

Lipschitz Constants of Functions of Neural Networks

Sunjoong Kim

Contents

1	Introduction	1
2	Lipschitz Constants of Functions Between Euclidean Spaces	2
2.1	Lipschitz Constants	2
2.2	Real-valued Functions of a Real Variable	4
2.3	Linear Functions	6
2.4	Affine Functions	13
2.5	Elementwise Application of Nonlinear Functions	14
2.6	Composites of Functions	15
3	Lipschitz Constants of Neural Networks	16
3.1	Rademacher Theorem	16
3.2	AutoLip : Upperbounding the Lipschitz constant by an algorithmic method	17
3.3	AutoGrad	22
4	Conclusion	23
	Bibliography	24

Abstract

Chapter 1

Introduction

Deep neural networks have made many advances including computer vision, language modeling, machine translation and text and picture generating. Still, one of the difficulties in applying deep learning algorithms to reality is that the algorithm often lacks its robustness. As an example, Szegedy et al. (2014) have found that a small perturbation of an input for true image may cause the classifier to misclasssify the image as false image.

To overcome the instabiltiy of training and to enhance robustness of generating models such as GANs, M. Arjovsky et al. (2017) proposed Wasserstein distance between distributions and restrict their attention to 1-Lipschitz function to the critic. As this example suggest, the Lipschitz constant can be a good metric to access the robustness of the algorithm.

The Lipschitz constant of a function measures the sensitiveness or the robustness, or the rate of changes of the function. In chapter 2, we define the optimal Lipschitz constant of the functions between euclidean spaces and explore the computation of this optimal constant for various functions including linear maps, affine maps and compositions of functions. However, if the function becomes more complex, it is difficult or almost impossible to calculate the optimal constant. So, we propose the algorithms to estimate the constant, using the Rademacher's theorem.

Chapter 2

Lipschitz Constants of Functions Between Euclidean Spaces

2.1 Lipschitz Constants

For vectors $\mathbf{x} = [x_1 \cdots x_n]^T$ and $\mathbf{y} = [y_1 \cdots y_m]^T$ in Euclidean spaces \mathbb{R}^n and \mathbb{R}^m , respectively, $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$ are the usual standard norms of \mathbf{x} and \mathbf{y} defined by

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}, \quad \|\mathbf{y}\| = \sqrt{y_1^2 + \cdots + y_m^2}. \quad (2.1.1)$$

Definition 2.1. A function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is called *Lipschitz continuous* if there is a nonnegative real number c such that

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq c\|\mathbf{x} - \mathbf{y}\| \quad (\mathbf{x}, \mathbf{y} \in \mathbb{R}^n) \quad (2.1.2)$$

The infimum of c which satisfies (2.1.2) is called the *Lipschitz constant* of f and is denoted by $\text{Lip}(f)$.

This definition can also be extended to functions from a metric space to another metric space if we replace the norms by the distances. It is useful to

know that there is an equivalent definition ;

$$\text{Lip}(f) = \sup_{\mathbf{x} \neq \mathbf{y}} \frac{\|f(\mathbf{x}) - f(\mathbf{y})\|}{\|\mathbf{x} - \mathbf{y}\|} \quad (2.1.3)$$

To prove the equality, let

$$\begin{aligned} L_1 &= \inf\{c \geq 0 : \|f(\mathbf{x}) - f(\mathbf{y})\| \leq c\|\mathbf{x} - \mathbf{y}\|, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n\} \\ L_2 &= \sup \left\{ \frac{\|f(\mathbf{x}) - f(\mathbf{y})\|}{\|\mathbf{x} - \mathbf{y}\|} : \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{y} \right\}. \end{aligned} \quad (2.1.4)$$

Let \mathcal{L}_1 and \mathcal{L}_2 be two the sets in the definition of L_1 and L_2 , respectively. Then, $\mathcal{L}_2 \subset \mathcal{L}_1$ since

$$\mathcal{L}_2 = \{c \geq 0 : \|f(\mathbf{x}) - f(\mathbf{y})\| = c\|\mathbf{x} - \mathbf{y}\|, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{y}\},$$

and it follows that $L_1 \leq L_2$. Pick $c \in \mathcal{L}_1$. Then, $\|f(\mathbf{x}) - f(\mathbf{y})\| \leq c\|\mathbf{x} - \mathbf{y}\|$ for all \mathbf{x} and \mathbf{y} in \mathbb{R}^n . Assuming $\mathbf{x} \neq \mathbf{y}$, we have

$$\frac{\|f(\mathbf{x}) - f(\mathbf{y})\|}{\|\mathbf{x} - \mathbf{y}\|} \leq c.$$

Taking supremum for all \mathbf{x} and \mathbf{y} ($\mathbf{x} \neq \mathbf{y}$), we have $L_2 \leq c$. and taking infimum for all $c \in \mathcal{L}_1$ yields $L_2 \leq L_1$.

