}$ stabilizes the computation numerically.

Multi-Head Dot-Product Self-Attention. In practice, transformers employ h parallel self-attention heads to jointly attend to information from different representation subspaces (Vaswani et al., 2017). Each head i has its own projection matrices:

$$W_{Q,i}, W_{K,i}, W_{V,i} \in \mathbb{R}^{d\times m_i}, \tag{48}$$

where $m_i = \frac{m}{h}$, h denotes the number of heads, and produces an output for head i:

$$\text{Attention}_i(x) \in \mathbb{R}^{n\times m_i}. \tag{49}$$

The outputs are concatenated:

$$\text{Attention}(x) = \left[\text{Attention}_1(x), \text{Attention}_2(x), \cdots, \text{Attention}_h(x)\right] \in \mathbb{R}^{n \times m}, \tag{50}$$

produces the final output.

2.6.1 Locally Lipschitz Continuous

Standard dot-product self-attention is locally Lipschitz continuous, and several works have derived explicit upper bounds for its local Lipschitz constant. Let $\mathcal{B}(x_0, \delta) = \{x \mid \|x - x_0\|_2 < \delta\}$ denote a ball centered x_0 with a radius δ. Hu et al. show that for the dot-product self-attention layer within a ball $\mathcal{B}_2(x, \delta)$ is locally Lipschitz continuous bounded by:

$$\text{Lip}\Big[\text{Attention}; \mathcal{B}_2(x, \delta)\Big] \leq n(n+1)\Big(\|x\|_2 + \delta\Big)^2 \Big[\|W_V\|_2 \|W_Q\|_2 \|W_K^\top\|_2 + \|W_V\|_2\Big], \tag{51}$$

where $x \in \mathbb{R}^{n \times d}$, n is the sequence length, d is the input dimension, $W_K, W_Q, W_V \in \mathbb{R}^{d \times m}$ are the *key*, *query* and *value* parameter matrices, and δ is a 2-norm radius (Hu et al., 2024, Theorem 2). Castin et al. provide a tighter bound by $\sqrt{n}$ up to a constant factor, showing that within a ball $\mathcal{B}_2(0, \delta)$, the Lipschitz constant of dot-product self-attention is bounded by:

$$\text{Lip}\Big[\text{Attention}; \mathcal{B}_2(0, \delta)\Big] \leq \sqrt{3}\|W_V\|_2 \left[\left\|\frac{W_K^\top W_Q}{\sqrt{d}}\right\|_2^2 \delta^4 (4n+1) + n\right]^{\frac{1}{2}}, \tag{52}$$

where A is the attention score matrix (Castin et al., 2024, Theorem 3.3).

Most recently, Yudin et al. (2025) refine the bound by explicitly incorporating the Jacobian of the softmax function, given by:

$$\mathcal{M}(P_x) := \nabla \operatorname{softmax}(x) = \operatorname{diag}(P_x) - P_x^\top P_x, \tag{53}$$

where:

$$P_x = \left(\frac{\mathrm{e}^{x_i}}{\sum_j \mathrm{e}^{x_j}}\right), \tag{54}$$

leading that the Lipschitz constant for a single head is bounded by:

$$\text{Lip}\Big[\text{Attention}\Big] \leq \|W_V\|_2 \left(\|P\|_2 + 2\|x\|_2^2 \; \|A\|_2 \; \max_i \|\mathcal{M}(P_{i,:})\|_2\right) \tag{55}$$

with:

$$P = \operatorname{softmax}\Big(xAx^\top\Big) \in \mathbb{R}^{n \times n} \qquad \text{and} \qquad A = \frac{W_Q W_K^\top}{\sqrt{d}} \in \mathbb{R}^{d \times d} \tag{56}$$

(Yudin et al., 2025, Theorem 3).

2.6.2 Globally Non-Lipschitz Continuous

It is worth noting that the Lipschitz constants of both single-head and multi-head dot-product self-attention are not globally bounded. This is because the Lipschitz bound for dot-product self-attention grows with the input sequence length, implying that the dot-product self-attention is not globally Lipschitz continuous (Kim et al., 2021, Theorem 3.1). Kim et al. show that, for a sequence length n, input dimension d, and h heads, the 2-norm Lipschitz bound is:

$$\text{Lip}\left[\text{Attention}\right] \leq \frac{\sqrt{n}}{\sqrt{d/h}} \Big[4\phi^{-1}(n-1) + 1\Big] \left(\sqrt{\sum_{h_i} \|W_{Q,h_i}\|_2^2 \; \|W_{V,h_i}\|_2^2}\right) \|W_O\|_2, \tag{57}$$

and the ∞-norm Lipschitz bound:

$$\mathrm{Lip}_{\infty}\left[\mathrm{Attention}\right] \leq \left[4\phi^{-1}(n-1) + \frac{1}{d/h}\right] \|W_O^{\top}\|_{\infty} \max_{h_i} \|W_{Q,h_i}\|_{\infty} \|W_{Q,h_i}^{\top}\|_{\infty} \max_{h_j} \|W_{V,h_j}^{\top}\|_{\infty}, \tag{58}$$

where:

$$\phi(x) = x \exp(x+1) \tag{59}$$

is an invertible univariate function on $x > 0$, W_{Q,h_i} is the *query* parameter matrix for head h_i, W_{V,h_i} is the *value* parameter matrix for head h_i, and W_O is the projection parameter matrix (Kim et al., 2021, Theorem 3.2). They further conclude that their results can lead that the 2-norm Lipschitz bound is at the scale:

$$\mathrm{Lip}[\mathrm{Attention}] \sim O(\sqrt{n}\log n), \tag{60}$$

and the ∞-norm Lipschitz bound is at the scale:

$$\mathrm{Lip}_{\infty}[\mathrm{Attention}] \sim O(\log n) \tag{61}$$

(Kim et al., 2021, Theorem 3.2). The bounds in (Kim et al., 2021, Theorem 3.2) are complemented by concurrent work (Vuckovic et al., 2020) with a measure-theoretical framework. Vuckovic et al. show that the 1-norm Lipschitz bound is at the scale:

$$\mathrm{Lip}_1[\mathrm{Attention}] \sim O(\sqrt{\log n}) \tag{62}$$

(Vuckovic et al., 2020, Theorem 29 & Corollary 30).

2.7 Lipschitz Continuity of Neural Networks

Definition 2.15 (Feedforward Network)**.** Let $f : \mathbb{R}^m \to \mathbb{R}^n$ denote a feedforward neural network consisting of L layers, defined as the composition:

$$f := h^{(L)} \circ \cdots \circ h^{(2)} \circ h^{(1)}, \tag{63}$$

where each layer $h^{(\ell)}$ is composed of a linear transformation $\phi^{(\ell)}(\cdot)$ followed by a non-linear activation function $\rho^{(\ell)}(\cdot)$:

$$h^{(\ell)} := \rho^{(\ell)} \circ \phi^{(\ell)}. \tag{64}$$

Proposition 2.16 (Lipschitz Constant of a Linear Layer)**.** *Let $\phi^{(\ell)} : \mathbb{R}^{m_\ell} \to \mathbb{R}^{n_\ell}$ denote a linear transformation implemented either as a convolutional layer:*

$$\phi^{(\ell)}(z^{(\ell-1)}) = \boldsymbol{\theta}^{(\ell)} \circledast z^{(\ell-1)} + b^{(\ell)}, \tag{65}$$

or as a fully connected (dense) layer:

$$\phi^{(\ell)}(z^{(\ell-1)}) = \boldsymbol{\theta}^{(\ell)} z^{(\ell-1)} + b^{(\ell)}, \tag{66}$$

where $\boldsymbol{\theta}^{(\ell)}$ is the weight matrix (or convolution kernel), $b^{(\ell)}$ is the bias term, $\circledast$ denotes the convolution operator, and $z^{(\ell-1)}$ is the output from the previous layer. The Lipschitz constant of $\phi^{(\ell)}$ with respect to the Euclidean norm is given by:

$$\mathrm{Lip}[\phi^{(\ell)}] = \|\boldsymbol{\theta}^{(\ell)}\|_2, \tag{67}$$

where $\|\boldsymbol{\theta}^{(\ell)}\|_2$ denotes the spectral-norm of the associated linear operator.

Proposition 2.17 (Spectral-Norm via Singular Value Decomposition)**.** *Let $\boldsymbol{\theta} \in \mathbb{R}^{n\times m}$ be a matrix, admitting a truncated singular value decomposition (SVD):*

$$\boldsymbol{\theta} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top, \quad \textit{with} \quad \sigma_1 > \sigma_2 > \cdots > \sigma_r > 0, \quad r = \operatorname{rank}(\boldsymbol{\theta}), \tag{68}$$

*where σ_i are the singular values, and $\boldsymbol{u}_i$, $\boldsymbol{v}_i$ are the left and right singular vectors, respectively. Then, the spectral-norm of $\boldsymbol{\theta}$ (*i.e.*, its operator norm induced by the Euclidean norm) is given by the largest singular value:*

$$\|\boldsymbol{\theta}\|_2 = \sigma_1 \tag{69}$$

(Horn & Johnson, 2012, Example 5.6.6, § 5).

Proposition 2.18 (Lipschitz Constant Upper Bound of Feedforward Neural Network)**.** *Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be an L-layer feedforward neural network composed of linear transformations $\phi^{(\ell)}$ and activation functions $\rho^{(\ell)}$. The Lipschitz constant of f with respect to the Euclidean norm is defined as:*

$$\operatorname{Lip}[f] = \sup_{\forall\ u\neq v} \frac{\|f(u) - f(v)\|_2}{\|u - v\|_2}, \tag{70}$$

which immediately admits an upper bound by applying Proposition 2.9 (Lipschitz Constant of Function Composition):

$$\operatorname{Lip}[f] \leq \prod_{\ell=1}^{L} \operatorname{Lip}[\rho^{(\ell)}] \prod_{\ell=1}^{L} \operatorname{Lip}[\phi^{(\ell)}] \tag{71}$$

(Luo et al., 2025a, Proposition 8, § 5).

Corollary 2.19 (Spectral-Norm Upper Bound of Feedforward Neural Network)**.** *For an ℓ-layer feedforward network f with only linear or convolutional layers and 1-Lipschitz activation functions, combining Proposition 2.16 and Proposition 2.18 immediately yields a spectral-norm product bound:*

$$\operatorname{Lip}[f] \leq \prod_{\ell=1}^{L} \|\theta^{(\ell)}\|_2, \tag{72}$$

where $\theta^{(\ell)}$ is the ℓ-th layer parameter matrix.

2.7.1 Lipschitz Continuity for a Directed Acyclic Graph (DAG)

To the best of our knowledge, the existing literature does not present an explicit Lipschitz bound for a general feedforward network with skip connections. Prior literature such as (Yoshida & Miyato, 2017; Virmaux & Scaman, 2018) focuses strictly on sequential architectures, while norm-based capacity measures such as the path norm (Neyshabur et al., 2015a) implicitly involve path enumeration but are not formulated as direct Lipschitz bounds. More recent path-metric bounds (Gonon et al., 2025) apply to arbitrary DAGs but are expressed in a rescaling-invariant metric form rather than the simple sum-over-paths structure we consider.

Figure 1 provides an illustrative example of a DAG network. To complement the survey, we derive an explicit bound in Theorem 2.20 using graph-theoretic method. To the best of our knowledge, this result has not previously appeared in the literature.

Theorem 2.20 (Lipschitz Bound for DAG Network)**.** *Let $G = (V, E)$ be a finite directed acyclic graph (DAG) representing a feedforward neural network with unique input node $s \in V$ and output node $t \in V$. Each node $v \in V$ is a module h_v that is Lipschitz continuous with constant $\operatorname{Lip}[h_v]$. An edge $(u \to v) \in E$ represents the computation of h_v immediately after h_u. Let $x_{v_i} : x \to x_{v_i}(x)$ represent the output at a node $v_i \in V$. The network is evaluated additively along incoming edges:*

$$x_s(x) = x, \tag{73}$$

$$x_v(x) = \sum_{(u\to v)\in E} h_v\big(x_u(x)\big), \quad v \neq s. \tag{74}$$

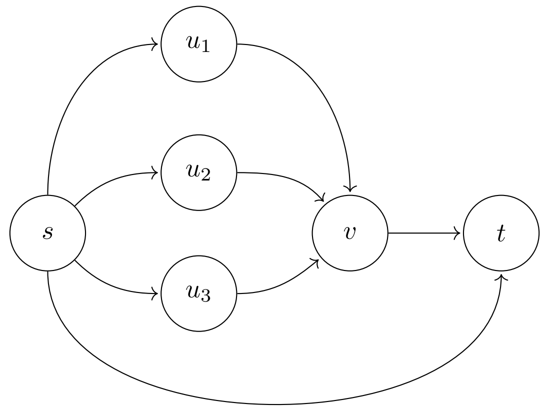

Figure 1: Example of Feedforward DAG Network. There are four computational paths: $s \to u_1 \to v \to t$, $s \to u_2 \to v \to t$, $s \to u_3 \to v \to t$, and $s \to t$. A node is a module in the DAG neural network f.

A computational path *p from s to t is an ordered sequence of nodes*

$$p = (v_0 = s, v_1, \ldots, v_{L_p} = t), \tag{75}$$

where $(v_{i-1} \to v_i) \in E$ for all $i = 1, \ldots, L_p$. Define for each edge $(u \to v)$ the constant

$$C_{(u \to v)} := \mathrm{Lip}[h_v], \tag{76}$$

and for a path p set

$$C_p := \prod_{i=1}^{L_p} \mathrm{Lip}[h_{v_i}] = \prod_{i=1}^{L_p} C_{(v_{i-1} \to v_i)}. \tag{77}$$

Let $\mathcal{P}$ be the set of all such paths from s to t. Then the Lipschitz constant of the overall network $f(x) := x_t(x)$ satisfies

$$\mathrm{Lip}[f] \;\le\; \sum_{p \in \mathcal{P}} C_p. \tag{78}$$

Proof. Fix any topological order of V. For $v \in V$, define the scalar

$$S(v) := \begin{cases} 1, & v = s, \\ \displaystyle\sum_{(u \to v) \in E} C_{(u \to v)}\, S(u), & v \neq s. \end{cases} \tag{79}$$

Claim: for all $v \in V$,

$$\mathrm{Lip}[x_v] \;\le\; S(v). \tag{80}$$

We prove equation 80 by induction along the topological order.

Base case ($v = s$). $x_s(x) = x$, hence $\mathrm{Lip}[x_s] = 1 = S(s)$.

Inductive step. Let $v \neq s$ and assume equation 80 holds for all predecessors u of v. Applying Minkowski inequality, for any x, y in the input space,

$$\|x_v(x) - x_v(y)\| = \left\| \sum_{(u\to v)} h_v\big(x_u(x)\big) - \sum_{(u\to v)} h_v\big(x_u(y)\big) \right\| \tag{81}$$

$$\leq \sum_{(u\to v)} \big\| h_v\big(x_u(x)\big) - h_v\big(x_u(y)\big) \big\| \tag{82}$$

$$\leq \sum_{(u\to v)} \mathrm{Lip}[h_v] \, \big\| x_u(x) - x_u(y) \big\| \tag{83}$$

$$\leq \sum_{(u\to v)} \mathrm{Lip}[h_v] \, \mathrm{Lip}[x_u] \, \|x - y\| \tag{84}$$

$$= \sum_{(u\to v)} C_{(u\to v)} \, \mathrm{Lip}[x_u] \, \|x - y\| \tag{85}$$

$$\leq \left(\sum_{(u\to v)} C_{(u\to v)} \, S(u) \right) \|x - y\| = S(v) \, \|x - y\|, \tag{86}$$

since inductive hypothesis:

$$\mathrm{Lip}[x_u] \leq S(u). \tag{87}$$

Thus $\mathrm{Lip}[x_v] \leq S(v)$. In particular,

$$\mathrm{Lip}[f] = \mathrm{Lip}[x_t] \;\leq\; S(t). \tag{88}$$

It remains to show $S(t) = \sum_{p\in\mathcal{P}} C_p$. More generally:

Lemma (Path expansion of S). For every $v \in V$,

$$S(v) = \sum_{p\in\mathcal{P}(s\to v)} \prod_{e\in p} C_e, \tag{89}$$

where $\mathcal{P}(s \to v)$ is the set of all directed paths from s to v.

