ABSTRACT

Title of dissertation: NONLINEAR ANALYSIS OF PHASE RETRIEVAL AND DEEP LEARNING

Dongmian Zou, Doctor of Philosophy, 2017

Dissertation directed by: Professor Radu Balan
Department of Mathematics

Nonlinearity causes information loss. The phase retrieval problem, or the phaseless reconstruction problem, seeks to reconstruct a signal from the magnitudes of linear measurements. With a more complicated design, convolutional neural networks use nonlinearity to extract useful features. We can model both problems in a frame-theoretic setting. With the existence of a noise, it is important to study the stability of the phaseless reconstruction and the feature extraction part of the convolutional neural networks. We prove the Lipschitz properties in both cases. In the phaseless reconstruction problem, we show that phase retrievability implies a bi-Lipschitz reconstruction map, which can be extended to the Euclidean space to accommodate noises while remaining to be stable. In the deep learning problem, we set up a general framework for the convolutional neural networks and provide an approach for computing the Lipschitz constants.

NONLINEAR ANALYSIS OF PHASE RETRIEVAL AND DEEP LEARNING

by

Dongmian Zou

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Radu Balan, Chair/Advisor
Professor John Benedetto
Professor Ramalingam Chellappa
Professor Prakash Narayan
Professor Kasso Okoudjou

To my family

Acknowledgments

First and foremost I would like to thank my advisor, Professor Radu Balan for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past four years. I still remember the first time I knocked at his door and he introduced the phaseless reconstruction problem to me. He has always made himself available for help and advice. I have always been impressed by his encyclopedic knowledge in so many fields in math, engineering and physics. He provides me with so many opportunities to travel and meet with different researchers. He always encourage me to talk to people and exchange ideas.

I would also like to thank Professor John Benedetto, Professor Ramalingam Chellappa, Professor Prakash Narayan and Professor Kasso Okoudjou for agreeing to serve on my thesis committee and for sparing their invaluable time reviewing the manuscript.

I would like to thank Dr. Maneesh Singh for his invitation to the topic of deep learning and mentorship during the past two years. He provided me with a great opportunity to work in real industrial environment to apply the theories in deep learning.

I would like to thank all the people in the Norbert Wiener Center for Harmonic Analysis and Its Applications. Thanks to Professor John Benedetto for his leading such an extraordinary group of researchers and for the fascinating lectures he gave on harmonic analysis and wavelets; he always makes me feel encouraged every time I meet him. Thanks to Professor Kasso Okoudjou for the introduction to graduate-

level real analysis and for his care during my graduate life. Thanks to Professor Wojciech Czaja for the lectures on Gabor analysis and the reference books he listed. They are great people and I really learned a lot from them. I would like also to thank Dr. Wei-Hsuan Yu, Dr. Mark Lai, Dr. James Murphy, Dr. Matt Begue and Dr. Chae Clark for all their suggestions as senior students from the NWC. Thanks to Weilin Li, Yiran Li, and Mark Magsino for being great study mates and for all the discussions that I benefited from.

I would also like to thank Professor Konstantina Trivisa for directing the fantastic program of AMSC. Thanks to Professor Eitan Tadmor and Professor Pierre-Emmanuel Jabin for directing the the Center for Scientific Computation and Mathematical Modeling. Thanks to all the professors in the classes I took in the first two years of graduate school. They built the foundation for my development in math and engineering. I would also like to acknowledge support from the staff members in the Math Department, the AMSC Program and the CSCAMM: Alverda McCoy, Jessica Sadler, Cristina Garcia, Celeste Regalado, Shaton Welton, Agi Alipio, Sara Moran, Anil Zenginoglu and Andrew Arensburger; they are always helpful.

I am also grateful to my friends Siming He, Tianyu Ma, Chen Qian, Luyu Sun, Zhenfu Wang, Jinhang Xue, Zhang Zhang for the wonderful graduate life in College Park. I thank all the people I meet during workshops and conferences.

I owe my deepest gratitude to my family - my mother and father who have always stood by me and who have always understood every decision I made in my life. Words cannot express the gratitude I owe them. I owe my grandfather for not accompanying him in his last days. I own my grandmother, for all the more years I

cannot spend by her side.

I would also like to express my deepest thanks to Jennifer Lin, for her love and her unwavering support on my academic path.

I would like to acknowledge financial support from NSF as research assistantship. Part of this thesis reports work done under the support of NSF fund DMS-1413249. I am grateful to all the financial support I received for travel and conferences during the past five years.