The optimal Lipschitz constant $\text{Lip}(f)$ is not only the infimum, but also the minimum of L satisfying (2.1.2). To show this, it suffices to show that $\text{Lip}(f) \in \mathcal{L}_1$, or that $c = \text{Lip}(f)$ satisfies the condition. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. If $\mathbf{x} = \mathbf{y}$, then (2.1.2) holds trivially. Assuming $\mathbf{x} \neq \mathbf{y}$, we have

$$\frac{\|f(\mathbf{x}) - f(\mathbf{y})\|}{\|\mathbf{x} - \mathbf{y}\|} \leq \text{Lip}(f)$$

because of the second definition (2.1.3) of $\text{Lip}(f)$. Multiplying both sides by $\|\mathbf{x} - \mathbf{y}\|$, we can conclude that $c = \text{Lip}(f)$ satisfies the condition (2.1.2).

Note also that $\text{Lip}(f) = 0$ if and only if f is a constant function. Suppose that f is a constant function. Then, the condition (2.1.2) holds vacuously and $\mathcal{L}_1 = [0, \infty)$; $\text{Lip}(f) = 0$. Suppose, on the contrary, that $\text{Lip}(f) = 0$. Pick $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Since $0 \in \mathcal{L}_1$, we have $\|f(\mathbf{x}) - f(\mathbf{y})\| = 0$. Thus, $f(\mathbf{x}) = f(\mathbf{y})$ and f is a constant function.

2.2 Real-valued Functions of a Real Variable

For univariate real function $f : \mathbb{R} \rightarrow \mathbb{R}$, the above condition (2.1.2) is equivalent to saying that the average rate of change is upper bounded by c ;

$$\left| \frac{f(x) - f(y)}{x - y} \right| \leq c. \quad (2.2.5)$$

If f is differentiable everywhere in the domain, the mean value theorem guarantees that this is equivalent to

$$|f'(x)| \leq c \quad (2.2.6)$$

for every $x \in \mathbb{R}$. Moreover, $\text{Lip}(f)$ is the least nonnegative number c satisfying the condition (2.2.6) for all x and we can express $\text{Lip} f$ in its exact form ;

$$\text{Lip}(f) = \sup_{x \in \mathbb{R}} |f'(x)| \quad (2.2.7)$$

In this sense, we can easily evaluate the optimal Lipschitz constant of the sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2.2.8)$$

and the hyperbolic tangent function

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \quad (2.2.9)$$

Since $0 < \sigma(x) < 1$ and $\sigma'(x) = \sigma(x)(1 - \sigma(x))$, we have $\sigma'(x) < \frac{1}{4}$ so that

$$\text{Lip}(\sigma) = \frac{1}{4}. \quad (2.2.10)$$

Since $\tanh(x) = 2\sigma(2x) - 1$, we have

$$\tanh'(x) = 4\sigma'(2x) < 1$$

and

$$\text{Lip}(\tanh) = 1. \quad (2.2.11)$$

The ReLU function

$$\text{ReLU}(x) = \max\{0, x\} \quad (2.2.12)$$

is not differentiable everywhere, but we can use (2.2.5) to conclude

$$\text{Lip}(\text{ReLU}) = 1. \quad (2.2.13)$$

Here is a list of the optimal Lipschitz constants of univariate activation functions, frequently used in machine learning and deep learning.

In the definition of $\text{Lip}(f)$ in (2.1.3), we can't replace the supremum by the maximum. Consider a function $f(x) = \ln(1 + e^x)$. This function is called the softplus activation function and is an antiderivative of the sigmoid function. It is a monotonically increasing function whose derivative is also increasing, with an asymptote $y = x$. Thus, $\text{Lip}(f) = 1$, but no x and y satisfy the equality $\frac{f(x)-f(y)}{x-y} = 1$.

Table 2.1: The optimal Lipschitz constants of univariate activation functions. $\alpha = 0.1$ for Leaky ReLU and $\alpha = 1$ for ELU.

Activation Functions	Formula	Lip(f)
Sigmoid	$\sigma(x) = \frac{1}{1+e^{-x}}$	$\frac{1}{4}$
Hyperbolic tangent	$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	1
Rectified Linear Unit	$\text{ReLU}(x) = \max\{0, x\}$	1
Leaky ReLU	$\text{LReLU}(x) = \max\{\alpha x, x\}$	1
Exponential Linear Unit	$\text{ELU}(x) = \begin{cases} x & (x \geq 0) \\ \alpha(e^x - 1) & (x < 0) \end{cases}$	1
Softplus	$f(x) = \frac{1}{\beta} \log(1 + e^{\beta x})$	1
Gaussian	$g(x) = e^{-x^2}$	$\sqrt{\frac{2}{e}}$

2.3 Linear Functions

2.3.1 Operator Norms $\|W\|$

Consider a linear function $f(\mathbf{x}) = W\mathbf{x}$ for some $m \times n$ matrix W . Because of the linearity, the condition (2.1.2) reduces to

$$\|W\mathbf{x}\| \leq c\|\mathbf{x}\|. \quad (\mathbf{x} \in \mathbb{R}^n) \quad (2.3.14)$$

The smallest c which satisfies (2.3.14) is called *the operator norm of W* and denoted by $\|W\|$. Thus, the optimal Lipschitz constant of f equals the operator norm of W ; $\|W\| = \text{Lip}(f)$. The definitions of the forms (2.1.4) are as follows;

$$\begin{aligned} \|W\| &= \inf \{c \geq 0 : \|W\mathbf{x}\| \leq c\|\mathbf{x}\|\} \\ &= \sup \left\{ \frac{\|W\mathbf{x}\|}{\|\mathbf{x}\|} : \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0} \right\} \end{aligned} \quad (2.3.15)$$

If we modify the second definition of (2.3.15) to

$$\|W\| = \sup \{ \|W\mathbf{x}\| : \|\mathbf{x}\| = 1 \}, \quad (2.3.16)$$

we can easily verify that the supremum is actually the maximum in this case. Because the set $\{\mathbf{x} : \|\mathbf{x}\| = 1\}$ is a compact subset of \mathbb{R}^m and the map $\mathbf{x} \mapsto \|W\mathbf{x}\|$ is continuous, the set in (2.3.16) has its maximum.