Proof of Lemma. Proceed again by induction in topological order. For $v = s$, $\mathcal{P}(s \to s)$ contains only the empty path with product 1, so equation 89 holds. Assume equation 89 holds for all predecessors u of v. Then

$$S(v) = \sum_{(u\to v)} C_{(u\to v)} \, S(u) = \sum_{(u\to v)} C_{(u\to v)} \left(\sum_{p\in\mathcal{P}(s\to u)} \prod_{e\in p} C_e \right) \tag{90}$$

$$= \sum_{(u\to v)} \sum_{p\in\mathcal{P}(s\to u)} \left(\prod_{e\in p} C_e \right) C_{(u\to v)}. \tag{91}$$

Concatenating each $p \in \mathcal{P}(s \to u)$ with the edge $(u \to v)$ yields a bijection onto $\mathcal{P}(s \to v)$, and multiplies the edge constants accordingly; thus equation 89 holds for v. □

Applying the lemma at $v = t$ gives $S(t) = \sum_{p\in\mathcal{P}} C_p$, and hence

$$\mathrm{Lip}[f] \;\leq\; S(t) = \sum_{p\in\mathcal{P}} \prod_{i=1}^{L_p} \mathrm{Lip}[h_{v_i}]. \tag{92}$$

□

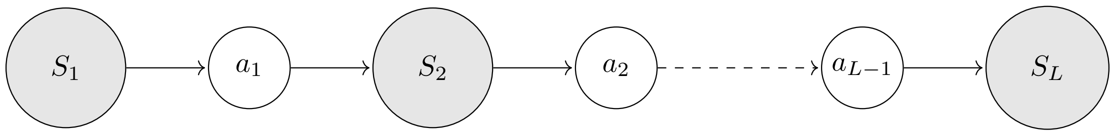

Figure 2: Topology of non-biconnected DAG network. If a DAG is separated into two sub-DAGs by the removal of a vertex a_i, then a_i is referred to as a *cut vertex* (or *articulation point*), and the DAG is said to be *non-biconnected*. This diagram shows the topology for a DAG that contains $L-1$ articulation points $a_1, a_2, \ldots, a_{L-1}$ and L corresponding sub-DAGs $S_1, S_2, \ldots, S_L$.

2.7.2 Lipschitz Constant of Non-Biconnected DAG

If a DAG is separated into two disconnected sub-DAGs by the removal of a vertex, then the removed vertex is referred to as *cut vertex* or *articulation point*. Modern deep learning architectures often give rise to computation DAGs that are not biconnected: the DAGs can be separated into multiple subgraphs by removing a small set of vertices.

Theorem 2.21 (Lipschitz Bound for Non-Biconnected DAG Network)**.** *Figure 2 shows the typical non-biconnected DAG topology found in modern deep learning architectures. The DAG contains a set of cut vertexes $a_1, a_2, \cdots, a_{L-1}$ and biconnected DAGs $S_1, S_2, \cdots, S_L$. Suppose the DAG network f in Figure 2 is:*

$$f = S_L \circ a_{L-1} \cdots a_2 \circ S_2 \circ a_1 \circ S_1, \tag{93}$$

then it is not difficult to show that the Lipschitz constant bound of this DAG is:

$$\mathrm{Lip}\,[f] \le \left(\prod_{i=1}^{L} \mathrm{Lip}\,[S_i]\right)\left(\prod_{j=1}^{L-1} \mathrm{Lip}\,[a_i]\right), \tag{94}$$

where each the Lipschitz bound of sub-DAG $\mathrm{Lip}\,[S_i]$ is given by Theorem 2.20 (Lipschitz Bound for DAG Network).

Remark 2.22. Theorem 2.21 (Lipschitz Bound for Non-Biconnected DAG Network) is particularly valuable for estimating the Lipschitz bound of a deep model. In contrast to path-based bound of Theorem 2.20 (Lipschitz Bound for DAG Network), which can be quite loose due to its combinatorial dependence on the number of paths, Theorem 2.21 yields significantly tighter and more structurally informed bounds.

2.7.3 Lipschitz Constant of Residual Network

Residual networks (He et al., 2016) address the optimization challenges of very deep neural architectures by introducing *residual modules*, which is illustrated in Figure 3. Each module m_{res} has the structure:

$$m_{\mathrm{res}}(x) = x + \phi(x), \tag{95}$$

which consists a non-identity unit ϕ and an identity skip connection. The resulting residual mapping $m_{\mathrm{res}}(x)$ maintains stable gradient flow, mitigates vanishing-gradient effects, and enables the training of substantially deeper networks without degradation. This simple architectural principle has proven highly effective and forms the backbone of many state-of-the-art deep learning models. Using Theorem 2.20 (Lipschitz Bound for DAG Network), it is not difficult to show that the Lipschitz constant bound of this structure is:

$$\mathrm{Lip}\,[m_{\mathrm{res}}] \le \mathrm{Lip}\,[s] + \mathrm{Lip}\,[\phi] = 1 + \mathrm{Lip}\,[\phi]\,, \tag{96}$$

which recovers the same results in (Behrmann et al., 2019b, Lemma 2, § 2), (Gouk et al., 2021, § 3.4) and Proposition 2.9 (Lipschitz Constant of Function Composition) by setting s to an identity map.

2.8 Complexity-Theoretic Generalization Bound

Notations. Let $\mathcal{H}$ be a hypothesis family and $\mathcal{H} \ni h : X \to Y$ be a hypothesis. Let $\ell : Y \times Y \to \mathbb{R}$ be a loss function. For each $\ell(h(x), y)$, we can associate ℓ and h with a map $G \ni g : X \times Y \to \mathbb{R}$, where G is the family of loss functions associated to with hypothesis family $\mathcal{H}$.

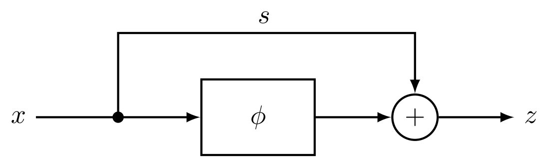

Figure 3: A residual module $m(x) = x + \phi(x)$ consists of a non-identity unit $\phi : x \mapsto \phi(x)$ and an identity skip connection unit $s : x \mapsto x$.

The generalization capacity of a neural network is fundamentally governed by its Lipschitz constant (Mohri et al., 2018). To quantify this capacity for a hypothesis, the concept of *generalization gap* is often used to measure the error between *true risk* and *empirical risk* (Definition 2.23). This gap for a model $h \in \mathcal{H}$ is proportional to its empirical Rademacher complexity of the hypothesis family $\mathcal{H}$, and the empirical Rademacher complexity is proportional to the Lipschitz constant $\text{Lip}[\mathcal{H}] := \sup_{h \in \mathcal{H}} \text{Lip}[h]$ of the hypothesis family:

$$\text{Generalization gap of } h \sim \text{Rademacher complexity} \sim O(\text{Lip}[\mathcal{H}]) \tag{97}$$

(Mohri et al. (2018, § 3); Shalev-Shwartz & Ben-David (2014, Part IV)). A smaller Lipschitz constant enforces smoother mappings, reducing the complexity of the function class and mitigating overfitting, thereby enhancing robustness to noise and adversarial perturbations (Neyshabur et al., 2015b; Gouk et al., 2021; Castin et al., 2024).

Definition 2.23 (Generalization Gap)**.** Let $S := \{(x_i, y_i)\}_{i=1}^m$ be a dataset where (x_i, y_i) is sampled *i.i.d.* from an unknown distribution $\mathcal{D}_X$. The true (population) risk for hypothesis h is defined as:

$$R(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}_X}[\ell(h(x), y)], \tag{98}$$

and the empirical risk on S is defined as:

$$R_S(h) := \frac{1}{m}\sum_{i=1}^{m} \ell(h(x_i), y_i) \tag{99}$$

(Mohri et al. (2018, § 3); Shalev-Shwartz & Ben-David (2014, Part IV)). Accordingly, the generalization gap on S is given by:

$$R(h) - R_S(h). \tag{100}$$

2.8.1 Rademacher Complexity and its Bounds

In learning theory, *Rademacher complexity* quantifies the expressiveness of a hypothesis family $G := \{g : X \to \mathbb{R}\}$ by measuring how well the G can fit data associated with random labels (Mohri et al., 2018, § 3.1). Let $S = (x_1, x_2, \ldots, x_m)$ be a dataset of size m with each $x_i \in X$ drawn i.i.d. from an unknown distribution $\mathcal{D}_X$. Let $h : X \to \mathbb{R}$ be a hypothesis that assigns a real-valued score to each input. Let $\sigma_1, \ldots, \sigma_m$ be m independent *Rademacher variables*, *i.e.*, random variables uniformly distributed over $\{-1, +1\}$ (Mohri et al., 2018, § 3.1). For random labels $\sigma_1, \ldots, \sigma_m$ assigned to S, the maximal correlation between the labels and the predictions of hypotheses in $\mathcal{H}$ is given by

$$\sup_{g \in G} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(x_i). \tag{101}$$

By assigning all possible random labels to the dataset S, we can thus define *Rademacher Complexity* for hypothesis family $\mathcal{G}$, stated in Definition 2.24 (Rademacher Complexity (Mohri et al., 2018, Definition 3.1, § 3)).

Definition 2.24 (Rademacher Complexity (Mohri et al., 2018, Definition 3.1, § 3)). The *empirical Rademacher complexity* of G with respect to S is defined as

$$\widehat{\mathfrak{R}}_S(G) := \mathbb{E}_\sigma \left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(x_i) \right], \tag{102}$$

and the (expected) *Rademacher complexity* is the expectation of $\widehat{\mathfrak{R}}_S(G)$ over all dataset S of size m:

$$\mathfrak{R}_m(G) := \mathbb{E}_{S \sim \mathcal{D}_X^m} \left[\widehat{\mathfrak{R}}_S(G) \right], \tag{103}$$

where $S \sim \mathcal{D}_X^m$ denotes m samples i.i.d. drawn from $\mathcal{D}$.

In the paradigm of machine learning, for the case of the hypothesis family $\ell \circ \mathcal{H}$ is the composition of a model $h \in \mathcal{H} := \{h : \mathbb{R}^n \to \mathbb{R}\}$ and loss function $\ell : \mathbb{R} \to \mathbb{R}$:

$$\ell \circ \mathcal{H} := \{\ell \circ h \mid h \in \mathcal{H}\}, \tag{104}$$

by Talagrand's Contraction Lemma (Mohri et al., 2018, Lemma 4.2, § 4), the empirical Rademacher complexity of $\ell \circ \mathcal{H}$ on S admits:

$$\widehat{\mathfrak{R}}_S(\ell \circ \mathcal{H}) \leq \mathrm{Lip}\,[\ell]\ \widehat{\mathfrak{R}}_S(\mathcal{H}). \tag{105}$$

For any $\delta > 0$, with probability at least $1 - \delta$, the generalization gap for $h \in \mathcal{H}$ on S is bounded by empirical Rademacher complexity of $\ell \circ \mathcal{H}$:

$$R(h) - R_S(h) \leq 2\,\widehat{\mathfrak{R}}_S(\ell \circ \mathcal{H}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \leq 2\,\mathrm{Lip}\,[\ell]\ \widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \tag{106}$$

(Mohri et al., 2018, Theorem 3.1 & Theorem 3.3, § 3). If $\mathcal{H}$ is a $K_{\mathcal{H}}$-Lipschitz continuous hypothesis family, that is:

$$\mathcal{H} := \{h : \mathbb{R}^n \to \mathbb{R} \mid \mathrm{Lip}\,[h] \leq K_{\mathcal{H}}\}, \tag{107}$$

and let:

$$\Phi := \{\phi : \mathbb{R}^n \to \mathbb{R}^n \mid \phi \text{ is a coordinate-wise identity map}\} \tag{108}$$

be an coordinate-wise identity hypothesis family. By using vector contraction theorem (Maurer, 2016, Equation 1 & Corollary 4), $\widehat{\mathfrak{R}}_S(\mathcal{H})$ holds an inequality:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) = \widehat{\mathfrak{R}}_S(\mathcal{H} \circ \Phi) \leq \sqrt{2} K_{\mathcal{H}}\,\widehat{\mathfrak{R}}_S(\Phi), \tag{109}$$

where $\widehat{\mathfrak{R}}_S(\Phi)$ is bounded by the 2-norm dataset diameter $\max_{x \in S} \|x\|_2$ of S:

$$\widehat{\mathfrak{R}}_S(\Phi) = \mathbb{E}_\sigma \left[\sup \frac{1}{m} \sum_{i=1}^{m} \sigma_i x_i \right] \leq \frac{\sqrt{\sup \sum_{i=1}^{m} \|x_i\|_2^2}}{m} = \frac{\max_{x \in S} \|x\|_2}{\sqrt{m}}, \tag{110}$$

this immediately yields:

$$R(h) - R_S(h) \leq O(K_{\mathcal{H}}). \tag{111}$$

The literature on norm-based generalization bounds (Bartlett et al., 2017; Golowich et al., 2018) also shows that the empirical Rademacher complexity of general neural networks with ReLU activations can be bounded by the product of the layerwise parameter 2-norms, which in turn provides an upper bound on the Lipschitz constant of the hypothesis family. For example, Bartlett et al. show that the lower bound of the empirical Rademacher complexity depends on an upper bound $K_{\mathcal{H}}^{+,\ell_2}$ of Lipschitz constant $K_{\mathcal{H}}$ of an L-layer network:

$$K_{\mathcal{H}} \leq K_{\mathcal{H}}^{+,\ell_2} := O\Big(\prod_{\ell=1}^{L} \|\theta^{(\ell)}\|_2\Big), \tag{112}$$

where $\|\theta^{(\ell)}\|_2$ is the 2-norm of the ℓ-th layer parameter matrix, this leads:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \geq O(K_{\mathcal{H}}^{+,\ell_2}), \tag{113}$$

(Bartlett et al., 2017, Theorem 3.4., § 3.3). The Lipschitz upper bound $K_{\mathcal{H}}^{+,\ell_2}$ is given by the product of the spectral-norms of parameter matrices (Proposition 2.18). Golowich et al. explicitly show that, for ReLU networks, the $\widehat{\mathfrak{R}}_S(\mathcal{H})$ is bounded by:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq O\Big(\frac{K_{\mathcal{H}}^{+,\ell_2}}{\sqrt{m}}\Big) \tag{114}$$

(Golowich et al., 2018, Theorem 2).

2.9 Training Dynamics of Lipschitz Continuity

The Lipschitz continuity of a parameterized network evolves as optimization updates its parameters. Consequently, the optimization process induces the dynamics of the network's Lipschitz continuity. While existing work in deep learning theory has largely focused on bounding the Lipschitz constants of neural networks, understanding how Lipschitz continuity evolves throughout training remains an open problem. Only a few studies have begun to explore this aspect. In particular, Luo et al. (2025a) introduces the concept of *optimization-induced dynamics* establishes a continuous-time stochastic framework that explicitly links the optimization process to the time-varying Lipschitz continuity bound of deep networks. The literature (Luo et al., 2025a) uses the operator-theoretical results regarding how the largest singular values of parameter matrices vary with respect to perturbations from the literature (Luo et al., 2025b), along with the theory from high-dimensional stochastic differential equations (SDEs) from stochastic analysis, for establishing a mathematical framework modeling the evolution of a spectral-norm based Lipschitz continuity upper bound. They further validate their theoretical framework with experiments across datasets and regularization scenarios.