Table of Contents

List of Abbreviations

$\mathbb{R}$	The set of real numbers
$\mathbb{C}$	The set of complex numbers
$\|\cdot\|_{L^p}$, $\|\cdot\|_{l^p}$, $\|\cdot\|_p$	The L^p-norm, l^p-norm and p-norm when there is no ambiguity
$\langle \cdot, \cdot \rangle$	The inner product
$\mathrm{supp}(f)$	The support of a function f
$\hat{f}$	The Fourier transform of f
f_λ	The dilation of f with scale λ
$\mathrm{Sym}(\mathcal{H})$	The space of symmetric operators on a Hilbert space $\mathcal{H}$
A^*	The conjugate transpose of a matrix A
$\mathrm{tr}(A)$	The trace of a matrix / operator A
$\rho(A)$	The spectrum of a matrix / operator A
$\mathrm{real}(z)$	The real part of a complex number z
$\mathrm{imag}(z)$	The imaginary part of a complex number z
$\bar{z}$	The complex conjugate of a complex number z
sinc	The sinc function defined by $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$
1D / 2D / 3D	one / two / three dimensional
CDI	Coherent Diffractive Imaging
CNN	Convolutional Neural Network
PCA	Principal Component Analysis
POVM	Positive Operator Valued Measure
PSD	Power Spectral Density
SNR	Signal to Noise Ratio
SSS	Strict Sense Stationary
SVM	Support Vector Machine
WSS	Wide Sense Stationary

Chapter 1: Background

1.1 Signal reconstruction without phase

Phase retrieval is a fundamental problem in signal reconstruction. In this problem we seek to recover the phase of a signal from the magnitude of linear measurements. It has important applications in X-ray crystallography, quantum information and speech recognition.

1.1.1 X-ray crystallography

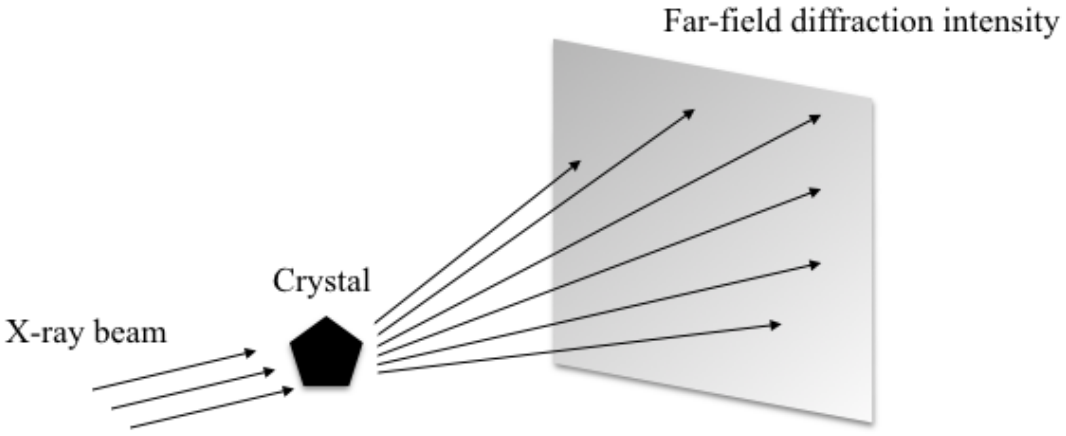

Figure 1.1: The experiment settings for X-ray crystallography.

X-ray crystallography is an important technique for determining the atomic and molecular structure of a crystal. Figure 1.1 illustrates the measurement process. The structure of the crystal causes the X-ray to diffract, and the diffracted pattern is produced on the far-field. The diffracted pattern only contains *the magnitude of the Fourier transform* of the crystal. A detailed description can be found in [49, 80, 90].

In X-ray crystallography, or more generally, coherent diffractive imaging (CDI), we only have information on the magnitude of the linearly transformed data. The phase information is difficult to get by experiment due to the high oscillation rate ($\sim 10^{15}$Hz) of the electromagnetic field. Therefore, we seek to regain the phase information from the magnitude measurements and reconstruct the crystal image.

X-ray crystallography set a start of the phase retrieval problems. The earliest and most popular method is the alternating projection algorithms proposed by Gerchberg and Saxton, later improved by Fienup [49]. There is no guarantee of convergence in that iteration method, and it does not work well for 3D images. The phase retrieval problem arouses interests of a lot of mathematicians, and is gradually generalized to a setting that more applications fit in.