The operator norm is indeed a norm of a vector space ;

Proposition 2.2. *The operator $\|\cdot\| : W \mapsto \|W\|$ satisfies the following three properties. Thus, the set $\mathcal{M}_{m,n}(\mathbb{R})$ of all m by n real matrices is a normed vector space with respect to this norm ;*

(a) $\|W\| \geq 0$ for all $W \in \mathcal{M}_{m,n}(\mathbb{R})$; $\|W\| = 0$ if and only if $W = 0$.

(b) $\|kW\| = |k| \cdot \|W\|$ for all $W \in \mathcal{M}_{m,n}(\mathbb{R})$ and $k \in \mathbb{R}$.

(c) $\|W + V\| \leq \|W\| + \|V\|$ for all $W, V \in \mathcal{M}_{m,n}(\mathbb{R})$.

For (a), that $\|W\| \geq 0$ is obvious. If $\|W\| = 0$, then $\|W\mathbf{x}\| \leq 0\|\mathbf{x}\| = 0$ for all $\mathbf{x} \in \mathbb{R}^n$. It follows that $W\mathbf{x} = 0$ for all $\mathbf{x} \in \mathbb{R}^n$, by the definition of vector norm, for which $W = 0$. Suppose, on the other hand, that $W = 0$. Then $W\mathbf{x} = 0$ for all $x \in \mathbb{R}^n$. Thus, $c = 0$ is the minimal value satisfying (2.3.14). To prove (b), we have

$$\begin{aligned} \|kW\| &= \inf \{ c : \|kWx\| \leq c\|x\| \} \\ &= \inf \{ c : |k| \cdot \|Wx\| \leq c\|x\| \} \\ &= \inf \left\{ c : \|Wx\| \leq \frac{c}{|k|} \|x\| \right\} \\ &= \inf \{ |k|b : \|Wx\| \leq b\|x\| \} \\ &= |k| \inf \{ b : \|Wx\| \leq b\|x\| \} \\ &= |k| \cdot \|W\|. \end{aligned}$$

Now, consider (c). For any $\mathbf{x} \in \mathbb{R}^n$, we have

$$\begin{aligned} \|(W + V)\mathbf{x}\| &= \|W\mathbf{x} + V\mathbf{x}\| \leq \|W\mathbf{x}\| + \|V\mathbf{x}\| \\ &\leq \|W\| \|\mathbf{x}\| + \|V\| \|\mathbf{x}\| = (\|W\| + \|V\|)\|\mathbf{x}\| \end{aligned}$$

By the minimality of $\|W + V\|$, we have $\|W + V\| \leq \|W\| + \|V\|$.

An analogue of (c) also holds for multiplication; if $W \in \mathcal{M}_{m,n}(\mathbb{R})$ and $V \in \mathcal{M}_{n,k}(\mathbb{R})$, then

$$\|WV\| \leq \|W\| \|V\|. \quad (2.3.17)$$

This is because of the inequality

$$\|(WV)\mathbf{x}\| = \|W(V\mathbf{x})\| \leq \|W\| \|V\mathbf{x}\| \leq \|W\| \|V\| \|\mathbf{x}\|$$

and the minimality of $\|WV\|$.

2.3.2 Evaluation of $\|W\|$ for square matrices

First, consider the case when W is a square matrix so that $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Let $\lambda_1, \dots, \lambda_n$ be (possibly repeated) eigenvalues of W and let $\tilde{\lambda} = \max\{|\lambda_i| : i = 1, \dots, n\}$. Then, the following inequality holds in general ;

$$\tilde{\lambda} \leq \|W\|. \quad (2.3.18)$$

For a proof, pick an eigenvalue λ_i and the corresponding eigenvector \mathbf{x}_i which is nonzero. Since $W\mathbf{x}_i = \lambda_i\mathbf{x}_i$, Then, we have $\|W\mathbf{x}_i\| = |\lambda_i| \|\mathbf{x}_i\|$ and

$$|\lambda_i| = \frac{\|W\mathbf{x}_i\|}{\|\mathbf{x}_i\|} \leq \|W\|.$$

Taking the maximum over i , we have the desired inequality. The above inequality (2.3.18) becomes equality when W is symmetric.

