Lipschitz Constants of Functions of Neural Networks

Sunjoong Kim

Contents

Abstract

Chapter 1

Introduction

(Lipschitz constant에 대한 간략한 설명 및 neural network과의 관계)

neural network의 Lipschitz constant는 그 neural network의 안정성(robustness)과 관련되어 있다. Lipschitz constant를 사용해 성능의 개선을 이룬 수많은 사례들이 있는데, 그 중 가장 대표적인 예는 Wasserstein GAN이다.

Lipschitz constant는 크게 global Lipschitz constant와 local Lipschitz constant로 나뉜다. local Lipschitz constant에 대해 다룬 논문들도 있지만, 이 논문에서는 global한 경우에 대해서만 다루겠다.

아주 간단한 neural network라고 하더라도 optimal Lipschitz constant를 구하는 것이 계산불가능할 수 있음이 알려져있다. (N-P hard) 따라서, optimal Lipschitz constant를 정확히 계산하는 것 보다도 optimal Lipschitz constant의 upper bound찾고, 그 upper bound를 계속해서 개선해나가는 알고리즘들이 알려져있다. 그것들이 3장에서 다루어질 것이다.

2장에서는 유클리드 공간 사이에 정의된 함수에 대한 Lipschitz에 대한 일반적인 사항들을 다룬다. 여기서는 일반적인 함수들에 대해서 다루기도 하지만, 특별히 머신러닝과 관련된 여러 함수들, affine function과 activation function들에 대해서도 다룬다. 3장에서는 우선, 2장에서 서술한 기본적인 원리들을 가지고 MLP의 Lipschitz constant가 어떻게 서술될 수 있는지 다룬다. 이어서 CNN과 RNN과 같은 조금 복잡한 구조에서는 Lipschitz constant가 어떻게 계산될 수 있을지 다룬다.

Chapter 2

Lipschitz Constants of Functions Between Euclidean Spaces

2.1 Lipschitz Constants

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

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

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

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

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

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 ;

$$\mathrm{Lip}(f) = \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{||f(\boldsymbol{x}) - f(\boldsymbol{y})||}{||\boldsymbol{x} - \boldsymbol{y}||} \tag{2.1.3}$$

To prove the equality, let

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

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

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

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

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

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

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

$$
\frac{||f(\boldsymbol{x}) - f(\boldsymbol{y})||}{||\boldsymbol{x} - \boldsymbol{y}||} \leq \mathrm{Lip}(f)
$$

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

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

2.2 Real-valued Functions of a Real Variable

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

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

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

$$|f'(x)| \leq c \tag{2.2.6}$$

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

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

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

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

and the hyperbolic tangent function

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

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

$$\mathrm{Lip}(\sigma) = \frac{1}{4}. \tag{2.2.10}$$

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

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

and

$$\mathrm{Lip}(\tanh) = 1. \tag{2.2.11}$$

The ReLU function

$$\mathrm{ReLU}(x) = \max\{0, x\} \tag{2.2.12}$$

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

$$\mathrm{Lip}(\mathrm{ReLU}) = 1. \tag{2.2.13}$$

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

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	$\mathrm{ReLU}(x) = \max\{0, x\}$	1
Leaky ReLU	$\mathrm{LReLU}(x) = \max\{\alpha x, x\}$	1
Exponential Linear Unit	$\mathrm{ELU}(x) = \begin{cases} x & (x \geq 0) \\ \alpha(e^x - 1) & (x < 0) \end{cases}$	1
Softplus	$f(x) = \frac{1}{\beta}\log(1 + e^{\beta x})$	1
Gaussian	$g(x) = e^{-x^2}$	$\sqrt{\frac{2}{e}}$

In the definition of 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 assymptote $y = x$. Thus, Lip(f) = 1, but no x and y satisfy the equality $\frac{f(x)-f(y)}{x-y} = 1$.

2.3 Linear Functions

2.3.1 Operator Norms $||W||$

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

$$||W\boldsymbol{x}|| \leq c||\boldsymbol{x}||. \qquad (\boldsymbol{x} \in \mathbb{R}^n) \tag{2.3.14}$$

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

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

If we modify the second definition of (2.3.15) to

$$||W|| = \sup\{||W\boldsymbol{x}|| : ||\boldsymbol{x}|| = 1\}, \tag{2.3.16}$$

we can easily verify that the supremum is actually the maximum in this case. Because the set $\{\boldsymbol{x} : ||\boldsymbol{x}|| = 1\}$ is a compact subset of $\mathbb{R}^m$ and the map $\boldsymbol{x} \mapsto ||W\boldsymbol{x}||$ 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 $||W\boldsymbol{x}|| \leq 0||\boldsymbol{x}|| = 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$. It follows that $W\boldsymbol{x} = 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$, by the definition of vector norm, for which $W = 0$ Suppose, on the other hand, that $W = 0$.