Consider an L-layer feed-forward neural network $f : \mathbb{R}^m \to \mathbb{R}$ with 1-Lipschitz activation functions (*e.g.*, ReLU). Let $\boldsymbol{\theta}^{(\ell)}(t) \in \mathbb{R}^{m_\ell \times n_\ell}$ be the ℓ-layer parameter matrix at time t and $\boldsymbol{\theta}(t)$ be the collection of $\boldsymbol{\theta}^{(\ell)}(t)$ for all layers. Let $\mathcal{L}_f(\boldsymbol{\theta}^{(\ell)}(t))$ be the loss expectation for the f with parameters $\boldsymbol{\theta}(t)$ at time t. Suppose that $\boldsymbol{\Sigma}_t^{(\ell)} \in \mathbb{R}^{(m_\ell n_\ell)\times(m_\ell n_\ell)}$ is the layer-wise covariance matrix of stochastic gradient noise, arising from mini-batch sampling. Let $K^{(\ell)}(t)$ be the Lipschitz constant at layer ℓ. Let $K(t)$ be the network Lipschitz spectral-norm bound. Luo et al. (2025a) show that the continuous-time dynamics of the parameters under stochastic gradient descent (SGD) with a sufficiently small learning rate $\eta > 0$ are given by a system of SDEs (Luo et al., 2025a, Definition 10):

$$\begin{cases} \mathrm{dvec}(\boldsymbol{\theta}^{(\ell)}(t)) = -\mathrm{vec}\left[\nabla^{(\ell)}\mathcal{L}_f(\boldsymbol{\theta}(t))\right]\, \mathrm{d}t + \sqrt{\eta}\left[\boldsymbol{\Sigma}_t^{(\ell)}\right]^{\frac{1}{2}}\, \mathrm{d}\boldsymbol{B}_t^{(\ell)} \\ K^{(\ell)}(t) = \|\boldsymbol{\theta}^{(\ell)}(t)\|_{op} \\ Z(t) = \sum_{l=1}^{L} \log K^{(\ell)}(t) \\ K(t) = \mathrm{e}^{Z(t)} \end{cases}$$

where:

- vec : $\mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ represents major-column vectorization operator.

- $\|\cdot\|_{op}$ the matrix operator (spectral) norm.
- $\nabla^{(\ell)}\mathcal{L}_f(\boldsymbol{\theta}(t))$ is the gradient of the loss with respect to the ℓ-th layer parameters.
- $\boldsymbol{B}_t^{(\ell)}$ is a standard $(m_\ell n_\ell)$-dimensional Wiener process.

This stochastic dynamical system characterizes, in continuous time, the evolution of both layer-wise and network-level Lipschitz continuity bounds induced by the optimization process. Luo et al. (2025a, Theorem 15) show that the layer-wise dynamics decompose into three *driving forces*:

$$\frac{\mathrm{d}K^{(\ell)}(t)}{K^{(\ell)}(t)} = \left(\mu^{(\ell)}(t) + \kappa^{(\ell)}(t)\right)\,\mathrm{d}t + \boldsymbol{\lambda}^{(\ell)}(t)^\top\,\mathrm{d}\boldsymbol{B}_t^{(\ell)}, \tag{115}$$

where $\mathrm{d}\,\boldsymbol{B}_t^{(\ell)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{m_\ell n_\ell}\mathrm{d}t)$ represents the increment of a standard Wiener process in $\mathbb{R}^{m_\ell n_\ell}$, and the three *driving forces* are:

1. **Optimization-induced drift**:

$$\mu^{(\ell)}(t) = \frac{\left\langle \boldsymbol{J}_{op}^{(\ell)}(t), -\mathrm{vec}\left[\nabla^{(\ell)}\mathcal{L}_f(\boldsymbol{\theta}(t))\right]\right\rangle}{\sigma_1^{(\ell)}(t)}, \tag{116}$$

   where $\boldsymbol{J}_{op}^{(\ell)}(t)$ is the Jacobian of the operator norm with respect to $\mathrm{vec}(\boldsymbol{\theta}^{(\ell)}(t))$ and $\sigma_1^{(\ell)}(t)$ is its largest singular value. This term captures the deterministic component of Lipschitz evolution driven by the mean gradient flow. This shows that the alignment of the gradient $\nabla^{(\ell)}\mathcal{L}_f(\boldsymbol{\theta}(t))$ and principal direction of the ℓ-th layer parameter matrix $\boldsymbol{J}_{op}^{(\ell)}(t)$ at time t determines the contribution of the optimization to the increment of the layer-wise Lipschitz constant.

2. **Noise–curvature entropy production**:

$$\kappa^{(\ell)}(t) = \frac{\eta}{2\,\sigma_1^{(\ell)}(t)}\left\langle \boldsymbol{H}_{op}^{(\ell)}(t), \boldsymbol{\Sigma}_t^{(\ell)}\right\rangle \geq 0, \tag{117}$$

   where $\boldsymbol{H}_{op}^{(\ell)}(t)$ is the Hessian of the operator norm and $\boldsymbol{\Sigma}_t^{(\ell)}$ is the gradient-noise covariance. This non-negative term quantifies irreversible growth of the Lipschitz bound due to the interaction between stochastic gradient noise and curvature.

3. **Diffusion-modulation intensity**:

$$\boldsymbol{\lambda}^{(\ell)}(t) = \frac{\sqrt{\eta}}{\sigma_1^{(\ell)}(t)}\left[\boldsymbol{\Sigma}_t^{(\ell)}\right]^{1/2\top}\boldsymbol{J}_{op}^{(\ell)}(t), \tag{118}$$

   which controls the variance of Lipschitz evolution by scaling the Wiener noise term from mini-batch sampling.

Luo et al. (2025a, Theorem 16) show that, at the network level, these quantities aggregate as:

$$\mu_Z(t) = \sum_{\ell=1}^{L}\mu^{(\ell)}(t), \qquad \kappa_Z(t) = \sum_{\ell=1}^{L}\kappa^{(\ell)}(t), \qquad \lambda_Z(t) = \left[\sum_{\ell=1}^{L}\|\boldsymbol{\lambda}^{(\ell)}(t)\|_2^2\right]^{\frac{1}{2}}, \tag{119}$$

governing the deterministic trend, irreversible growth, and stochastic fluctuations of the overall Lipschitz bound $K(t)$.

Their framework is particularly useful for interpreting the behaviors of neural networks, such as the near-convergence behavior, noisy supervision and mini-batch sampling trajectories. Luo et al. (2025a, § 8) show that:

1. The Lipschitz constant bound irreversibly increase since the term *noise-curvature entropy production* $\boldsymbol{\kappa}_Z(t)$ is *non-negative*, which increases the system entropy (Luo et al., 2025a, § 8.3).
2. The magnitude of supervision noise affects the Lipschitz bound (Luo et al., 2025a, § 8.4 & 8.5). In particular, larger supervision noise leads to a lower Lipschitz bound for the network. This is because the supervision noise shrinks the *optimization-induced drift* $\boldsymbol{\mu}_Z(t)$.
3. Mini-batch trajectories do not affect the variance of the Lipschitz bound if batch size is sufficiently large (Luo et al., 2025a, § 8.6).

Remark 2.25. The dynamical analysis of Lipschitz continuity in deep learning models remains an open research problem. For instance, the dynamics under small-batch training remains poorly understood.

3 Estimation Methods

Proposition 2.18 gives an upper bound of the Lipschitz constant of a network (Miyato et al., 2018; Luo et al., 2025a). However, exact computation of the Lipschitz constant

$$K = \sup_{x \neq y} \frac{\|f(x) - f(y)\|_2}{\|x - y\|_2} = \sup_x \|\nabla f(x)\|_2 \tag{120}$$

is generally NP-hard (Virmaux & Scaman, 2018; Jordan & Dimakis, 2020). This section surveys major estimation and bounding techniques, ranging from statistical sampling to certified convex relaxations. We group them into *Power Iteration*, *Extreme Value Theory*, *Derivative Bound Propagation*, *Spectral Alignment*, *Convex Optimization Relaxations*, and *Exact/Relaxed MILP formulations*.

3.1 Power Iteration for Single Linear Unit

Power Iteration (Mises & Pollaczek-Geiringer, 1929) is also known as the Von Mises iteration (Mises & Pollaczek-Geiringer, 1929). It is the standard way to estimate the Lipschitz constant for a linear layer — fully-connected or convolutional, by approximating the spectral-norm of weight matrices, which bounds the layer Lipschitz constant (Proposition 2.16) Yoshida & Miyato (2017); Miyato et al. (2018); Sedghi et al. (2019); Kim et al. (2021); Luo et al. (2025a). Another desirable property of *power iteration* in optimization is that *power iteration* is differentiable. For example, Yoshida & Miyato (2017) and Miyato et al. (2018) use *power iteration* for computing the largest singular values of parameter matrices, in which the differentiability of *power iteration* allows the gradient propagation for gradient based optimization.

Suppose that $W \in \mathbb{R}^{m \times m}$ is the parameter matrix of a fully-connected layer or a convolutional layer, then the largest singular value $\sigma_1(W)$ is its operator norm:

$$\mathrm{Lip}[W] = \sigma_1(W). \tag{121}$$

Computing the exact value of $\sigma_1(W)$ for a large $m \times m$ matrix W is computationally expensive, as classical algorithms require $O(m^3)$ time. However, the largest singular value can be approximated using *power iteration* by assuming the existence of a uniquely dominant singular value (Golub & Van Loan, 2013, § 7.3). Starting with random unit vectors u_0, v_0, one step of power iteration updates

$$v_{k+1} \;\leftarrow\; \frac{W^\top u_k}{\|W^\top u_k\|_2}, \qquad u_{k+1} \;\leftarrow\; \frac{W v_k}{\|W v_k\|_2}, \tag{122}$$

and uses

$$\hat{\sigma}_1^{(k+1)} = u_{k+1}^\top W v_{k+1} \tag{123}$$

as a differentiable estimate of $\sigma_1(W)$ at step $k+1$. At the iteration number T, the estimated Lipschitz constant of W is approximated as:

$$\mathrm{Lip}[W] \approx u_T^\top W v_T, \tag{124}$$

where T is the number of iterations. The estimation error is proportional to:

$$|\sigma_1 - \hat{\sigma}_1^{(T)}| = O\Big(\Big[\frac{\sigma_2}{\sigma_1}\Big]^T\Big), \tag{125}$$

where σ_1 and σ_2 are the largest, and second largest singular values, respectively (Golub & Van Loan, 2013, Equation 7.3.5, § 7.3).

3.2 Extreme Value Theory

CLEVER (Weng et al., 2018b) — *Cross-Lipschitz Extreme Value for nEtwork Robustness* — converts the problem of estimating the attack-independent robustness into the problem of estimating the *local* Lipschitz constant by sampling gradient norms in a neighborhood of the input (Weng et al., 2018b). This is because the q-norm Lipschitz constant of a network $g : \mathbb{R}^m \to \mathbb{R}$ is determined by:

$$\mathrm{Lip}_q\,[g] = \sup_x \|\nabla\ g(x)\|_q, \tag{126}$$

for any inputs x and y:

$$|g(x) - g(y)| \leq \mathrm{Lip}_q\,[f]\,\|x - y\|_p \tag{127}$$

(Weng et al., 2018b, Lemma 3.1) where $1 \leq p, q \leq \infty$ satisfy the Hölder conjugacy relation (Rudin, 1976, § 6):

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{128}$$

Adversarial Perturbation Bound. Weng et al. (2018b) further show that the local Lipschitz constant can be used for deriving the adversarial perturbation bound. Let

$$f : \mathbb{R}^m \to \mathbb{R}^k \tag{129}$$

be a k-class classifier and f_i be the i-th prediction. Let:

$$c(x) = \underset{1\leq i\leq k}{\arg\max}\, f_i(x) \tag{130}$$

be the prediction. Then the p-norm adversarial perturbation $\|\delta\|_p$ to an input x_0 is bounded above by:

$$\|\delta\|_p \leq \min_{j\neq c} \frac{f_c(x_0) - f_j(x_0)}{\mathrm{Lip}_q\,[f_j]}, \tag{131}$$

where the local Lipschitz constant $\mathrm{Lip}_q\,[f_j]$ can be estimated through:

$$\mathrm{Lip}_q\,[f_j] = \max_{x\in B_p(x_0,\delta)} \|\nabla\ f_j(x)\|_q \tag{132}$$

and $B_p(x_0, \delta)$ is a p-norm ball centered at x_0 with a radius δ (Weng et al., 2018b, Theorem 3.2).

Distribution of Extreme Value $\max_{x\in B_p(x_0,\delta)} \|\nabla\ f_j(x)\|_q$**.** Estimating $\max_{x\in B_p(x_0,\delta)} \|\nabla_x\ f_j(x)\|_q$ can be through sampling $x \in B_p(x_0, \delta)$. Suppose the n samples are $\{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$. Their gradient q-norms are $\{\|\nabla\ f_j(x^{(1)})\|_q, \|\nabla\ f_j(x^{(2)})\|_q, \cdots, \|\nabla\ f_j(x^{(n)})\|_q\}$. Weng et al. (2018b) show that the extreme value of the gradient q-norm:

$$Y := \lim_{n\to\infty} \max_{x^{(i)}\in B_p(x_0,\delta)} \Big\{\|\nabla\ f_j(x^{(1)})\|_q, \|\nabla\ f_j(x^{(2)})\|_q, \cdots, \|\nabla\ f_j(x^{(n)})\|_q\Big\} \tag{133}$$

can only be a distribution P_Y of three distribution classes — *Gumbel class (Type I)*, *Fréchet class (Type II)* and *Reverse Weibull class (Type III)* — according to Fisher-Tippett-Gnedenko Theorem (Weng et al., 2018b, Theorem 4.1). The extreme value $\max \|\nabla\ f_j(x^{(1)})\|_q$ is thus converted to estimating the mean of the distribution P_Y:

$$\mathrm{Lip}_q\,[f_j] \approx \mathbb{E}_{Y\sim P_Y}\,[Y]\,. \tag{134}$$

3.3 Coordinate-Wise Gradient

Fast-Lip (Weng et al., 2018a) derives a lower bound of local Lipschitz constant in the form of coordinate-wise gradients by analyzing the activation patterns of ReLU networks. Let n_k denote the number of neurons at the k-th layer of an m-layer network and Let n_0 be the input dimension. Let $\phi_k : \mathbb{R}^{n_0} \to \mathbb{R}^{n_k}$ be the map from input to the output of k-th layer. Let σ be the coordinate-wise activation function. Then the relation between the $(k-1)$-th layer and the k-th layer can be written as:

$$\phi_k(x) = \sigma\Big(W^{(k)}\phi_{k-1}(x) + b^{(k)}\Big), \tag{135}$$

where $W^{(k)} \in \mathbb{R}^{n_k \times n_{k-1}}$ is the k-th layer parameter matrix, and $b^{(k)} \in \mathbb{R}^{n_k}$ is the bias. Set $f(x) = \phi_m(x)$ and let $f_j(x)$ denote the j-th output of f. Weng et al. (2018a) start from analyzing the ReLU activation patterns and then deriving gradient bounds under this setting for bounding local Lipschitz constant.

Activation Patterns of ReLU Networks. Let $l_r^{(k)}$ and $u_r^{(k)}$ denote the lower and upper bound for the r-th neuron in the k-th layer and let $z_r^{(k)}$ be the pre-activation at the k-th layer, given as:

$$z_r^{(k)} = W_{r,:}^{(k)}\phi_{k-1}(x) + b_r^{(k)}, \tag{136}$$

where $W_{r,:}^{(k)}$ denotes the r-th row of $W^{(k)}$ (Weng et al., 2018a, § 3.2). For the neurons indexed by $[n_k] := \{1, 2, \cdots, n_k\}$, then there are only three activation patterns:

1. **Always Activated.** Neurons are always activated: $\mathcal{I}_k^+ := \{r \in [n_k] \mid u_r^{(k)} \geq l_r^{(k)} \geq 0\}$.
2. **Always Inactivated.** Neurons are always inactivated: $\mathcal{I}_k^- := \{r \in [n_k] \mid l_r^{(k)} \leq u_r^{(k)} \leq 0\}$.
3. **Either Activated or Inactivated.** Neurons are either activated or inactivated: $\mathcal{I}_k := \{r \in [n_k] \mid l_r^{(k)} \leq 0 \leq u_r^{(k)}\}$.