Suppose that W is real symmetric. Then, W is orthogonally diagonaliz-

able in the sense that, there exists an orthogonal matrix

$$V = [\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_n],$$

such that

$$W = V\Lambda V^T,$$

where $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_n\}$ and λ_i is an eigenvalue of W with the corresponding eigenvector \mathbf{v}_i . Note that $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ forms an orthonormal basis for \mathbb{R}^n . For any $\mathbf{x} \in \mathbb{R}^n$, there exist real numbers c_1, \dots, c_n such that $\mathbf{x} = c_1\mathbf{v}_1 + \dots + c_n\mathbf{v}_n$. Since

$$\begin{aligned} W\mathbf{x} &= c_1W\mathbf{v}_1 + \dots + c_nW\mathbf{v}_n \\ &= c_1\lambda_1\mathbf{v}_1 + \dots + c_n\lambda_n\mathbf{v}_n, \end{aligned}$$

we have

$$\begin{aligned} \|W\mathbf{x}\|^2 &= \|c_1W\mathbf{v}_1 + \dots + c_nW\mathbf{v}_n\|^2 \\ &= |c_1\lambda_1|^2\|\mathbf{v}_1\|^2 + \dots + |c_n\lambda_n|^2\|\mathbf{v}_n\|^2 \\ &= c_1^2\lambda_1^2 + \dots + c_n^2\lambda_n^2. \end{aligned}$$

Thus,

$$\frac{\|W\mathbf{x}\|^2}{\|\mathbf{x}\|^2} = \frac{c_1^2\lambda_1^2 + \dots + c_n^2\lambda_n^2}{c_1^2 + \dots + c_n^2} \leq \tilde{\lambda}^2,$$

for which $\|W\mathbf{x}\|/\|\mathbf{x}\| \leq \tilde{\lambda}$. Therefore, we have the reverse inequality $\|W\| \leq \tilde{\lambda}$.

As examples consider the following matrices W_1, W_2, W_3, W_4, W_5 defined

by

$$W_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad W_2 = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}, \quad W_3 = \begin{bmatrix} 4 & 2 \\ 2 & 7 \end{bmatrix},$$

$$W_4 = \begin{bmatrix} 1 & 3 \\ 2 & 0 \end{bmatrix}, \quad W_5 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

W_1 is the identity matrix and thus $||W|| = 1$. Indeed, the characteristic polynomial of W_1 is $(\lambda - 1)^2 = 0$, $\lambda_1 = \lambda_2 = 1$ and thus the maximum absolute value of them is 1. The matrix W_2 stretches a given vector \mathbf{x} twice in the x-axis direction and three times in the y-axis direction. Thus, $||W_2||$ should be 3. Indeed, $|W_2 - \lambda I| = (\lambda - 2)(\lambda - 3)$ and $\tilde{\lambda} = 3$. W_3 is the last example that is symmetric ; $|W_3 - \lambda I| = (\lambda - 3)(\lambda - 8)$ and $\tilde{\lambda} = 8$.

W_4 is not symmetric and we can't evaluate $||W_4||$ for now. Still, we have, $\lambda_1 = -2$, $\lambda_2 = 3$ and $||W_4|| \geq 3$ by (2.3.18). The reverse inequality is not valid since the eigenvectors $x_1 = [1 \ -1]^T$ and $x_2 = [3 \ 2]^T$ are not perpendicular. W_5 is not symmetric either, but we can postulate that $||W_5|| = 1$ since

$$W_5 \mathbf{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ 0 \end{bmatrix}$$

and

$$\frac{||W_5 \mathbf{x}||}{||\mathbf{x}||} = \frac{|x_2|}{\sqrt{x_1^2 + x_2^2}} \leq 1.$$

Since W_5 has the only eigenvector 0, the inequality (??) still holds.

Here is a short description of what we've done in this subsection.

Lemma 2.3. *Let $W \in \mathcal{M}_n(\mathbb{R})$ be a real symmetric matrix. Then, the operator norm $||W||$ is the maximal absolute eigenvalue of W .*

2.3.3 Evaluation of $\|W\|$ for rectangular matrices

Almost all matrices that we encounter in machine learning problem are neither square matrices, nor are they symmetric matrices. But, we can always calculate the Lipschitz constant or the operator norm of any rectangular matrices. Aside from the properties listed in the Proposition 2.2, there is a useful criteria that help find the optimal Lipschitz constant of any rectangular matrix.

Theorem 2.4. *Let $W \in \mathcal{M}_{m,n}(\mathbb{R})$ be an m by n real matrix. Then*

$$\|W\| = \sqrt{\|W^T W\|}. \quad (2.3.19)$$

Note that the norm on the left hand side is the operator norm on $\mathcal{M}_{m,n}(\mathbb{R})$ and the norm on the right hands side is the operator norm on $\mathcal{M}_n(\mathbb{R})$. Note also that the matrix $W^T W$ is symmetric in that $(W^T W)^T = W^T (W^T)^T = W^T W$ and we are always able to calculate the norm of $\|W^T W\|$ by making use of the lemma 2.3.