Then $W\boldsymbol{x} = 0$ 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 $\boldsymbol{x} \in \mathbb{R}^n$, we have

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

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

$$||WV|| \leq ||W||\,||V||. \tag{2.3.17}$$

This is because of the inequality

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

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 $f : \mathbb{R}^n \to \mathbb{R}^n$. Let $\lambda_1, \cdots, \lambda_n$ be (possibly repeated) eigenvalues of W and let $\tilde{\lambda} = \max\{|\lambda_i| : i = 1, \cdots, n\}$. Then, the following inequality holds in general ;

$$\tilde{\lambda} \leq ||W||. \tag{2.3.18}$$

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

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

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 diagonalizable in the sense that, there exists an orthogonal matrix

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

such that

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

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

$$\begin{aligned} W\boldsymbol{x} &= c_1 W\boldsymbol{v}_1 + \cdots + c_n W\boldsymbol{v}_n \\ &= c_1\lambda_1\boldsymbol{v}_1 + \cdots + c_n\lambda_n\boldsymbol{v}_n, \end{aligned}$$

we have

$$\begin{aligned} ||W\boldsymbol{x}||^2 &= ||c_1 W\boldsymbol{v}_1 + \cdots + c_n W\boldsymbol{v}_n||^2 \\ &= |c_1\lambda_1|^2||\boldsymbol{v}_1||^2 + \cdots + |c_n\lambda_n|^2||\boldsymbol{v}_n||^2 \\ &= {c_1}^2{\lambda_1}^2 + \cdots + {c_n}^2{\lambda_n}^2. \end{aligned}$$

Thus,

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

for which $||W\boldsymbol{x}||/||\boldsymbol{x}|| \leq \tilde{\lambda}$. 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 polyinomial 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 $\boldsymbol{x}$ 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

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

and

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

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*

$$||W|| = \sqrt{||W^T W||}. \tag{2.3.19}$$

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 $||\boldsymbol{x}||$ of a vector $\boldsymbol{x} \in \mathbb{R}^n$ in (2.1.1) can also be defined by means of the inner product on $\mathbb{R}^n$;

$$||\boldsymbol{x}|| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}. \tag{2.3.20}$$

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

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

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 $\boldsymbol{x} \in \mathbb{R}^n$. Then

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

and thus

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

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

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

Squaring both sides and making use of (2.3.17) yield

$$||W||^2 = ||W^T W|| \leq ||W^T||\,||W||. \tag{2.3.21}$$

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

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

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

$$||W|| = ||W^T||. \tag{2.3.22}$$

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

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

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

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

This is because the condition (2.1.2)

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

applied to (2.4.23) becomes

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

2.5 Elementwise Application of Nonlinear Functions

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

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

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

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

for which

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

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

$$\mathrm{Lip}(F) \leq \mathrm{Lip}(f). \tag{2.5.26}$$

2.6 Composites of Functions

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

$$\mathrm{Lip}(g \circ f) \leq \mathrm{Lip}(f)\mathrm{Lip}(g) \tag{2.6.27}$$

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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

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

By the minimality of $\mathrm{Lip}(g \circ f)$, 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 involoves several functions. That is, if $f_i : \mathbb{R}^{n_{i-1}} \to \mathbb{R}^{n_i}$ are functions between euclidean spaces, with its optimal Lipschitz constant L_i, for $i = 1, 2, \cdots, p$, the composite

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

is Lipschitz constinuous and

$$\mathrm{Lip}(f_p \circ \cdots \circ f_1) \leq \prod_{i=1}^{k} L_i. \tag{2.6.28}$$

Chapter 3

Lipschitz Constants of Neural Networks

Now we turn to the problem of finding the optimal Lischitz 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 true value of the Lipschitz constant whenever the function f is given. Indeed, even for simple algorithm such as two-layered MLP, it takes beyond polynomial time to evaluate the optimal constant. Instead of struggling to find analytic solutions, we present a systematic algorithm for each architectuare 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 we impose is not the global one (2.1.2) ; it only need to be locally Lipschitz in a sense that the bounding constant c exists for each $\boldsymbol{x}$. It has two conclusions : the differentiability (a.e.) and the formula for the optimal constant.

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

$$Lip(f) = \sup_{\boldsymbol{x} \in \mathbb{R}^n} ||D_{\boldsymbol{x}}(f)|| \tag{3.1.1}$$

Here, $D_{\boldsymbol{x}}f$ is the Jacobian matrix of f or a matrix A satisfying

$$\lim_{\boldsymbol{y}\to\boldsymbol{x}} \frac{||f(\boldsymbol{y}) - f(\boldsymbol{x}) - A(\boldsymbol{y}-\boldsymbol{x})||}{||\boldsymbol{y}-\boldsymbol{x}||} = 0.$$

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

The first statement of the theorem — the differentiability — is usually refered 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 measure theory including [9] covered this issue. In the process, concepts like absolute continuity, Radon-Nykodym theorem, total variation are involved. A proof for multi-dimensional case is found, for example, in [11].

The second statement, or the optimal Lipschitz constant is the one that we can use for our purpose. We think of the matrix whose entries are partial derivatives at point $\boldsymbol{x}$, compute the matrix norm and take the supremum over all $\boldsymbol{x} \in \mathbb{R}^n$. But, it doesn't mean that we can always find the accurate value $\mathrm{Lip}(f)$, given f and $\boldsymbol{x}$. To evaluate $\mathrm{Lip}(f)$, we should find the supremum, or the maximum if it exists. The theorem converts the problem of evaluating the constant to the problem of optimization problem.

If $f : \mathbb{R} \to \mathbb{R}$, then the previous condition becomes (2.2.7). If $f : \mathbb{R}^n \to \mathbb{R}$,

3.2 Multi-Layer Perceptron

3.3 Dropout and Normalization

3.4 CNN

3.5 RNN

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] Keith Conrad. *Computing the norm of a matrix.* `https://kconrad.math.uconn.edu/blurbs/linmultialg/matrixnorm.pdf`

[9] Rudin, Walter. *Real and Complex Analysis.* : McGraw-Hill Science/Engineering/Math, 1986.

[10] Herbert Federer. *Geometric measure theory.* Classics in Mathematics. Springer-Verlag Berlin Heidelberg, 1969.

[11] Villani, Cédric. *Optimal transport: old and new.* Vol. 338. Berlin: springer, 2009.