Gradient Analysis of ReLU Networks. Gradient norm carries the information for local Lipschitz constant. Weng et al. (2018a) then analyze the coordinate-wise gradients of ReLU networks with a manner of layer-wise. Using the activation patterns of ReLU networks, the k-th layer output can be rewritten into:

$$\phi_k(x) = \Lambda^{(k)}\Big(W^{(k)}\phi_{k-1}(x) + b^{(k)}\Big), \tag{137}$$

where $\Lambda^{(k)}$ is the activation pattern matrix:

$$\Lambda_{r,r}^{(k)} = \begin{cases} 1 \text{ or } 0, & \text{if } r \in \mathcal{I}_k \\ 1, & \text{if } r \in \mathcal{I}_k^+ \\ 0, & \text{if } r \in \mathcal{I}_k^- \end{cases}. \tag{138}$$

Let $\Lambda_a^{(k)}$ be the diagonal activation matrix for neurons in the k-th layer that are always activated and set $\Lambda_u^{(k)} := \Lambda^{(k)} - \Lambda_a^{(k)}$. Starting by analyzing the bound of the gradient $\nabla\phi_k(x)$ coordinate-wisely in a 2-layer ReLU network, Weng et al. (2018a) show that an inequality for gradient q-norm holds:

$$\max_{x \in B_p(x_0,\epsilon)} \Big|[\nabla f_j(x)]_k\Big| \leq \max\Big(C_{j,k}^{(1)} + L_{j,k}^{(1)},\ C_{j,k}^{(1)} + U_{j,k}^{(1)}\Big), \tag{139}$$

where:

$$C_{j,k}^{(1)} = W_{j,:}^{(2)}\Lambda_a^{(1)}W_{:,k}^{(1)}, \quad L_{j,k}^{(1)} = \sum_{i \in \mathcal{I}_1, W_{j,i}^{(2)}W_{i,k}^{(2)} < 0} W_{j,i}^{(2)}W_{i,k}^{(2)}, \quad \text{and} \quad U_{j,k}^{(1)} = \sum_{i \in \mathcal{I}_1, W_{j,i}^{(2)}W_{i,k}^{(2)} > 0} W_{j,i}^{(2)}W_{i,k}^{(2)}. \tag{140}$$

Then the gradient q-norm can be bounded above by:

$$\mathrm{Lip}_q\left[f_j; B_p(x_0,\epsilon)\right] = \max_{x \in B_p(x_0,\epsilon)} \|\nabla f_j(x)\|_q \leq \left[\sum_k \Big(\max_{x \in B_p(x_0,\epsilon)} \|\nabla f_j(x)\|_k\Big)\right]^{\frac{1}{q}}, \tag{141}$$

yielding a local Lipschitz constant upper bound. This method is referred to as **Fast-Lip** in Weng et al. (2018a).

3.4 Spectral Alignment

SeqLip (Virmaux & Scaman, 2018) refines the upper bound of an L-layer sequential network f with 1-Lipschitz activation functions:

$$\mathrm{Lip}\,[f]^{+} = \prod_{\ell=1}^{L} \|W^{(\ell)}\|_2, \tag{142}$$

where $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell+1}}$ is the ℓ-th layer parameter matrix, and $\mathrm{Lip}\,[f]^{+}$ is referred to as **AutoLip** bound in Virmaux & Scaman (2018). By taking into account the spectral alignment between the parameter matrices of two consecutive layers, the bound is then refined as:

$$\mathrm{Lip}\,[f] \leq \mathrm{Lip}\,[f]^{+} \underbrace{\left(\prod_{\ell}^{L-1} \sqrt{(1 - r_\ell - r_{\ell+1}) \max_{t^{(\ell)} \in [0,1]^{n_\ell}} \langle t^{(\ell)} \cdot v_1^{(\ell+1)},\ u_1^{(\ell)} \rangle^2 + r_\ell + r_{\ell+1} + r_\ell r_{\ell+1}}\right)}_{\text{spectral alignment}}, \tag{143}$$

where $t^{(\ell)}[0,1]^{n_\ell}$ represents the ℓ-th layer gradient patterns from activation functions for n_ℓ neurons — the coordinate-wise gradient of an activation function such as ReLU falls in $[0,1]$, $u_1^{(\ell)}$ is the first left-singular vector for $W^{(\ell)}$, $v_1^{(\ell+1)}$ is the first right-singular vector for $W^{(\ell+1)}$, and $r_\ell = \frac{\sigma_2^{(\ell)}}{\sigma_1^{(\ell)}}$ is the ratio of the second largest singular value to the first largest singular value for $W^{(\ell+1)}$ (Virmaux & Scaman, 2018, Theorem 3).

Virmaux & Scaman (2018) also show that, if $u, v\mathbb{R}^n$ are two independent random vectors taken uniformly from $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$, the following limit:

$$\lim_{n\to\infty} \max_{t\in[0,1]^n} \Big| \langle t \cdot u\ , v \rangle \Big| = \frac{1}{\pi} \tag{144}$$

holds *almost surely*. Under this independent assumption, the bound reduces to:

$$\mathrm{Lip}\,[f] \leq \frac{\mathrm{Lip}\,[f]^{+}}{\pi^{L-1}} \tag{145}$$

(Virmaux & Scaman, 2018, Lemma 2). In real setting, the singular vectors are not independent across layers.

3.5 Convex Optimization Relaxation

LipSDP (Fazlyab et al., 2019) interpret activation functions as gradients of convex potential functions, satisfying certain properties described by quadratic constraints. Therefore, Lipschitz constant estimation problem is treated as a semidefinite program (SDP) problem.

Let $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ be an L-layer feed-forward network, recursively defined as:

$$\begin{cases} x^{(0)} = x \\ x^{(k+1)} = \phi\Big(W^{(k)x^{(k)}} + b^{(k)}\Big) \end{cases}, \tag{146}$$

where $x^{(\ell)}$ is the output at the ℓ-th layer, $W^{(\ell)} \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$ is the parameter matrix at the ℓ-th layer, $b^{(\ell)} \in \mathbb{R}^{n_{\ell+1}}$ is the bias at the ℓ-th layer, and ϕ is the coordinate-wise activation function. For a single layer network:

$$f(x) = W^{(1)} \phi\big(W^{(0)} x + b^{(0)} + b^{(1)}\big), \tag{147}$$

Fazlyab et al. (2019) shows that the Lipschitz constant of f is given by a SDP problem:

$$\mathrm{Lip}\,[f] \leq \sqrt{\rho}, \tag{148}$$

defined by:

$$\begin{aligned} &\text{minimize} && \rho && (149)\\ &\text{subject to} && M(\rho, T) \preceq 0 \quad T \in T_n, && (150)\end{aligned}$$

where ϕ is *slope-restricted* on $[\alpha, \beta]$:

$$\alpha \leq \frac{\phi_j(x) - \phi_j(y)}{x - y} \leq \beta \qquad \forall x, y \in \mathbb{R} \tag{151}$$

and:

$$M(\rho, T) := \begin{bmatrix} -2\alpha\beta(W^{(0)})^\top T W^{(0)} & (\alpha+\beta)(W^{(0)})^\top \\ (\alpha+\beta)TW^{(0)} & -2T + (W^{(1)})^\top W^{(1)} \end{bmatrix} \preceq 0 \tag{152}$$

holds true for $T \in T_n$ (Fazlyab et al., 2019, Theorem 1). For more general L-layer fully-connected network, Fazlyab et al. (2019) show that $M(\rho, T)$ is:

$$M(\rho, T) = \begin{bmatrix} A \\ B \end{bmatrix}^\top \begin{bmatrix} -2\alpha\beta T & (\alpha+\beta)T \\ (\alpha+\beta)T & -2T \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} + \begin{bmatrix} -\rho I_{n_0} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (W^{(L)})^\top W^{(L)} \end{bmatrix} \preceq 0, \tag{153}$$

where:

$$A = \begin{bmatrix} W^{(0)} & 0 & \cdots & 0 & 0 \\ 0 & W^{(1)} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & W^{(L-1)} & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & I_{n_1} & 0 & \cdots & 0 \\ 0 & 0 & I_{n_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & I_{n_L} \end{bmatrix}, \tag{154}$$

and:

$$C = \begin{bmatrix} 0 & \cdots & 0 & W^{(L)} \end{bmatrix}, \qquad b = \begin{bmatrix} (b^{(0)})^\top \cdots (b^{(L-1)})^\top \end{bmatrix}^\top \tag{155}$$

(Fazlyab et al., 2019, Theorem 2).

3.6 Integer Programming for ReLU Networks

It is well known that computing the Lipschitz constant of an arbitrary scalar- or vector-valued function is NP-hard (Virmaux & Scaman, 2018; Jordan & Dimakis, 2020). A common approximation is to estimate the Lipschitz constant using gradient norms, where the Jacobians can be explicitly derived via the chain rule. However, nondifferentiabilities, *e.g.*, those arising in ReLU networks, can introduce inaccuracies into these estimates. An alternative approach is to formulate the Lipschitz bounding problem as an integer programming problem, which avoids such inaccuracies by directly bounding the Jacobians.

LipMIP (Jordan & Dimakis, 2020) is a method that provably exactly compute 2-norm and ∞-norm Lipschitz constants of non-smooth ReLU networks. We summarize their method by simplifying their notations. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a ReLU network. Let σ be activation function. Then, an L-layer ReLU network is recursively defined as:

$$\begin{cases} f(x) = \sigma(Z_L(x)) \\ Z^{(\ell)}(x) = W^{(\ell)}\, \sigma\Big(Z^{(\ell-1)}(x)\Big) + b^{(\ell)} \\ Z^{(0)} = x \end{cases}, \tag{156}$$

where $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ is the ℓ-th layer parameter matrix, and $Z^{(\ell)}(x)$ is the output of the ℓ-th layer neurons. Let $\partial f(x)$ be the Clarke sub-differential convex hull of f at point x — see Section 2.4 (Sub-Differential Convex Functions). For example:

$$\partial\sigma(0) = [0, 1]. \tag{157}$$

Let $\partial f(X)$:

$$\partial f(X) = \Big\{ g \mid \; g \in \partial f(x), x \in X \Big\} \tag{158}$$

be the set-valued Clarke sub-differentials at set X. Let $\nabla^\sharp f(\bullet)$ be the pushforward set that the sub-gradients of ReLU activation at value 0 come from the hull set $\partial\sigma(0)$:

$$\nabla^\sharp f(\bullet) = \Big\{ \nabla f(\bullet) \mid \text{the sub-differentials of activations come from the hull set } \partial\sigma(0) \Big\}. \tag{159}$$

Thus for a ReLU network, the *feasible set* for its gradients on X are given as:

$$\nabla^\sharp f(X) = \Big\{ G \in \nabla^\sharp f(x) \mid x \in X \Big\}. \tag{160}$$

Jordan & Dimakis (2020) show that the push forward set sub-gradient $\nabla^\sharp f(x)$ is the Clarke sub-gradient hull $\partial f(x)$, that is:

$$\nabla^\sharp f(x) = \partial f(x) \tag{161}$$

(Jordan & Dimakis, 2020, Theorem 2). Then the local $p \to q$ Lipschitz constant of f on X is given as:

$$\mathrm{Lip}_{p\to q}\left[f; X\right] = \sup_{G\in\nabla^\sharp f(X)} \sup_{\|v\|_p \le 1} \frac{\|G^\top v\|_q}{\|v\|_p} = \sup_{G\in\nabla^\sharp f(X)} \|G^\top\|_q. \tag{162}$$

Finding the Lipschitz constant on X can be formulated as an integer programming problem on the feasible set $\nabla^\sharp f(X)$. Let $n = \sum_{\ell=1}^{L} n_\ell$ be the number of neurons in f. The set:

$$\Big\{ \|G^\top\|_q \mid G \in \nabla^\sharp f(X) \Big\} \tag{163}$$

is then represented as a mixed-integer programming problem for maximizing $\|G^\top\|_q$ by solving the combinations of input $x \in X$ and ReLU's sub-gradient $a \in [0,1]^n$ on a mixed-integer polytope in $X \times [0,1]^n$, such that $\nabla^\sharp f(x)$ is represented with some A, B and c:

$$\nabla^\sharp f(x) = Ax + Ba < c \tag{164}$$

by analyzing gradients with chain rules on ReLU networks (Jordan & Dimakis, 2020, Definition 6 & Lemma 1). This problem can be solved by off-the-shelf MIP solvers.

4 Regularization Approaches

A variety of techniques have been developed to explicitly or implicitly enforce Lipschitz continuity in deep neural networks. These approaches can be categorized into:

(i) Section 4.1: **Weight Regularization** — constraining or regularizing weight matrices during initialization and training;

(ii) Section 4.2: **Gradient Regularization** — penalizing or normalizing gradient norms during training;

(iii) Section 4.3: **Activation Regularization** — designing or constraining activation functions to be Lipschitz-bounded;

(iv) Section 4.4: **Class-Margin Regularization** — implicitly enforcing Lipschitz constraints via maximizing class decision boundaries;

(v) Section 4.5: **Architectural Regularization** — enforcing Lipschitz constraints by designing architectures, particularly for transformers.

4.1 Weight Regularization

This subsection reviews methods that control the Lipschitz constant by directly constraining or regularizing the network's weight parameters, either at initialization or during training.

4.1.1 Spectral-Norm Regularization

Spectral-norm regularization constrains the Lipschitz constant of a neural network by bounding the spectral-norm of its weight matrices, which is the largest singular value of the matrix (Proposition 2.16). Yoshida & Miyato (2017) propose normalizing the weight matrix W of a linear layer by:

$$W \leftarrow \frac{W}{\|W\|_2}, \tag{165}$$

ensuring a Lipschitz constant of at most 1 for the layer (Yoshida & Miyato, 2017, § 3). This method, later adopted by Miyato et al. in generative adversarial networks (GANs) (Goodfellow et al., 2020), for improving the generations (Miyato et al., 2018, § 2.1).

4.1.2 Weight Clipping

Weight clipping is a straightforward method to enforce Lipschitz continuity by limiting the magnitude of weight matrices. This is because for a matrix W, the operator-norm variation $\|W\|_{op}$ in ℓ_2 for W under perturbation ΔW is bounded above by:

$$\|W\|_{op} = \sigma_1(W) = \langle u_1 v_1^\top,\ \Delta W\rangle \leq \|u_1 v_1^\top\|_2\, \|\Delta W\|_2, \tag{166}$$

with Cauchy–Schwarz inequality, where $\sigma_1(W)$ is the largest singular value of W, u_1, v_1 are the left- and right-singular vectors corresponding the largest singular value (Luo et al., 2025b, Lemma 5.1).

In the setting of GANs, Arjovsky et al. (2017) show that the discriminator f of a GAN evaluates the Wasserstein distance $W(\mathbb{P}, \mathbb{Q})$ between two distributions $\mathbb{P}$ and $\mathbb{Q}$. The Kantorovich-Rubinstein duality (Villani et al., 2008) tells that:

$$W(\mathbb{P}, \mathbb{Q}) = K \sup_{\mathrm{Lip}[f] \leq K} \mathbb{E}_{x\sim\mathbb{P}}\big[f(x)\big] - \mathbb{E}_{y\sim\mathbb{Q}}\big[f(y)\big] \tag{167}$$

(Arjovsky et al., 2017, Equation 2). Therefore reducing the Lipschitz constant K can reduce the Wasserstein distance $W(\mathbb{P}, \mathbb{Q})$. Arjovsky et al. (2017) propose clipping weights to a fixed range:

$$W \leftarrow \mathrm{clip}(W, -c, +c) \tag{168}$$

(Arjovsky et al., 2017, Algorithm 1), where c is a small positive constant (*e.g.*, $c = 0.01$). Weight clipping is computationally efficient, but can lead to exploding or vanishing gradients if c is too small, reducing model capacity, as criticized in the a later literature (Gulrajani et al., 2017).