Before the proof of (2.3.19), we illustrate elementary properties of the operator norm of matrices. The euclidean norm $\|\mathbf{x}\|$ of a vector $\mathbf{x} \in \mathbb{R}^n$ in (2.1.1) can also be defined by means of the inner product on \mathbb{R}^n ;

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}. \quad (2.3.20)$$

The inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$ has the following properties and we omit the proofs of them.

$$\begin{aligned} \langle \mathbf{x}, W\mathbf{y} \rangle &= \langle W^T \mathbf{x}, \mathbf{y} \rangle, \\ \langle k\mathbf{x}, \mathbf{y} \rangle &= k\langle \mathbf{x}, \mathbf{y} \rangle, \\ |\langle \mathbf{x}, \mathbf{y} \rangle| &\leq \|\mathbf{x}\| \|\mathbf{y}\|. \end{aligned}$$

Now we are ready for the proof of (2.3.19). If $W = 0$, then the equation

(2.3.19) holds trivially. Suppose that $W \neq 0$. Let $\mathbf{x} \in \mathbb{R}^n$. Then

$$\begin{aligned} \|W\mathbf{x}\|^2 &= \langle W\mathbf{x}, W\mathbf{x} \rangle \\ &= \langle W^T W\mathbf{x}, \mathbf{x} \rangle \\ &\leq \|W^T W\mathbf{x}\| \|\mathbf{x}\| \\ &\leq \|W^T W\| \|\mathbf{x}\|^2, \end{aligned}$$

and thus

$$\|W\mathbf{x}\| \leq \sqrt{\|W^T W\|} \|\mathbf{x}\|.$$

for all $\mathbf{x} \in \mathbb{R}^n$. By the minimality of $\|W\|$, we have

$$\|W\| \leq \sqrt{\|W^T W\|}$$

Squaring both sides and making use of (2.3.17) yield

$$\|W\|^2 = \|W^T W\| \leq \|W^T\| \|W\|. \quad (2.3.21)$$

Since $\|W\| \neq 0$, we have

$$\|W\| \leq \|W^T\|.$$

Substituting W by its transpose, we get the reverse inequality. Therefore,

$$\|W\| = \|W^T\|. \quad (2.3.22)$$

By (2.3.22), the equation (2.3.21) reduces to

$$\|W\|^2 = \|W^T W\|,$$

and this proves (2.3.19).

For example, we can evaluate the operator norm of W_4 and W_5 in the

previous subsection. We have

$$W_4^T W_4 = \begin{bmatrix} 1 & 3 \\ 2 & 0 \end{bmatrix}, \quad W_5^T W_5 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

$W_4^T W_4$ has its eigenvalues $7 \pm \sqrt{13}$, thus $\|W_4\| = \sqrt{7 + \sqrt{13}}$. And $W_5^T W_5$ has eigenvalues 0 and 1 ; $\|W_5\| = \sqrt{1} = 1$ as expected. As a rectangular matrix, we may think of

$$W_6 = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 0 & 1 \end{bmatrix}.$$

Since

$$W_6^T W_6 = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix},$$

$\lambda = 1, 6$ and $\|W_6\| = \sqrt{6}$.

2.4 Affine Functions

An affine function

$$f(\mathbf{x}) = W\mathbf{x} + \mathbf{b} \tag{2.4.23}$$

has the same optimal Lipschitz constant as that of its linear part ;

$$\text{Lip}(f) = \|W\| \tag{2.4.24}$$

This is because the condition (2.1.2)

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq c\|\mathbf{x} - \mathbf{y}\|$$

applied to (2.4.23) becomes

$$||W(\mathbf{x} - \mathbf{y})|| \leq c||\mathbf{x} - \mathbf{y}||.$$

2.5 Elementwise Application of Nonlinear Functions

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be any function and let $F : \mathbb{R}^m \rightarrow \mathbb{R}^m$ be applying f to each element of the input $\mathbf{x} = [x_1 \cdots x_m]^T$:

$$F(\mathbf{x}) = F\left(\begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}\right) = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_m) \end{bmatrix}. \quad (2.5.25)$$

If f is a Lipschitz continuous function with its optimal constant L . Then,

$$\begin{aligned} ||F(\mathbf{x}) - F(\mathbf{y})||^2 &= \left\| \begin{bmatrix} f(x_1) - f(y_1) \\ \vdots \\ f(x_m) - f(y_m) \end{bmatrix} \right\|^2 = \sum_{i=1}^m |f(x_i) - f(y_i)|^2 \\ &\leq \sum_{i=1}^m L^2 |x_i - y_i|^2 = L^2 ||\mathbf{x} - \mathbf{y}||^2, \end{aligned}$$

for which

$$||F(\mathbf{x}) - F(\mathbf{y})|| \leq L||\mathbf{x} - \mathbf{y}||.$$

Thus, $\text{Lip}(F)$ is upper bounded by $L = \text{Lip}(f)$;

$$\text{Lip}(F) \leq \text{Lip}(f). \quad (2.5.26)$$

2.6 Composites of Functions

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^m \rightarrow \mathbb{R}^k$ be Lipschitz continuous functions. Then, the composite function $g \circ f$ is also Lipschitz continuous and the optimal constant is bounded by the product of those of f and g ;

$$\text{Lip}(g \circ f) \leq \text{Lip}(f)\text{Lip}(g) \quad (2.6.27)$$

This is readily proved by the following reason. Let L and M be the optimal Lipschitz constants for f and g respectively. Then, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\begin{aligned} \|(g \circ f)(\mathbf{x}) - (g \circ f)(\mathbf{y})\| &= \|g(f(\mathbf{x})) - g(f(\mathbf{y}))\| \\ &\leq M\|f(\mathbf{x}) - f(\mathbf{y})\| \\ &\leq LM\|\mathbf{x} - \mathbf{y}\|. \end{aligned}$$

By the minimality of $\text{Lip}(g \circ f)$, we have (2.6.27).

Note that (2.3.17) is a specific case of (2.6.27). Here, the inequality can be strict. As an example, $\|W_2\|\|W_4\| = 3 \times \sqrt{7 + \sqrt{13}} > \sqrt{1 + \sqrt{37}} = \|W_2 W_4\|$. Note also that (2.6.27) can be generalized further, to the case when the composition involves several functions. That is, if $f_i : \mathbb{R}^{n_{i-1}} \rightarrow \mathbb{R}^{n_i}$ are functions between euclidean spaces, with its optimal Lipschitz constant L_i , for $i = 1, 2, \dots, p$, the composite

$$f_p \circ \dots \circ f_1 : \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_{p+1}}$$

is Lipschitz continuous and

$$\text{Lip}(f_p \circ \dots \circ f_1) \leq \prod_{i=1}^p L_i. \quad (2.6.28)$$

Chapter 3

Lipschitz Constants of Neural Networks

Now we turn to the problem of finding the optimal Lipschitz constant of various architectures frequently appearing in machine learning and deep learning.

For locally Lipschitz function, it is possible to express the Lipschitz constant of f in its exact form, as the Rademacher Theorem suggests. But, it is not always easy or feasible to find the exact value of Lipschitz constant whenever the function f is given. In fact, although the function is of simple form such as 2-layered MLP, the problem of evaluating the Lipschitz function is NP-hard. Instead of struggling to find analytic solutions, we present a systematic algorithm for each architecture of neural network.

3.1 Rademacher Theorem

Here is a theorem that can be applied to all Lipschitz functions between euclidean spaces. The Lipschitz condition that we impose is not the global one (2.1.2) ; it only need to be locally Lipschitz. It has two conclusions : the differentiability (a.e.) and the formula for the optimal constant.

Theorem 3.1. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a locally Lipschitz function. Then, f is differentiable almost everywhere, and*

$$\text{Lip}(f) = \sup_{\mathbf{x} \in \mathbb{R}^n} \|D_{\mathbf{x}}(f)\| \quad (3.1.1)$$

The first statement — differentiability — is usually referred to the main conclusion of this theorem. But the proof of it is so lengthy that we omit the proof here. As for one-dimensional case, many of the textbooks of analysis including [8] covered this issue. In the process, concepts like absolute continuity, Radon-Nykodym theorem, total variation are involved.

The second statement, or the optimal Lipschitz constant is the one that we can use for our purpose. Note first that, $D_{\mathbf{x}}f$ is the Jacobian matrix of f or a matrix A satisfying

$$\lim_{\mathbf{y} \rightarrow \mathbf{x}} \frac{\|f(\mathbf{y}) - f(\mathbf{x}) - A(\mathbf{y} - \mathbf{x})\|}{\|\mathbf{y} - \mathbf{x}\|} = 0,$$

and that $\|D_{\mathbf{x}}(f)\|$ is nothing but the operator norm we've defined in the previous chapter.

3.2 AutoLip : Upperbounding the Lipschitz constant by an algorithmic method

Here is an algorithm called *AutoLip* that evaluate an upperbound for the Lipschitz constant. The algorithm can be applied to any deterministic (resp. stochastic, e.g. bayesian neural networks) neural network with the computation graph, which is a composite of elementary functions.

The only difficulty in the Rademacher theorem is the presense of the supremum ; it is not easy or sometimes impossible to evaluate the exact value of the supremum or the maximum of a function. But, it is possible, in most case, to find the maximum in the relatively simple function. In other

words, the supremum for the Jacobian matrix can be calculated for most elementary functions.

For a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we've used the notation for the jacobian matrix of f as

$$D_{\mathbf{x}}(f)$$

in the previous section. In this section, we occasionally use

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$$

to represent the same matrix, provided that $\mathbf{y} = f(\mathbf{x})$. In this context, $\|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\|$ represents the operator norm for the matrix. Furthermore, if y is a function of a univariate variable x , we usually denote the ordinary derivative of y with respect to x as $\frac{dy}{dx}$. But, in order to unify the notation, we employ $\frac{\partial y}{\partial x}$.

3.2.1 An example for univariate case

Consider, for example, a univariate function f_{ω} with a parameter ω defined by

$$f_{\omega}(x) = \frac{1}{1 + e^{-x}} + |2x + \omega \cos x|.$$

In order to construct a systematic way to estimate the constant and to make use of automatic differentiation in implementing the backpropagation by Tensorflow or Pytorch, we think of a computation graph for f_{ω} . (Figure 3.1)

We can bound the Lipschitz constant of f_{ω} from above, by upperbounding the partial derivatives of $\theta_1, \theta_2, \dots, \theta_7$ consecutively where $\theta_7 = f_{\omega}(x)$. Since $\theta_1 = g_1(\theta_0) = \omega$, we have $\frac{\partial \theta_1}{\partial x} = 0$. The next parameter θ_2 is a function of two variables ; $\theta_2 = g_2(\theta_0, \theta_1) = \cos \theta_0$. It is legal to apply the chain rule to get

$$\frac{\partial \theta_2}{\partial x} = \frac{\partial \theta_2}{\partial \theta_0} \frac{\partial \theta_0}{\partial x} + \frac{\partial \theta_2}{\partial \theta_1} \frac{\partial \theta_1}{\partial x} = -\sin x$$