4.1.3 Orthogonal Weight

Parseval Networks (Cisse et al., 2017). Parseval networks enforce approximate orthonormality on weight matrices to bound their spectral-norms (Cisse et al., 2017). For a weight matrix $W^{(\ell)}$ in layer ℓ, Cisse et al. (2017) minimize the 2-norm of the deviation from orthonormality:

$$\|W^{(\ell)\top} W^{(\ell)} - \mathrm{I}\|_F, \tag{169}$$

ensuring $\|W^{(\ell)}\|_2 \leq 1$ (Cisse et al., 2017, § 4.2). For a feedforward network with L layers and 1-Lipschitz activations (*e.g.*, ReLU), the global Lipschitz constant is bounded by:

$$K \leq \prod_{\ell=1}^{L} \|W^{(\ell)}\|_2 \leq 1 \tag{170}$$

(Proposition 2.18). Cisse et al. (2017) argue that the robustness to adversarial attacks is enhanced since the reduced sensitivity to perturbations.

4.2 Gradient Regularization

Gradient-based regularization methods enforce Lipschitz continuity by constraining the gradient norm of the loss function or network output with respect to inputs, ensuring smooth decision boundaries and robustness. Note that, for a network f on convex X, its 2-norm Lipschitz constant on X is given by:

$$\mathrm{Lip}\,[f] = \sup_{x \in X} \|\nabla f(x)\|_2, \tag{171}$$

see 2.5 (Lipschitz Constant Bounds Gradient Norm). Theoretically, constraining $\|\nabla f(x)\|_2$ can bound Lipschitz constant of f.

For example, Gulrajani et al. (2017) argue that bounding Lipschitz constant by weight clipping can lead exploding or vanishing gradients, and reduce capacity of the critic in a GAN. They then introduce a regularization method by directly bounding the gradients of the critic network D through:

$$\mathbb{E}_{x \sim \mathbb{P}_X}\left[\left(\|\nabla D(x)\|_2 - 1\right)^2\right], \tag{172}$$

where $\mathbb{P}_X$ is the distribution of training sample set X (Gulrajani et al., 2017, § 4).

4.3 Activation Regularization

Activation regularization methods modulates Lipschitz continuity by constraining the properties of activation functions or their outputs. Such approaches may impose explicit norm bounds, design activations that are inherently norm-preserving, or apply penalties to activation magnitudes, thereby limiting the contribution of nonlinearities to the overall Lipschitz constant of the network.

4.3.1 Group-Sorting Activation

Bounding the Lipschitz constant of a network by ensuring 1-Lipschitz for affine units through spectral-norm constraints can improve robustness (Yoshida & Miyato, 2017; Miyato et al., 2018). However, this often comes at the cost of reduced expressiveness (Anil et al., 2019). Anil et al. (2019) first show that, to remain expressive in a spectral-norm constrained network, the network must preserve the gradient norms at each layer (Anil et al., 2019). ReLU networks can only satisfy this for positive values, this is because:

$$\frac{\partial\, \mathrm{ReLU}(x)}{\partial\, x} = \begin{cases} 1, & x > 0 \\ (0,1), & \text{in sub-differential sense}\ . \\ 0, & x < 0 \end{cases} \tag{173}$$

To preserve the gradient norms at activation layer, Anil et al. (2019) use a general purpose 1-Lipschitz activation function $\mathbf{GroupSort}(x)$ which is homogeneous:

$$\mathbf{GroupSort}(\alpha\, x) = \alpha\, \mathbf{GroupSort}(x) \tag{174}$$

(Anil et al., 2019). **GroupSort** activation separates a pre-activation x into k groups:

$$x = (\underbrace{x_1, \cdots, x_{g_1}}_{\text{group 1}}, \underbrace{x_{g_1+1}, \cdots, x_{g_2}}_{\text{group 2}}, \cdots \underbrace{x_{g_{k-1}+1}, \cdots, x_{g_k}}_{\text{group k}}), \tag{175}$$

and sorts each group into ascending order:

$$\mathbf{GroupSort}(x) = (\underbrace{x_{s_1}, \cdots, x_{s_{g_1}}}_{\text{group 1}}, \underbrace{x_{s_{g_1}+1}, \cdots, x_{s_{g_2}}}_{\text{group 2}}, \cdots \underbrace{x_{s_{g_{k-1}}+1}, \cdots, x_{s_{g_k}}}_{\text{group k}}), \tag{176}$$

where the i-th sorted group satisfies:

$$x_{s_{g_{i-1}}+1} \leq x_{s_{g_{k-1}}+2} \leq \cdots \leq x_{s_{g_k}} \tag{177}$$

(Anil et al., 2019). For the pre-activation in a linear unit:

$$Wx \tag{178}$$

where $W \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$, if the linear unit is with spectral-norm constraint such that:

$$\|W\|_2 = 1, \tag{179}$$

GroupSort activation directly gives rise an 1-Lipschitz constant activation:

$$\sup_{\|x\|_2=1} \|\textbf{GroupSort}(Wx)\|_2 = \sup_{\|x\|_2=1} \|Wx\|_2 = \|W\|_2. \tag{180}$$

When the group size is 2, Anil et al. (2019) refers to this special case as **MaxMin**, which is equivalent to the Orthogonal Permutation Linear Unit (OPLU) (Chernodub & Nowicki, 2017); when the group size is the entire input, this is referred to as **FullSort** in the literature (Chernodub & Nowicki, 2017).

To train a network with **GroupSort** activations and guarantee that all linear units are exactly 1-Lipschitz during optimization instead of bounded by 1 (Cisse et al., 2017; Gulrajani et al., 2017), Chernodub & Nowicki (2017) approximate the updated parameter matrix W_0 with a closest orthogonal matrix through a differentiable, iterative algorithm, referred to as *Björck Orthogonalization* (Björck & Bowie, 1971). This algorithm is defined by:

$$W_{k+1} = W_k\Big(I + \frac{1}{2}Q_k + \cdots + (-1)^p \binom{-\frac{1}{2}}{p} Q_k^p\Big), \tag{181}$$

where W_k is the k-th iterative result, p is a chosen hyper-parameter and Q_k is:

$$Q_k = I - W_k^\top W_k \tag{182}$$

(Chernodub & Nowicki, 2017, § 4.2.1). As a result, the iteration at step T leads to:

$$W_T^\top W_T \approx I, \tag{183}$$

which provably ensures the linear unit with parameter matrix W is 1-Lipschitz continuous:

$$\|W_T\|_2 \approx 1. \tag{184}$$

4.3.2 Contraction Activation and Invertible Residual Map

Flow-based models or flows learn to transform a source distribution p_X to target distribution p_Z, consisting of invertible networks (Rezende & Mohamed, 2015). Ensuring that the transformation is invertible and its Jacobian are computable is a straightforward way for ensuring invertibility:

$$\log p_X(x) = \log p_Z(z) + \log|\det J_F(x)| \tag{185}$$

where $F : \mathbb{R}^m \to \mathbb{R}^m$ is an invertible map and $J_F(x)$ is the Jacobian of F at x (Berg et al., 2018; Dinh et al., 2017). Different from explicit computation of Jacobian, Behrmann et al. (2019a) propose a method for inverting residual networks (He et al., 2016) by ensuring the Lipschitz constants be less than 1, so that the residual networks are *contraction maps* (Behrmann et al., 2019a; Perugachi-Diaz et al., 2021).

Inverting Residual Layer. Let

$$F(x) = x + g(x) \tag{186}$$

be a residual layer where $F : \mathbb{R}^m \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}^m$ and $x \in \mathbb{R}^m$. Let g be a contraction map:

$$\mathrm{Lip}\,[g] < 1. \tag{187}$$

Suppose that the inversion F^{-1} exists and set $y = F(x)$, so that:

$$F^{-1}(x) = x = y - g(x). \tag{188}$$

Set:

$$T(x) = y - g(x), \tag{189}$$

note that:

$$\mathrm{Lip}\,[T] = \mathrm{Lip}\,[g] < 1 \tag{190}$$

is a contraction mapping. Using *Banach contraction principle* or *Banach fixed-point theorem*, set $x_0 = y$, the $F^{-1}(y)$ can be approximated through:

$$x_{k+1} = y - g(x_k) \tag{191}$$

iteratively. The **essential condition** for inverting g is that g must be a contraction map:

$$\mathrm{Lip}\,[g] < 1. \tag{192}$$

Contraction Activation. Chen et al. (2019) first introduce **LipSwish** activation function for inverting residual networks, defined as:

$$\mathbf{LipSwish}(x) = \frac{x\sigma(\beta\, x)}{1.1}, \tag{193}$$

where σ is the sigmoid function, and $\beta > 0$ is a learnable negative constant, initialized with 0.5. **LipSwish** has a Lipschitz constant at most 1. However, the negative axis of **LipSwish** has zero gradients almost everywhere. In a 1-Lipschitz continuous network, the Jacobian norm across two consecutive layers is reduced to be at most one, which limits the network's expressiveness and referred to as *gradient norm attenuation* problem (Anil et al., 2019; Li et al., 2019). To mitigate this problem, Perugachi-Diaz et al. (2021) use specially designed activation function **CLipSwish** by concatenating two **LipSwish** functions. The **CLipSwish** is therefore given as:

$$\Phi(x) = \begin{bmatrix} \mathbf{LipSwish}(x) \\ \mathbf{LipSwish}(-x) \end{bmatrix}, \qquad \mathbf{CLipSwish}(x) = \frac{\Phi(x)}{\mathrm{Lip}\,[\Phi]} \leq 1, \tag{194}$$

where $\mathrm{Lip}\,[\Phi] \approx 1.004$ (Perugachi-Diaz et al., 2021, § 3.4). This concatenation overcomes the gradient attenuation problem introduced by **LipSwish** since the negative axis has zero gradients almost everywhere while ensuring the activation is a contraction map.

4.4 Class-Margin Regularization

The goal of adversarial defense in classification task is to ensure that, for a k-class classifier $f : \mathbb{R}^d \to \mathbb{R}^k$, and an input $x \in \mathbb{R}^d$, a bounded perturbation $\|\epsilon_x\|_2 < c$ does not alter the prediction:

$$\underset{i\in[k]=\{1,2,\cdots,k\}}{\arg\max} \quad f_i(x + \epsilon_x), \tag{195}$$

where $f_i(x)$ denotes the output at the i-th coordinate. The margin $m(x)$ is defined as:

$$m(x) = f_i(x) - \max_{j\neq i} f_j(x), \tag{196}$$

where i is the ground-truth (Tsuzuku et al., 2018, § 4.1). Then, a guaranteed 2-norm perturbation bound does not alter the prediction, if:

$$\|\epsilon_x\|_2 \leq \frac{m(x)}{\sqrt{2}\,\mathrm{Lip}\,[f]} \tag{197}$$

(Szegedy et al., 2014; Cisse et al., 2017). This implies that the class margin must be at least:

$$m(x) = f_i(x) - \max_{j \neq i} f_j(x) \geq \sqrt{2}\,\mathrm{Lip}\,[f] \tag{198}$$

(Tsuzuku et al., 2018, Proposition 1 & 2). Suppose i is the ground-truth for x, in training stage, Perugachi-Diaz et al. (2021) adjust each class margin $j \neq i$ by:

$$f_j(x) \leftarrow f_j(x) + \sqrt{2}\,c\,\mathrm{Lip}\,[f]\,, \tag{199}$$

where c is a guaranteed perturbation bound (Perugachi-Diaz et al., 2021, Algorithm 1). For each linear unit $\phi : \mathbb{R}^m \to \mathbb{R}^n$ in the network, and let $u \in \mathbb{R}^m$ be drawn i.i.d. from $\mathcal{N}(0, 1)$. Perugachi-Diaz et al. (2021) use a general method for estimating the Lipschitz constant of ϕ iteratively by:

$$u_k \leftarrow \frac{u_k}{\|u_k\|_2}, \qquad v_k \leftarrow \phi(u_k), \qquad u_{k+1} \leftarrow \frac{1}{2}\frac{\partial\,\|v_k\|_2^2}{\partial\,u}. \tag{200}$$

At the end of the iteration T, the Lipschitz constant of ϕ is given by:

$$\mathrm{Lip}\,[\phi] \approx \|u_T\|_2, \tag{201}$$

almost surely (Perugachi-Diaz et al., 2021).

4.5 Architectural Regularization

Lipschitz continuity can also be constrained through architectural design choices, including activation functions, parameter structures, initialization schemes, and optimization. As a complement to regularization approaches that do not fall into one single category, we survey these approaches in this section.

4.5.1 Orthogonalizing Convolution

Parameter orthogonality in neural networks can directly yield 1-Lipschitz continuous (Proposition 2.8) (Lipschitz Constant of Semi-Orthogonal Matrix)). This is because, for a matrix $W \in \mathbb{R}^{m \times n}$, an semi-orthogonal matrix W preserves the 2-norm:

$$W^\top W = I_n \quad \text{or} \quad WW^\top = I_m \quad \Longrightarrow \quad \mathrm{Lip}\,[W] = 1, \tag{202}$$

if $m \geq n$ and $W^\top W = I_n$, W is an isometric map; if $m < n$ and $WW^\top = I_m$, W is a contraction map on the kernel $\ker(W)$ strictly (Proposition 2.8) (Lipschitz Constant of Semi-Orthogonal Matrix)).

Extending Semi-Orthogonality to Convolution. Let $W \in \mathbb{R}^{c_{out} \times c_{in} \times n \times n}$ be a convolutional filter where c_out is the number of output channels, c_in is the number of input channels, and $n \times n$ are the input dimensions. The action on an input $x \in \mathbb{R}^{cc_{in} \times n \times n}$ is denoted by:

$$W \circledast x : \mathbb{R}^{c_{in} \times n \times n} \to \mathbb{R}^{c_{out} \times n \times n}. \tag{203}$$

Using the norm preserving concept for 1-Lipschitz maps on ℓ_2, for $\forall\ x \in \mathbb{R}^{cc_{in} \times n \times n}$, if the 2-norms are preserved:

$$\|W \circledast x\|_2 = \|x\|_2, \tag{204}$$

then the convolutional filter W is referred to as semi-orthogonal (Trockman & Kolter, 2021).

Filter Orthogonalization via Cayley Transform. However, convolutional operation is not matrix multiplication, to apply the existing results from matrix theory, Trockman & Kolter (2021) harness the *Convolutional Theorem* (Jain, 1989) for converting the convolutional operation in spatial domain to multiplication in frequency domain by:

$$\mathscr{F}\,[W \circledast x]\,[:, i, j] = \mathscr{F}\,[W]\,[:, :, i, j]\;\mathscr{F}\,[x]\,[:, :, i, j] = \widehat{W}[:, i, j]\widehat{x}[:, i, j], \tag{205}$$

where $\mathscr{F}$ represents Fourier transform, $\widehat{W} = \mathscr{F}[W]$, $\widehat{x} = \mathscr{F}[x]$, $[:, :, i, j]$ and $[:, i, j]$ are slicing operations by fixing indices (i, j)) (Trockman & Kolter, 2021, Equantion 3 & 4, § 4). Note that $\widehat{W}[:, :, i, j]$ is not orthogonal or semi-orthogonal, Trockman & Kolter (2021) use a bijective *Cayley Transform* for deriving an orthogonal matrix $Q[:, :, i, j]$ from $\widehat{W}[:, :, i, j]$ by:

$$Q[:, :, i, j] = (I - A[:, :, i, j])(I + A[:, :, i, j])^{-1}, \tag{206}$$

where:

$$A[:, :, i, j] = \widehat{W}[:, :, i, j] - \widehat{W}[:, :, i, j]^*, \tag{207}$$

and $A[:, :, i, j]$ is referred to as *skew-symetric* matrix (Trockman & Kolter, 2021). Conceptually, the convolution on frequency domain is then computed through:

$$\begin{aligned} \mathscr{F}[W \circledast x][:, i, j] &= Q[:, :, i, j]\, \widehat{x}[:, i, j] && (208)\\ &= (I - A[:, :, i, j])(I + A[:, :, i, j])^{-1}\, \widehat{x}[:, i, j], && (209) \end{aligned}$$

then $Q[:, :, i, j]\, \widehat{x}[:, i, j]$ is inverted back to spatial domain by:

$$\mathscr{F}^{-1}\left[Q[:, :, i, j]\, \widehat{x}[:, i, j]\right], \tag{210}$$

which leads to the convolution 1-Lipschitz continuous.