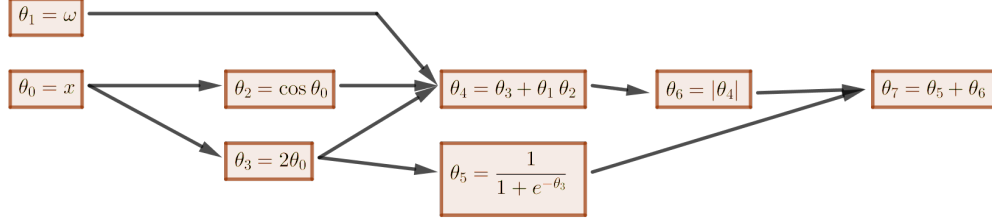


Figure 3.1: A computational graph for the function $f_\omega(x) = \frac{1}{1+e^{-2x}} + |2x + \omega \cos x|$.

Then, the *operator norm* or the Lipschitz constant $\left\| \frac{\partial \theta_2}{\partial x} \right\|$ of θ_2 is

$$\left\| \frac{\partial \theta_2}{\partial x} \right\| = \sup_x (-\sin x) = 1.$$

Again, since $\theta_3 = g_3(\theta_0, \theta_1, \theta_2) = 2\theta_0$, we have

$$\frac{\partial \theta_3}{\partial x} = \frac{\partial \theta_3}{\partial \theta_0} \frac{\partial \theta_0}{\partial x} + \frac{\partial \theta_3}{\partial \theta_1} \frac{\partial \theta_1}{\partial x} + \frac{\partial \theta_3}{\partial \theta_2} \frac{\partial \theta_2}{\partial x} = 2,$$

and thus $\left\| \frac{\partial \theta_3}{\partial x} \right\| = 2$. Note that the value of partial derivatives evaluated above, such as $\frac{\partial \theta_1}{\partial x}$ and $\frac{\partial \theta_2}{\partial x}$, can be applied to calculate $\frac{\partial \theta_3}{\partial x}$. To represent that $\left\| \frac{\partial \theta_3}{\partial x} \right\|$ is upper bounded by 2, we write $L_3 = 2$. Likewise, let $L_1 = 0$ and $L_2 = 1$.

For the next parameter $\theta_4 = g_4(\theta_0, \theta_1, \theta_2, \theta_3) = \theta_3 + \theta_1 \theta_2$, we have

$$\begin{aligned} \frac{\partial \theta_4}{\partial x} &= \frac{\partial \theta_4}{\partial \theta_0} \frac{\partial \theta_0}{\partial x} + \frac{\partial \theta_4}{\partial \theta_1} \frac{\partial \theta_1}{\partial x} + \frac{\partial \theta_4}{\partial \theta_2} \frac{\partial \theta_2}{\partial x} + \frac{\partial \theta_4}{\partial \theta_3} \frac{\partial \theta_3}{\partial x} \\ &= 0 \times 1 + \theta_2 \times 0 + \theta_1 \times (-\sin x) + 1 \times 2 \end{aligned}$$

and

$$\begin{aligned} \left\| \frac{\partial \theta_4}{\partial x} \right\| &\leq |\omega| \cdot \left\| \frac{\partial \theta_2}{\partial x} \right\| + 1 \times \left\| \frac{\partial \theta_3}{\partial x} \right\| \\ &\leq |\omega| \cdot L_2 + L_3 = |\omega| + 2. \end{aligned}$$

Thus, $L_4 = |\omega| + 2$.

In similar fashions, $\theta_5 = g_5(\theta_0, \theta_1, \dots, \theta_4) = \frac{1}{1+e^{-\theta_3}}$, $\theta_6 = g_6(\theta_0, \theta_1, \dots, \theta_5) = |\theta_4|$ and $\theta_7 = g_7(\theta_0, \theta_1, \dots, \theta_6) = \theta_5 + \theta_6$ have

$$\begin{aligned} \frac{\partial \theta_5}{\partial x} &= \frac{\partial \theta_5}{\partial \theta_3} \frac{\partial \theta_3}{\partial x} \times 2 & \left\| \frac{\partial \theta_5}{\partial x} \right\| &\leq \frac{1}{4} \times 2 = \frac{1}{2} \\ \frac{\partial \theta_6}{\partial x} &= \frac{\partial \theta_6}{\partial \theta_4} \frac{\partial \theta_4}{\partial x} & \left\| \frac{\partial \theta_6}{\partial x} \right\| &\leq \left\| \frac{\partial \theta_6}{\partial \theta_4} \right\| \left\| \frac{\partial \theta_4}{\partial x} \right\| = |\omega| + 2 \\ \frac{\partial \theta_7}{\partial x} &= \frac{\partial \theta_7}{\partial \theta_5} \frac{\partial \theta_5}{\partial x} + \frac{\partial \theta_7}{\partial \theta_6} \frac{\partial \theta_6}{\partial x} & \left\| \frac{\partial \theta_7}{\partial x} \right\| &\leq \left\| \frac{\partial \theta_5}{\partial x} \right\| + \left\| \frac{\partial \theta_6}{\partial x} \right\| \leq |\omega| + \frac{5}{2}. \end{aligned}$$

Thus, the Lipschitz constant for the function f_ω is upper bounded by $\omega + \frac{5}{2}$.