4.5.2 Orthogonality with Lie Unitary Group

Lezcano-Casado & Martínez-Rubio (2019) introduce a method, stemming from Lie group theory, through the exponential map, for ensuring 1-Lipschitz by construction. Orthogonal constraints networks can improve the robustness and generalization capabilities (Huang et al., 2018; Bansal et al., 2018). For a matrix A, if $AA^* = I$, then the matrix A is referred to as an unitary matrix. Unitary matrices allow to solve the exploding or vanish gradients (Arjovsky et al., 2016; 2017). In the sense of Lie algebra, these unitary matrices form a special orthogonal group on field $\mathbb{R}$:

$$SO(n) = \left\{A \in \mathbb{R}^{n \times n} \mid A^\top A = I\right\} \tag{211}$$

and unitary group on field $\mathbb{C}$:

$$U(n) = \left\{A \in \mathbb{C}^{n \times n} \mid A^* A = I\right\} \tag{212}$$

(Lezcano-Casado & Martínez-Rubio, 2019, § 3.1). We are interested of converting a parameter matrix $W \in \mathbb{R}^{n \times n}$ to on $SO(n)$ or $U(n)$. Lezcano-Casado & Martínez-Rubio (2019) harness the concept of *skew-symmetric matrix* group (Definition 2.7 (Skew-Hermitian & Skew-Symmetric Matrix)) as a bridge:

$$\mathfrak{so}(n) = \{B \in \mathbb{R}^{n \times n} \ : \ B = -B^\top\} \tag{213}$$

and:

$$\mathfrak{u}(n) = \{B \in \mathbb{R}^{n \times n} \ : \ B = -B^*\}. \tag{214}$$

Taking:

$$B = W - W^\top \tag{215}$$

can easily send a parameter matrix W to $\mathfrak{u}(n)$.

Connecting to Lie Orthogonal Group. Then there exist an exponential *surjective* map between two groups:

$$\exp : \mathfrak{so}(n) \to SO(n) \tag{216}$$

and:

$$\exp : \mathfrak{u}(n) \to U(n), \tag{217}$$

which is defined as:

$$\exp(B) = I + B + \frac{1}{2}B^2 + \cdots \tag{218}$$

(Lezcano-Casado & Martínez-Rubio, 2019, § 3.2). Then a network $f(x; W)$ is parameterized on the Lie group $SN(n)$ by an algebraic mapping chain:

$$W \in \mathbb{R}^{n \times n} \xrightarrow{B=W-W^\top} B \in \mathfrak{so}(n) \xrightarrow{A=\exp(A)} A \in SO(n), \tag{219}$$

and the optimization problem becomes:

$$\min_A \; f(x; A) \Longleftrightarrow \min_B \; f(x; \exp(B)) \tag{220}$$

(Lezcano-Casado & Martínez-Rubio, 2019, § 3.3).

Optimizing on Lie Orthogonal Group. The Lie group $SO(n)$ forms a Riemannian manifold, to optimize a function f on the Riemannian manifold $SO(n)$:

$$f : \mathcal{X} \times SO(n) \to \mathbb{R}, \tag{221}$$

where $\mathcal{X}$ is input space, one may use *Riemannian gradient descent* (Absil et al., 2009). Given a point on the manifold $A \in SO(n)$, $\mathcal{T}_A\, SO(n)$ is the tangent space at A with induced metric, and a gradient $\Omega \in \mathcal{T}_A\, SO(n)$, then the geodesic update is:

$$A \leftarrow A \,\exp\left(-\eta\, A^* \,\Omega\right), \tag{222}$$

where η is a learning rate $\eta > 0$ (Lezcano-Casado & Martínez-Rubio, 2019). Computing the Riemannian exponential map is expensive, Lezcano-Casado & Martínez-Rubio (2019) use a first-order approximation, *i.e.*, *Cayley map*, for the update instead:

$$A \leftarrow A\, \phi\left(-\eta\, A^* \,\Omega\right), \tag{223}$$

where ϕ is the *Cayley map*:

$$\phi(A) = (I + \frac{1}{2}A)(I - \frac{1}{2}A)^{-1} \tag{224}$$

(Wisdom et al., 2016; Vorontsov et al., 2017). Finally, this induces an update rule on $B \in \mathfrak{so}(n)$:

$$\exp(B) \leftarrow \exp\Big(B - \eta\, \nabla\, f\,(x; \exp(B))\Big) \tag{225}$$

where $\nabla\, f\,(x; \exp(B))$ is the usual gradient with Euclidean metric. In a specially designed recurrent neural network (RNN), Lezcano-Casado & Martínez-Rubio (2019) transform the matrix exponential maps skew-symmetric matrices to orthogonal matrices transforming an optimization problem with orthogonal constraints (Lezcano-Casado & Martínez-Rubio, 2019). Particularly, Lezcano-Casado & Martínez-Rubio (2019) use Padé approximants and the scale-squaring trick to compute machine-precision approximations of the matrix exponential and its gradient (Lezcano-Casado & Martínez-Rubio, 2019). As a result, the optimization remains in the group and the Lipschitz constants of the updated matrices remain 1 by design.

4.5.3 Lipschitz Continuous Transformer

Qi et al. (2023) demonstrate that enforcing Lipschitz continuity in Transformer architecture (Vaswani et al., 2017) is more crucial for ensuring training stability in contrast to the tricks such as *learning rate warmup*, *layer normalization*, *attention formulation*, and *weight initialization* (Qi et al., 2023). Their proposed method, referred to as **LipsFormer**, replaces the Transformer parts with their Lipschitz continuous counterparts: *CenterNorm* for *LayerNorm*, *spectral initialization* for *Xavier initialization*, *scaled cosine similarity attention* for *dot-product attention*, and *weighted residual shortcut* for *residual connection* (Qi et al., 2023). These modifications result in a Transformer architecture with a provably bounded Lipschitz constant.

CenterNorm instead of LayerNorm. LayerNorm (Ba et al., 2016) is widely used in Transformer, defined as:

$$\mathrm{LN}(x) = \gamma \odot z + \beta, \tag{226}$$

and:

$$z = \frac{y}{\mathrm{std}(y)}, \qquad y = \Big(I_d - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\top\Big)x, \tag{227}$$

where $x \in \mathbb{R}^d$ is input, I_d is an identity matrix in $\mathbb{R}^{d\times d}$, $\mathbf{1}_d$ is an all-ones vector in $\mathbb{R}^d$, $\mathrm{std}(y)$ is the standard deviation of y, $\odot$ is element-wise product, α and β are learnable parameters initialized to 1 and 0 respectively (Qi et al., 2023). The Jacobian of z with respect to x is given as:

$$\frac{\partial z}{\partial x} = \frac{1}{\mathrm{std}(y)}\Big(I_d - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\top\Big)\Big(I_d - \frac{yy^\top}{\|y\|_2^2}\Big), \tag{228}$$

which shows that **LayerNorm** is not Lipschitz continuous when $\mathrm{std}(y) \to 0$. To address this problem, Qi et al. (2023) introduce **CenterNorm**, defined as:

$$\mathrm{CN}(x) = \gamma \odot \frac{d}{d-1}\Big(I_d - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\top\Big)x + \beta, \tag{229}$$

which admits a Lipschitz constant:

$$\mathrm{Lip}\,[\mathrm{CN}] = \frac{d}{d-1} \approx 1 \tag{230}$$

for sufficient large d, $\alpha = 1$, and $\beta = 0$ (Qi et al., 2023).

Scaled Cosine Similarity Attention (SCSA). Kim et al. (2021) show that standard dot-product self-attention is not globally Lipschitz continuous since the Lipschitz constant depends on sequence length which is not bounded (Kim et al., 2021), see also Section 2.6 (Lipschitz Continuity of Dot-Product Self-Attention). To address this problem, Qi et al. (2023) replace standard dot-product attention part with **SCSA**, defined as:

$$\mathrm{SCSA}(x; v, \tau) = vPV, \qquad P = \mathrm{softmax}(\tau\, QK^\top), \tag{231}$$

where K, Q, V are row-wise normalized in ℓ_2 (Qi et al., 2023, § 4.1.2).

Remark 4.1. It is worth noting that their method reduces the Lipschitz constant of the attention module; however, the module remains not globally Lipschitz continuous since the Lipschitz constant still depends on sequence length n:

$$\mathrm{Lip}_2\,[SCSA] \le O(n), \qquad \text{and} \qquad \mathrm{Lip}_\infty\,[SCSA] \le O(n^2) \tag{232}$$

(Qi et al., 2023, Theorem 1).

Weighted Residual Shortcut. For a residual module:

$$h(x) = x + g(x), \tag{233}$$

where $g : \mathbb{R}^d \to \mathbb{R}^d$, $h : \mathbb{R}^d \to \mathbb{R}^d$ and $x \in \mathbb{R}^d$, the Lipschitz constant of h is bounded by:

$$\mathrm{Lip}\left[h\right] \leq 1 + \mathrm{Lip}\left[g\right], \tag{234}$$

see Section 2.7.3 (Lipschitz Constant of Residual Network). To further reduce the Lipschitz constant introduced by residual module, Qi et al. (2023) replaces the standard residual module with:

$$\mathrm{WRS}(x) = x + \alpha \odot g(x), \tag{235}$$

where α is a learnable parameter, initialized to a small value $[0.1, 0.2]$ (Qi et al., 2023). **Note that the result:**

$$\mathrm{Lip}\left[\mathrm{WRS}\right] = 1 + \max(\alpha) \tag{236}$$

in the Qi et al. (2023, § 4.1.3) is incorrect mathematically. The Lipschitz constant of WRS in residual module should be bounded by:

$$\mathrm{Lip}\left[\mathrm{WRS}\right] \leq 1 + \max(\alpha). \tag{237}$$

Spectral-based Weight Initialization. Qi et al. (2023) further normalize the parameters at initialization by spectral-norm. Let W be a parameter matrix, the W is normalized by:

$$W \leftarrow \frac{W}{\|W\|_2} \tag{238}$$

(Qi et al., 2023, § 4.1.4).

4.5.4 Jacobian Norm Minimization

Constraining the Lipschitz constant of Transformer architecture for stable optimization and robustness remains an open research problem. Yudin et al. (2025) first derive an explicit tight local Lipschitz constant for self-attention module in Transformer architecture by analyzing the Jacobian of softmax (Yudin et al., 2025). Building on top of the Jacobian analsis, they present **JaSMin** (**Ja**cobian **S**oftmax norm **Min**imization) for enhancing Transformer's robustness by constraining the local Lipschitz constant (Yudin et al., 2025).

Bounding Lipschitz through Softmax Jacobian. Consider the softmax function:

$$\mathrm{softmax} : \mathbb{R}^d \to \mathbb{R}^d, \tag{239}$$

then the Jacobian of the softmax function is:

$$\mathcal{M}(P_x) := \nabla\,\mathrm{softmax}\,(x) = \mathrm{diag}(P_x) - P_x^\top P_x, \tag{240}$$

where:

$$P_x = \left(\frac{\mathrm{e}^{x_i}}{\sum_j \mathrm{e}^{x_j}}\right). \tag{241}$$

Yudin et al. (2025) incorporate this Jacobian expression of the softmax function into self-attention, leading a Lipschitz constant bound for a single head:

$$\mathrm{Lip}\Big[\mathrm{Attention}\Big] \leq \|W_V\|_2 \left(\|P\|_2 + 2\|x\|_2^2 \; \|A\|_2 \;\; \max_i \|\mathcal{M}(P_{i,:})\|_2 \right) \tag{242}$$

with:

$$P = \operatorname{softmax}\Big(xAx^\top\Big) \in \mathbb{R}^{n\times n} \qquad \text{and} \qquad A = \frac{W_Q W_K^\top}{\sqrt{d}} \in \mathbb{R}^{d\times d} \tag{243}$$

(Yudin et al., 2025, Theorem 3). It worthy noting that the local Lipschitz constant contains a term:

$$\operatorname{Lip}\Big[\text{Attention}\Big] \leq O\Big(\max_i \|\mathcal{M}(P_{i,:})\|_2\Big), \tag{244}$$

which shows that softmax Jacobian norm determines an upper bound of the local Lipschitz constant.

Jacobian Softmax Norm Minimization. To approximate the Jacobian norm more efficiently, Yudin et al. (2025) introduce a surrogate function g that captures the partial ordering of the singular values of softmax mappings. For $x \in \mathbb{R}^n_{>0}$, let $x_{(k)}$ be the k-th largest component of x. Define:

$$g_k(x) := x_{(k)}(1 - x_{(k)} + x_{(k+1)}), \tag{245}$$

for $k = 1, 2, \cdots, n-1$ (Yudin et al., 2025, Definition 1). Let $A = \operatorname{diag}(x) - xx^\top$, then the following inequality holds true:

$$x_{(1)} \geq g_1(x) \geq \sigma_1(A) \geq x_{(2)} \geq g_2(x) \geq \sigma_2(A) \geq x_{(n)} \geq g_n(x) \geq \sigma_n(A) \geq 0 \tag{246}$$

(Yudin et al., 2025, Theorem 4).

Building on the derived bound, Yudin et al. (2025) propose an efficient regularization loss expressed in terms of the function g with two forms:

$$\mathcal{L}_{\text{JaSMin}(k=0)} = \sum_{\ell=1}^{L}\sum_{h_i=1}^{h} \max_j \log\Big[g_1\Big(P^{\ell,i}_{j,:}\Big)\Big], \tag{247}$$

and:

$$\mathcal{L}_{\text{JaSMin}(k)} = \sum_{\ell=1}^{L}\sum_{h_i=1}^{h} \max_j \log\left[\frac{g_1\Big(P^{\ell,i}_{j,:}\Big)}{g_k\Big(P^{\ell,i}_{j,:}\Big)}\right], \tag{248}$$

where $P^{\ell,i}_{j,:}$ represents the softmax map for the ℓ-th attention module ($1 \leq \ell \leq L$), i-th head ($1 \leq i \leq h$), and the j-th row of the map (Yudin et al., 2025, § 4). They also show that for ${}^{g_1}/_{g_k}$ is bounded below γ:

$$\frac{g_1\Big(P^{\ell,i}_{j,:}\Big)}{g_k\Big(P^{\ell,i}_{j,:}\Big)} \leq \gamma, \tag{249}$$

for $1 \leq \gamma \leq \frac{k}{4}$. Then the softmax Jacobian 2-norm is bounded by:

$$\|\mathcal{M}(P^{\ell,i}_{j,:})\|_2 \leq O\Big(\frac{\gamma}{k}\Big) \tag{250}$$

(Yudin et al., 2025, Proposition 1).

5 Certifiable Robustness

Certifiable robustness, a cornerstone of trustworthy deep learning, guarantees that a neural network's predictions remain consistent under adversarial perturbations within a specified norm ball. Unlike empirical defenses like adversarial training, which lack formal guarantees, Lipschitz-based certification methods leverage the Lipschitz constant to bound sensitivity to input changes, ensuring provable robustness. While some literature, such as (Clevert et al., 2016), have been introduced previously from the perspectives — such as **theoretical foundations**, **estimation methods** and **regularization approaches**, we may revisit some of them in this section from the perspective of **certifiable robustness**.

5.1 Formal Robustness Guarantees

In the setting of an n-class classifier f with Lipschitz constant K, a central question in adversarial defense is to determining the largest perturbation norm that leaves the prediction unchanged.

Theorem 5.1 (Lipschitz Margin Robustness Radius in ℓ_2-Normed Spaces)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^n$ be a classifier. Let f_i be the i-th output. Let $c = \arg\max_i f_i(x)$ be the predicted class for x. Define the (one-versus-rest) margin*

$$m(x) \;:=\; f_c(x) - \max_{j\neq c} f_j(x). \tag{251}$$

Assume that f is K-Lipschitz from the input ℓ_2 norm to the output ℓ_2 norm, i.e.*,*

$$\|f(x) - f(y)\|_2 \;\le\; K\,\|x - y\|_2 \quad \textit{for all } x, y. \tag{252}$$

Then for any perturbation δ with

$$\|\delta\|_2 \;<\; \frac{m(x)}{\sqrt{2}\,K}, \tag{253}$$

the prediction is unchanged:

$$\arg\max_i f_i(x+\delta) = c. \tag{254}$$

Remark 5.2. This result is particularly useful for analyzing the adversarial defense problem. Similar results have been discussed in the literature (Szegedy et al., 2014; Hein & Andriushchenko, 2017; Cisse et al., 2017; Tsuzuku et al., 2018; Perugachi-Diaz et al., 2021).