3.2.2 AutoLip

Consider $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, a composite function of elementary functions $g_0, g_1, g_2, \dots, g_K$, so that

$$f = g_K \circ \dots \circ g_2 \circ g_1 \circ g_0.$$

For every $\mathbf{x} \in \mathbb{R}^n$, define the intermediate variables $\theta_0 = g_0(\mathbf{x})$, $\theta_1 = g_1(g_0(\mathbf{x}))$, \dots , $\theta_K = g_K(g_{K-1}(\dots g_1(g_0(\mathbf{x})))) = f(\mathbf{x})$. For a notational purpose, let g_0 be the identity function and think of g_k as a function of θ_0, θ_1 ,

\dots, θ_{k-1} for $1 \leq k \leq K$ and write

$$\begin{aligned}\theta_1 &= g_1(\theta_0) \\ \theta_2 &= g_2(\theta_0, \theta_1) \\ &\vdots \\ \theta_K &= g_K(\theta_0, \theta_1, \dots, \theta_{K-1}).\end{aligned}$$

An upper bound L_k for the lipschitz constant $\sup_{\mathbf{x}} \left\| \frac{\partial \theta_k}{\partial \mathbf{x}} \right\|$ of θ_k , can be evaluated inductively, where L_K is an upper bound for $\text{Lip}(f)$.

Let $L_0 = 1$. Suppose that k is an integer such that $1 \leq k \leq K$ and we have an upper bound L_i of the Lipschitz constant

$$\sup_{\mathbf{x}} \left\| \frac{\partial \theta_i}{\partial \mathbf{x}} \right\| \leq L_i$$

for $1 \leq i \leq k-1$. Since

$$\begin{aligned}\frac{\partial \theta_k}{\partial \mathbf{x}} &= \frac{\partial \theta_k}{\partial \theta_0} \frac{\partial \theta_0}{\partial \mathbf{x}} + \frac{\partial \theta_k}{\partial \theta_1} \frac{\partial \theta_1}{\partial \mathbf{x}} + \dots + \frac{\partial \theta_k}{\partial \theta_{k-1}} \frac{\partial \theta_{k-1}}{\partial \mathbf{x}} \\ &= \sum_{i=1}^{k-1} \frac{\partial \theta_k}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mathbf{x}},\end{aligned}$$

it follows that

$$\left\| \frac{\partial \theta_k}{\partial \mathbf{x}} \right\| \leq \sum_{i=1}^{k-1} \left\| \frac{\partial \theta_k}{\partial \theta_i} \right\| \left\| \frac{\partial \theta_i}{\partial \mathbf{x}} \right\|$$

and that

$$\sup_{\mathbf{x}} \left\| \frac{\partial \theta_k}{\partial \mathbf{x}} \right\| \leq \sum_{i=1}^{k-1} \max \left\| \frac{\partial \theta_k}{\partial \theta_i} \right\| \cdot L_i.$$

Let L_k be the left hand side of the above inequality and proceed for $k+1$.

After K iterations, we have an upper bound L_K of the lipschitz constant for $\theta_K = f(\mathbf{x})$. (**Algorithm 1**)

Algorithm 1: AutoLip	
<p>Input: the function $f = g_K \circ \dots \circ g_1$, where $\theta_k = (g_k \circ \dots \circ g_1)(\mathbf{x})$ for $1 \leq k \leq K$</p> <p>Output: An upper bound L for $\text{Lip}(f)$</p> <pre> 1 $k \leftarrow 1$; 2 $L_1 \leftarrow 1$; 3 while $k < K$ do 4 $L_k \leftarrow \sum_{i=1}^{k-1} \left(\max \left\ \frac{\partial \theta_k}{\partial \theta_i} \right\ \right) L_i$; 5 $k \leftarrow k + 1$; 6 end 7 $L \leftarrow L_K$</pre>	

3.2.3 AutoLip applied to 2-layered-perceptron

3.3 AutoGrad

Chapter 4

Conclusion

Bibliography

- [1] Scaman, K. (2018, May 28). Lipschitz regularity of deep neural networks: analysis and. . . arXiv.org. Retrieved September 19, 2022, from <https://arxiv.org/abs/1805.10965>
- [2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.
- [3] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, pages 214–223, 2017.
- [4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
- [5] Francis H. Clarke. *On the Inverse function Theorem*, Pacific Journal of Mathematics, Vol 64, No. 1.
- [6] Matt Jordan and Alexandros G. Dimakis *Exactly Computing the Local Lipschitz Constant of ReLUNetworks*
- [7] Bhowmick, Aritra & D’Souza, Meenakshi & Raghavan, G.. (2021). LipBaB: Computing Exact Lipschitz Constant of ReLU Networks. 10.1007/978-3-030-86380-7_13.

- [8] Rudin, Walter. *Real and Complex Analysis.* : McGraw-Hill Science/Engineering/Math, 1986.
- [9] Herbert Federer. *Geometric measure theory.* Classics in Mathematics. Springer-Verlag Berlin Heidelberg, 1969.