Proof. Let $e_i \in \mathbb{R}^n$ denote the i-th standard basis vector, *i.e.*, e_i has a 1 in the i-th position and 0 elsewhere. Then

$$f_i(x) = e_i^\top f(x). \tag{255}$$

For each $j \neq c$, define the pairwise margin function:

$$g_j(x) := f_c(x) - f_j(x) = (e_c - e_j)^\top f(x). \tag{256}$$

By Hölder's inequality,

$$\begin{aligned} |g_j(x) - g_j(y)| &= |(e_c - e_j)^\top (f(x) - f(y))| && (257)\\ &\le \|e_c - e_j\|_2 \cdot \|f(x) - f(y)\|_2 && (258)\\ &\le \sqrt{2} \cdot K \cdot \|x - y\|_2, && (259) \end{aligned}$$

so each g_j is $\sqrt{2}K$-Lipschitz:

$$\mathrm{Lip}[g_j] = \sqrt{2}K. \tag{260}$$

Note that the margin $m(x)$ can be expressed as

$$m(x) = f_c(x) - \max_{j\neq c} f_j(x) = \min_{j\neq c} g_j(x). \tag{261}$$

Now let $\delta \in \mathbb{R}^d$ be a perturbation satisfying

$$\|\delta\|_2 < \frac{m(x)}{\sqrt{2}K}. \tag{262}$$

Then for all $j \neq c$,

$$g_j(x+\delta) \geq g_j(x) - \mathrm{Lip}[g_j] \cdot \|\delta\|_2 \tag{263}$$

$$> m(x) - \sqrt{2}K \cdot \frac{m(x)}{\sqrt{2}K} = 0. \tag{264}$$

Hence, $f_c(x+\delta) > f_j(x+\delta)$ for all $j \neq c$, which implies

$$\arg\max_i f_i(x+\delta) = c. \tag{265}$$

□

Theorem 5.3 (Lipschitz Margin Robustness Radius in Conjugate-Normed Spaces)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^n$ be a classifier with predicted class $c = \arg\max_i f_i(x)$ and one-versus-rest margin*

$$m(x) := f_c(x) - \max_{j\neq c} f_j(x). \tag{266}$$

Assume f is K-Lipschitz from the input ℓ_p norm to the output ℓ_q norm:

$$\|f(x) - f(y)\|_q \;\leq\; K\,\|x-y\|_p. \tag{267}$$

Let p be the dual exponent of q, i.e., $\frac{1}{q} + \frac{1}{p} = 1$*. Then the classification of x is certifiably robust to any perturbation δ with*

$$\|\delta\|_p \;<\; r_{\mathrm{cert}}(x) := \frac{m(x)}{2^{\frac{1}{p}}\,K}, \tag{268}$$

where $e_i \in \mathbb{R}^n$ denotes the i-th standard basis vector — e.g.*, one-hot vector.*

Remark 5.4. Similar results have been discussed in the literature (Hein & Andriushchenko, 2017; Weng et al., 2018b).

Proof. Let $e_i \in \mathbb{R}^n$ denote the i-th standard basis vector, *i.e.*, e_i has a 1 in the i-th position and 0 elsewhere. Then

$$f_i(x) = e_i^\top f(x). \tag{269}$$

For each $j \neq c$, define the pairwise margin function:

$$g_j(x) := f_c(x) - f_j(x) = (e_c - e_j)^\top f(x). \tag{270}$$

By Hölder's inequality and the Lipschitz property of f,

$$|g_j(x) - g_j(y)| \leq \|e_c - e_j\|_p\,\|f(x) - f(y)\|_q \tag{271}$$

$$\leq K\,\|e_c - e_j\|_p\,\|x-y\|_p. \tag{272}$$

Thus the Lipschitz constant of g_j is:

$$\mathrm{Lip}[g_j] \leq K\,\|e_c - e_j\|_p. \tag{273}$$

If:

$$\|x' - x\|_p < \frac{g_j(x)}{\mathrm{Lip}[g_j]}, \tag{274}$$

for all $j \neq c$, then $g_j(x') > 0$ for all $j \neq c$, preserving the predicted class. Taking the minimum over j yields

$$r_{\mathrm{cert}}(x) = \min_{j\neq c} \frac{g_j(x)}{\mathrm{Lip}[g_j]} \geq \frac{m(x)}{K\,\max_{j\neq c}\|e_c - e_j\|_p} = \min_{j\neq c} \frac{g_j(x)}{\mathrm{Lip}[g_j]} \geq \frac{m(x)}{K\,2^{\frac{1}{p}}}. \tag{275}$$

□

Remark 5.5. Note that:

$$\alpha = \max_{j \neq c} \|e_c - e_j\|_p = 2^{\frac{1}{p}}, \tag{276}$$

common output norm choices yield:

$$\ell_2 \text{ output norm:} \quad \alpha = \sqrt{2} \qquad \Rightarrow \qquad r_{\text{cert}} \geq \frac{m(x)}{\sqrt{2}\,K}, \tag{277}$$

$$\ell_\infty \text{ output norm:} \quad \alpha = 1 \qquad \Rightarrow \qquad r_{\text{cert}} \geq \frac{m(x)}{K}, \tag{278}$$

$$\ell_1 \text{ output norm:} \quad \alpha = 2 \qquad \Rightarrow \qquad r_{\text{cert}} \geq \frac{m(x)}{2K}. \tag{279}$$

Note that the number of classes n affects the certified radius only through the margin $m(x)$; it does not appear explicitly in the bound.

5.2 Certifiable Robustness via Global Lipschitz Bound

For a network f with 1-Lipschitz activations and $W^{(\ell)}$ is the ℓ-th layer parameter matrix, the spectral-norm product bounds the Lipschitz constant of f globally:

$$\text{Lip}[f] \leq \prod_{\ell=1}^{L} \|W^{(\ell)}\|_2, \tag{280}$$

providing a *certified global bound*. This is computationally cheap but typically loose. **SeqLip** (Virmaux & Scaman, 2018) refines this spectral product by incorporating singular-vector alignment between layers, also producing certified global bounds. **LipSDP** (Fazlyab et al., 2019) interprets activation functions as gradients of convex potential functions, satisfying certain properties described by quadratic constraints. Therefore, Lipschitz constant estimation problem is treated as a semi-definite program (SDP) problem.

5.3 Certifiable Robustness via Local Lipschitz Bound

Local Lipschitz bounds are computed over a neighborhood. **Fast-Lip** (Weng et al., 2018a) analyze the activation patterns of ReLU networks and derive a gradient norm based local Lipschitz bound. **CLEVER** (Weng et al., 2018b) estimates local constants via extreme value theory on sampled gradient norms; this improves tightness but does not yield a formal certificate. MILP-based methods (Jordan & Dimakis, 2020) encode the exact piecewise-linear constraints of ReLU networks within the ball, yielding exact local Lipschitz constants (and thus exact certified radii), but are limited to small networks.

5.4 Certifiable Robustness in Language Models

Proper Distance Metrics for Text. Language models typically operate on text sequences as inputs, for which the conventional notion of a Lipschitz constant over $\mathbb{R}$ with ℓ_p norm does not directly apply. The *Levenshtein distance* (Levenshtein, 1966) measures the number of character replacements, insertions or deletions needed in order to transform a sequence x to sequence y. It is a proper metric that satisfies all axioms of a metric space:

1. **Non-Negativity.** $d_{\text{lev}}(x, y) \geq 0$;
2. **Identity.** $d_{\text{lev}}(x, y) = 0 \iff x = y$;
3. **Symmetry.** $d_{\text{lev}}(x, y) = d_{\text{lev}}(y, x)$;
4. **Triangle Inequality.** $d_{\text{lev}}(x, z) \leq d_{\text{lev}}(x, y) + d_{\text{lev}}(y, z)$;

where $d_{\text{lev}}(x, y)$ denotes the Levenshtein distance between two sequence x and y. Thereby it provides a principled foundation for discussing Lipschitz continuity in the context of language models. For tokenized text — where each vocabulary item is represented as a one-hot vector — a variant of the *Levenshtein distance*, known as the *ERP distance*, has been proposed for comparing sequences of one-hot vectors (Chen & Ng, 2004).

Rocamora et al. (2024) introduce **LipsLev**, a method for training convolutional text classifiers that are *1-Lipschitz* in the sense of Levenshtein distance for Lipschitz constant based defense (Rocamora et al., 2024). That is, for a classifier $f : S \to \mathbb{R}^c$, two one-hot vector-valued sequences $x \in \mathbb{R}^{n\times d} \subset S$ and $x \in \mathbb{R}^{m\times d} \subset S$ where n, m are sequence lengths and d is one-hot vector dimension, the classification margin must hold:

$$|f_i(x) - f_j(y)| \leq K_{\text{lev}}\ d_{\text{lev}}(x, y) \tag{281}$$

for a constant $K_{\text{lev}} > 1$, so that the prediction does not change under perturbation radius $d_{\text{lev}}(x, y) < R$ (Rocamora et al., 2024, Theorem 3.3). They then train a 1-Lipschitz classifier over text sequences — in the sense of the Levenshtein distance — by normalizing each layer's outputs through division by its corresponding Lipschitz constant (Rocamora et al., 2024, § 3.3).

6 Conclusions

Lipschitz continuity serves as a cornerstone for enhancing the robustness, generalization, and optimization stability of deep neural networks, providing a rigorous mathematical framework to quantify and control sensitivity to input perturbations. This comprehensive survey synthesizes key insights across theoretical foundations, estimation techniques, regularization approaches, and certifiable robustness, highlighting their interconnections and practical implications. By unifying disparate research threads, we address a critical gap in the literature, offering a cohesive resource for researchers and practitioners. Challenges remain in balancing bound tightness with computational scalability, preserving model expressivity under constraints, and extending certifications to diverse norms and large-scale architectures like transformers, paving the way for future advancements in trustworthy deep learning systems.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Fatemeh Amerehi and Patrick Healy. Label augmentation for neural networks robustness. In Vincenzo Lomonaco, Stefano Melacci, Tinne Tuytelaars, Sarath Chandar, and Razvan Pascanu (eds.), *Proceedings of The 3rd Conference on Lifelong Learning Agents*, volume 274 of *Proceedings of Machine Learning Research*, pp. 620–640. PMLR, 29 Jul–01 Aug 2025. URL https://proceedings.mlr.press/v274/amerehi25a.html.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. 2016. URL https://arxiv.org/abs/1606.06565.

Cem Anil, James Lucas, and Roger Grosse. Sorting out Lipschitz function approximation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 291–301. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/anil19a.html.

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1120–1128, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/arjovsky16.html.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep cnns? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 4266–4276, Red Hook, NY, USA, 2018. Curran Associates Inc.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6241–6250, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Joern-Henrik Jacobsen. Invertible residual networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 573–582. PMLR, 09–15 Jun 2019a. URL https://proceedings.mlr.press/v97/behrmann19a.html.

Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Joern-Henrik Jacobsen. Invertible residual networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 573–582. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/behrmann19a.html.

Rianne van den Berg, Leonard Hasenclever, Jakub M Tomczak, and Max Welling. Sylvester normalizing flows for variational inference. UAI, 2018.

Å. Björck and C. Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971. doi: 10.1137/0708036. URL https://doi.org/10.1137/0708036.

Stephen Boyd, John Duchi, Mert Pilanci, and Lieven Vandenberghe. Notes for ee364b. Stanford University, 2022. URL https://stanford.edu/class/ee364b/lectures/subgradients_notes.pdf. Lecture notes for Spring 2021–22.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5817–5840. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/castin24a.html.

Arthur Cayley. Sur quelques propriétés des déterminants gauches. 1846.

Lei Chen and Raymond Ng. On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pp. 792–803. VLDB Endowment, 2004. ISBN 0120884690.

Ricky T. Q. Chen, Jens Behrmann, David K Duvenaud, and Joern-Henrik Jacobsen. Residual flows for invertible generative modeling. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/5d0d5594d24f0f955548f0fc0ff83d10-Paper.pdf`.

Artem Chernodub and Dimitri Nowicki. Norm-preserving orthogonal permutation linear unit activation functions (oplu), 2017. URL `https://arxiv.org/abs/1604.02313`.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL `http://jmlr.org/papers/v24/22-1144.html`.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning (ICML)*, pp. 854–863. PMLR, 2017. URL `https://arxiv.org/abs/1704.08847`.

Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *ICLR*, 2016.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan

Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL `https://arxiv.org/abs/2501.12948`.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=HkpbnH9lx`.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In T. Leen, T. Dietterich, and V. Tresp (eds.), *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL `https://proceedings.neurips.cc/paper_files/paper/2000/file/44968aece94f667e4095002d140b5896-Paper.pdf`.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/95e1533eb1b20a97777749fb94fdb944-Paper.pdf`.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 297–299. PMLR, 06–09 Jul 2018. URL `https://proceedings.mlr.press/v75/golowich18a.html`.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

Antoine Gonon, Nicolas Brisebarre, Elisa Riccietti, and Rémi Gribonval. A rescaling-invariant lipschitz bound based on path-metrics for modern reLU network parameterizations. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=T8VLY1KuOz`.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021. doi: 10.1007/s10994-020-05929-w. URL `https://doi.org/10.1007/s10994-020-05929-w`.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5769–5779, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/e077e1a544eec4f0307cf5c3c721d944-Paper.pdf`.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

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

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Xixu Hu, Runkai Zheng, Jindong Wang, Cheuk Hang Leung, Qi Wu, and Xing Xie. Specformer: Guarding vision transformer robustness via maximum singular value penalization. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LIV*, pp. 345–362, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72948-5. doi: 10.1007/978-3-031-72949-2_20. URL https://doi.org/10.1007/978-3-031-72949-2_20.

Lei Huang, Xianglong Liu, Bo Lang, Adams Wei Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Anil K Jain. Fundamentals of digital image processing. 1989.

Matt Jordan and Alexandros G. Dimakis. Exactly computing the local lipschitz constant of relu networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5562–5571. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kim21i.html.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

Mario Lezcano-Casado and David Martínez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3794–3803. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lezcano-casado19a.html.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Qiyang Li, Saminul Haque, Cem Anil, James Lucas, Roger B Grosse, and Joern-Henrik Jacobsen. Preventing gradient attenuation in lipschitz constrained convolutional networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1ce3e6e3f452828e23a0c94572bef9d9-Paper.pdf.

Róisín Luo, James McDermott, Christian Gagné, Qiang Sun, and Colm O'Riordan. Optimization-induced dynamics of Lipschitz continuity in neural networks. 2025a. URL https://arxiv.org/abs/2506.18588.

Róisín Luo, Colm O'Riordan, and James McDermott. Higher-order singular-value derivatives of real rectangular matrices. *Journal of Mathematical Analysis and Applications*, 556(2):130236, 2025b. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2025.130236. URL https://www.sciencedirect.com/science/article/pii/S0022247X25010170.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings*, pp. 3–17, Berlin, Heidelberg, 2016. Springer-Verlag. ISBN 978-3-319-46378-0. doi: 10.1007/978-3-319-46379-7_1. URL https://doi.org/10.1007/978-3-319-46379-7_1.

RV Mises and Hilda Pollaczek-Geiringer. Praktische verfahren der gleichungsauflösung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 9(1):58–77, 1929.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Pravin Nair. Softmax is 1/2-lipschitz: A tight bound across all ℓ_p norms. *arXiv preprint arXiv:2510.23012*, 2025.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401, Paris, France, 03–06 Jul 2015a. PMLR. URL https://proceedings.mlr.press/v40/Neyshabur15.html.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pp. 1376–1401. PMLR, 2015b.

Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/10ce03a1ed01077e3e289f3e53c72813-Paper.pdf.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Yura Perugachi-Diaz, Jakub Tomczak, and Sandjai Bhulai. Invertible densenets with concatenated lipswish. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17246–17257. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.

Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, and Lei Zhang. Lipsformer: Introducing lipschitz continuity to vision transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cHf1DcCwcH3.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2018. URL https://openreview.net/forum?id=SkBYYyZRZ.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/rezende15.html.

Elias Abad Rocamora, Grigorios Chrysos, and Volkan Cevher. Certified robustness in NLP under bounded levenshtein distance. In *ICML 2024 Next Generation of AI Safety Workshop*, 2024. URL https://openreview.net/forum?id=5llRwQO3CU.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 3 edition, 1976.

Walter Rudin. *Real and complex analysis*. McGraw-Hill, Inc., 1987.

Hanie Sedghi, Vineet Gupta, and Philip M Long. The singular values of convolutional layers. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Hygvb2C5FX.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, Cambridge, UK, 2014. ISBN 9781107057135. URL https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/copy.html.

Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017. doi: 10.1109/TSP.2017.2708039.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman,

Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina

Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin

Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Pbj8H_jEHYv.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: scalable certification of perturbation invariance for deep neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6542–6551, Red Hook, NY, USA, 2018. Curran Associates Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d54e99a6c03704e95e6965532dec148b-Paper.pdf.

Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3570–3578. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/vorontsov17a.html.

James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. A mathematical theory of attention. *arXiv preprint arXiv:2007.02876*, 2020.

Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for ReLU networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5276–5285. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/weng18a.html.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. In *International Conference on Learning Representations*, 2018b. URL https://openreview.net/forum?id=BkUHlMZ0b.

Scott Wisdom, Thomas Powers, John Hershey, Jonathan Le Roux, and Les Atlas. Full-capacity unitary recurrent neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/d9ff90f4000eacd3a6c9cb27f78994cf-Paper.pdf.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

Nikolay Yudin, Alexander Gaponov, Sergei Kudriashov, and Maxim Rakhuba. Pay attention to attention distribution: A new local lipschitz bound for transformers. *arXiv preprint arXiv:2507.07814*, 2025. URL https://arxiv.org/pdf/2507.07814.

A Proofs: Lipschitz Constants of Activation Functions

A.1 Proof: Lipschitz Constant of Sigmoid

Proof. Consider the sigmoid function:

$$\text{Sigmoid}(x) = f(x) = (1 + e^{-x})^{-1}. \tag{282}$$

Let $u = 1 + e^{-x}$, so $f(x) = u^{-1}$. The derivative is:

$$f'(x) = -\frac{1}{u^2} \cdot \frac{du}{dx}. \tag{283}$$

Compute $\frac{du}{dx}$:

$$u = 1 + e^{-x}, \quad \frac{du}{dx} = -e^{-x}. \tag{284}$$

Thus:

$$f'(x) = -\frac{1}{(1+e^{-x})^2} \cdot (-e^{-x}) = \frac{e^{-x}}{(1+e^{-x})^2}. \tag{285}$$

Express the derivative in terms of $f(x)$:

$$1 - f(x) = \frac{e^{-x}}{1+e^{-x}}, \tag{286}$$

and:

$$f'(x) = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = f(x)(1-f(x)). \tag{287}$$

Thus:

$$|f'(x)| = f(x)(1-f(x)). \tag{288}$$

Maximize the Derivative. Since $0 < f(x) < 1$, we maximize $g(z) = z(1-z)$ for $z = f(x) \in (0,1)$:

$$g'(z) = 1 - 2z = 0 \implies z = \frac{1}{2}, \tag{289}$$

thus:

$$g\left(\frac{1}{2}\right) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}. \tag{290}$$

Find when $f(x) = \frac{1}{2}$:

$$\frac{1}{1+e^{-x}} = \frac{1}{2} \implies e^{-x} = 1 \implies x = 0, \tag{291}$$

at $x = 0$:

$$f'(0) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}. \tag{292}$$

As $x \to \infty$, $e^{-x} \to 0$, so $f'(x) \to 0$. As $x \to -\infty$, $e^{-x} \to \infty$, so:

$$f'(x) \approx \frac{e^{-x}}{e^{-2x}} = e^{x} \to 0. \tag{293}$$

Hence:

$$\text{Lip}\left[\text{Sigmoid}(x)\right] = \sup_x |f'(x)| = \frac{1}{4}. \tag{294}$$

□

A.2 Proof: Lipschitz Constant of Tanh

Proof. The tanh is defined as:

$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{295}$$

for all $x, y \in \mathbb{R}$.

Let $g(x) = \sinh(x)$, $h(x) = \cosh(x)$, so $f(x) = \frac{g(x)}{h(x)}$. The derivative is:

$$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h(x)^2}. \tag{296}$$

Since $g'(x) = \cosh(x)$, $h'(x) = \sinh(x)$, and $h(x)^2 = \cosh^2(x)$:

$$f'(x) = \frac{\cosh(x) \cdot \cosh(x) - \sinh(x) \cdot \sinh(x)}{\cosh^2(x)} = \frac{\cosh^2(x) - \sinh^2(x)}{\cosh^2(x)}. \tag{297}$$

Using the identity $\cosh^2(x) - \sinh^2(x) = 1$:

$$f'(x) = \frac{1}{\cosh^2(x)} = \operatorname{sech}^2(x). \tag{298}$$

Since $\cosh(x) \geq 1$, we have:

$$|f'(x)| = \operatorname{sech}^2(x). \tag{299}$$

Express the derivative in terms of $f(x)$:

$$f'(x) = \operatorname{sech}^2(x) = 1 - \tanh^2(x) = 1 - f(x)^2. \tag{300}$$

Since $-1 \leq f(x) \leq 1$, we have $|f'(x)| = 1 - f(x)^2 \leq 1$.

Maximize the Derivative. Maximize $|f'(x)| = \operatorname{sech}^2(x)$:

$$\cosh(x) = \frac{e^x + e^{-x}}{2}. \tag{301}$$

At $x = 0$:

$$\cosh(0) = \frac{e^0 + e^0}{2} = 1, \quad \operatorname{sech}^2(0) = \frac{1}{\cosh^2(0)} = 1. \tag{302}$$

As $|x| \to \infty$, $\cosh(x) \approx \frac{e^{|x|}}{2}$, so:

$$\operatorname{sech}^2(x) \approx \frac{4}{e^{2|x|}} \to 0. \tag{303}$$

The supremum of $|f'(x)|$ is 1 at $x = 0$. Alternatively, since $f(x)^2 \leq 1$, the supremum of $1 - f(x)^2$ occurs when $f(x) = 0$:

$$\tanh(0) = 0, \quad f'(0) = 1 - 0^2 = 1. \tag{304}$$

Hence:

$$\operatorname{Lip}[\tanh(x)] = \sup_x |f'(x)| = 1. \tag{305}$$

□

A.3 Proof: Lipschitz Constant of Softplus

Proof. Consider the softplus function:

$$f(x) = \ln(1 + e^x). \tag{306}$$

The derivative is:

$$f'(x) = \frac{d}{dx} \ln(1 + e^x) = \frac{e^x}{1 + e^x}. \tag{307}$$

Since $e^x > 0$ and $1 + e^x > 1$, we have $f'(x) > 0$, so:

$$|f'(x)| = \frac{e^x}{1 + e^x}. \tag{308}$$

Maximize the Derivative. To find the Lipschitz constant, we need to compute:

$$K = \sup_{x \in \mathbb{R}} \frac{e^x}{1 + e^x}. \tag{309}$$

Notice that $\frac{e^x}{1+e^x}$ is the sigmoid function, which ranges between 0 and 1. Analyze its behavior:

1. As $x \to \infty$, e^x grows large, so:

$$\frac{e^x}{1 + e^x} \approx \frac{e^x}{e^x} = 1. \tag{310}$$

2. As $x \to -\infty$, $e^x \to 0$, so:

$$\frac{e^x}{1 + e^x} \to \frac{0}{1} = 0. \tag{311}$$

To confirm the supremum, consider the function $g(x) = \frac{e^x}{1+e^x}$. Its derivative is:

$$g'(x) = \frac{e^x(1 + e^x) - e^x \cdot e^x}{(1 + e^x)^2} = \frac{e^x}{(1 + e^x)^2}. \tag{312}$$

Since $g'(x) > 0$ for all x, $g(x)$ is strictly increasing, approaching 0 as $x \to -\infty$ and 1 as $x \to \infty$. Thus:

$$\sup_{x \in \mathbb{R}} \frac{e^x}{1 + e^x} = 1 \tag{313}$$

Hence:

$$\text{Lip}\left[\text{Softplus}(x)\right] = \sup_x |f'(x)| = 1. \tag{314}$$

□

A.4 Proof: Lipschitz Constant of Swish

Proof. Let

$$f(x) = x\,\sigma(x), \quad \sigma(x) = \frac{1}{1 + e^{-x}}. \tag{315}$$

The derivative is

$$g(x) := f'(x) = \sigma(x) + x\,\sigma(x)\big(1 - \sigma(x)\big). \tag{316}$$

Since $\sigma(-x) = 1 - \sigma(x)$, we have

$$g(-x) = 1 - g(x). \tag{317}$$

Using $\sigma(x) = \frac{1}{2}\big(1 + \tanh\big(\frac{x}{2}\big)\big)$ and $\sigma(x)(1 - \sigma(x)) = \frac{1}{4}\,\mathrm{sech}^2(\frac{x}{2})$,

$$g(x) = \frac{1}{2}\Big(1 + \tanh\frac{x}{2}\Big) + \frac{x}{4}\,\mathrm{sech}^2\frac{x}{2}. \tag{318}$$

Differentiating,

$$g'(x) = \frac{1}{4}\,\mathrm{sech}^2\frac{x}{2}\,\Big(2 - x\,\tanh\frac{x}{2}\Big), \tag{319}$$

so the maximizer $x^\star > 0$ satisfies

$$x^\star\,\tanh\frac{x^\star}{2} = 2. \tag{320}$$

Using $\tanh\Big(\frac{x^\star}{2}\Big) = \frac{2}{x^\star}$ and

$$\mathrm{sech}^2\frac{x^\star}{2} = \frac{(x^\star)^2 - 4}{(x^\star)^2}. \tag{321}$$

Thus

$$K = g(x^\star) = \frac{1}{2} + \frac{x^\star}{4}. \tag{322}$$

Since $g(-x) = 1 - g(x)$, $x^\star$ gives the global maximum of $|g|$, solving $x^\star\,\tanh\frac{x^\star}{2} = 2$ gives:

$$x^\star \approx 2.3993572805\cdots. \tag{323}$$

Hence:

$$\mathrm{Lip}\,[\mathrm{Siwsh}(x)] = \sup_x |f'(x)| \approx 1.09983932\cdots \tag{324}$$

□

A.5 Proof: Lipschitz constant of GELU

Proof. Let:

$$f(x) = x\,\Phi(x), \tag{325}$$

where:

$$\Phi(x) = \tfrac{1}{2}\big(1 + \mathrm{erf}(x/\sqrt{2})\big) \tag{326}$$

is the standard normal CDF and:

$$\phi(x) = \tfrac{1}{\sqrt{2\pi}}e^{-x^2/2} \tag{327}$$

is its PDF.

Differentiate:

$$g(x) := f'(x) = \Phi(x) + x\,\phi(x). \tag{328}$$

Note the symmetry $\Phi(-x) = 1 - \Phi(x)$ and $\phi(-x) = \phi(x)$, hence

$$g(-x) = 1 - g(x). \tag{329}$$

Compute the critical points:

$$g'(x) = \phi(x) + \phi(x) + x\,\phi'(x) = 2\phi(x) - x^2\phi(x) = \phi(x)\,(2 - x^2). \tag{330}$$

Since $\phi(x) > 0$, we have $g'(x) > 0$ for $|x| < \sqrt{2}$ and $g'(x) < 0$ for $|x| > \sqrt{2}$. Hence g attains its unique global maximum at $x^\star = \sqrt{2}$ and minimum at $-\sqrt{2}$.

Since $g(-x) = 1 - g(x)$, the minimum equals $1 - g(\sqrt{2})$, so indeed $\sup_x |g(x)| = g(\sqrt{2})$.

Evaluate at:

$$x = \sqrt{2} \approx 1.414224\cdots, \tag{331}$$

hence:

$$\begin{aligned}
\mathrm{Lip}\,[\mathrm{GELU}(x)] &= \sup_x |f'(x)| && (332)\\
&= g(\sqrt{2}) && (333)\\
&= \Phi(\sqrt{2}) + \sqrt{2}\,\phi(\sqrt{2}) && (334)\\
&= \tfrac{1}{2}\big(1 + \mathrm{erf}(1)\big) + \frac{e^{-1}}{\sqrt{\pi}} && (335)\\
&\approx 1.128904145. && (336)
\end{aligned}$$

□

A.6 Proof: Lipschitz constant of Softmax

Proof. Let

$$\mathrm{softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, \tag{337}$$

and:

$$p := \mathrm{softmax}(z) \in \Delta^{n-1}, \tag{338}$$

where Δ^{n-1} is the probability simplex in $\mathbb{R}^n$ (*i.e.*, the set of all probability vectors of length n):

$$\Delta^{n-1} \;:=\; \big\{\, p \in \mathbb{R}^n \;\big|\; p_i \geq 0,\; \sum_{i=1}^{n} p_i = 1 \,\big\}. \tag{339}$$

The Jacobian is

$$J(z) = \nabla\,\mathrm{softmax}(z) = \mathrm{diag}(p) - p\,p^\top, \tag{340}$$

which is symmetric positive semidefinite (PSD) and satisfies:

$$J(z)\,\mathbf{1} = 0. \tag{341}$$

Hence the Lipschitz constant K is:

$$K = \sup_{z\in\mathbb{R}^n} \|J(z)\|_2 = \sup_{p\in\Delta^{n-1}} \lambda_{\max}\big(\mathrm{diag}(p) - p\,p^\top\big). \tag{342}$$

where $\lambda_{\max}$ is the largest singular value of $\mathrm{diag}(p) - p\,p^\top$.

For any unit vector $v \in \mathbb{R}^n$,

$$v^\top J(z)\,v = v^\top \mathrm{diag}(p)\,v - (p^\top v)^2 = \sum_{i=1}^n p_i v_i^2 - \Big(\sum_{i=1}^n p_i v_i\Big)^2 = \mathrm{Var}_{i\sim p}[v_i]. \tag{343}$$

Therefore

$$\|J(z)\|_2 = \sup_{\|v\|_2=1} \mathrm{Var}_{i\sim p}[v_i]. \tag{344}$$

By Popoviciu's inequality on variances, for any random variable X supported in $[a, b]$,

$$\mathrm{Var}[X] \le \frac{(b-a)^2}{4}, \tag{345}$$

with equality attained by a two–point distribution at the endpoints.

Applying this to $X = v_i$ under p, we get

$$\mathrm{Var}_{i\sim p}[v_i] \le \frac{\big(\max_i v_i - \min_i v_i\big)^2}{4}. \tag{346}$$

Maximizing the RHS over $\|v\|_2 = 1$ yields $\max_i v_i - \min_i v_i \le \sqrt{2}$, with equality for

$$v = \frac{1}{\sqrt{2}}(e_a - e_b) \quad \text{for some distinct } a, b, \tag{347}$$

so that

$$\sup_{\|v\|_2=1} \mathrm{Var}_{i\sim p}[v_i] \le \frac{(\sqrt{2})^2}{4} = \frac{1}{2}. \tag{348}$$

This upper bound is (arbitrarily) attainable by choosing p supported on the two indices a, b with $p_a = p_b = \frac{1}{2}$ and $p_i \to 0$ for $i \notin \{a, b\}$ (which is approached by softmax logits $z_a = z_b \gg z_{i\notin\{a,b\}}$). In that case,

$$J = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} \end{bmatrix} \oplus 0_{n-2}, \tag{349}$$

whose largest eigenvalue is $\frac{1}{2}$.

Hence:

$$\mathrm{Lip}\,[\mathrm{Softmax}(x)] = \sup_z \|J(z)\|_2 = \frac{1}{2}, \tag{350}$$

independent of $n \ge 2$.

□