1.1.2 Quantum information

In quantum tomography, the quantum state is identified from the statistics of the measurements [39, 62]. Suppose we have a d-dimensional Hilbert space $\mathcal{H}$. A quantum state is a Hermitian matrix ρ with trace $\mathrm{tr}(\rho) = 1$.

The quantum measurements are generally described by positive operator val-

ued measures (POVM's). That is,

$$\mathbf{A} = \{A_1, \cdots, A_m\} \tag{1.1}$$

where each A_k, $k = 1, \cdots, m$, is Hermitian. The measurement gives

$$\mathbf{A}(\rho) = (\mathrm{tr}\rho A_1, \cdots, \mathrm{tr}\rho A_m) \ . \tag{1.2}$$

If ρ is a pure state, that is, $\rho = |\psi\rangle\langle\psi|$, then the recovery of the state falls into the problem of generalized phase retrieval, as studied in [103]. Moreover, if we choose to use rank-one POVM's ($A_k = |f_k\rangle\langle f_k|$, $k = 1, \cdots, m$), then we have

$$\mathrm{tr}\rho A_k = |\langle\psi|f_k\rangle|^2 \ . \tag{1.3}$$

In this case, we see that the quantum measurements give (squared) magnitude of the linear measurements of the state. The reconstruction of the underlying state is thus modeled by the phase retrieval problem.

1.1.3 Audio signal processing

The phase retrieval problem finds its application in audio processing as well [16]. In speech recognition, sampled speech signals are first transformed to the time-frequency domain via discrete windowed Fourier transform. In most signal enhancement algorithms, we only modify the amplitude of the transformed signal in the time-frequency domain and keep its noisy phase.

To illustrate, we take the example from [16]. The speech signal is sampled as $\{x(t) : t = 0, 1, \cdots, T-1\}$. The fast windowed Fourier transform gives for

$\omega = 0, 1, \cdots, M-1$ that

$$X(k,w) = \sum_{t=0}^{M-1} g(t)x(t+kN)e^{-2\pi i\omega t/M}, \qquad k = 0, 1, \cdots, \frac{T-M}{N} \ . \tag{1.4}$$

where g is the analysis window function, M is the window size and N is the time step. In the Ephraim-Malah noise reduction method, we apply a nonlinear transform on $|X(k,\omega)|$ that reads

$$\begin{aligned} Y(k,\omega) \ = \ & \frac{\sqrt{\pi}}{2}\frac{v(k,\omega)}{\gamma(k,\omega)} \exp\left(-\frac{v(k,\omega)}{2}\right)\left[(1+v(k,\omega))I_0\left(\frac{v(k,\omega)}{2}\right)+\right. \\ & \left. v(k,\omega)I_1\left(\frac{v(k,\omega)}{2}\right)\right] \cdot \left|X(k,\omega)\right| \ , \end{aligned} \tag{1.5}$$

where I_0 and I_1 are Bessel functions and v and γ are estimates to some SNRs. The enhanced speech signal is then

$$x^\sharp(t) = \sum_{k=0}^{(T-M)/N} \sum_{\omega=0}^{M-1} Y(k,\omega)\frac{X(k,\omega)}{|X(k,\omega)|}e^{2\pi i\omega(t-kN)/M}h(t-kN) \ , \tag{1.6}$$

where h is the synthesis window function. We see that in this example, in the representation domain, we do some manipulation without using the information from the phase. In fact, the phase in this case is noisy and it is desired that we do reconstruction without phase.

1.2 Signal classification using convolutional neural networks

Convolutional neural networks (CNNs) have enjoyed conspicuous success in various pattern recognition tasks [68, 70, 78, 95]. CNNs are artificial neural networks that use convolution in place of general matrix multiplication in at least one place

[54]. The convolution operation, which we denote by $*$, is understood to be

$$\begin{aligned} f * g(t) &= \int f(s)g(t-s)ds \\ &= \int f(t-s)g(s)ds \ , \end{aligned} \tag{1.7}$$

where the integral domain and the measure are to be specified depending on the detailed problem settings. In the discrete 2D case, the convolution reads as a left multiplication by a Toepliz matrix.

In CNNs, we use convolution between a signal and a filter. The filter in this case is called the *kernel*. In most applications, we take kernels of much smaller size than the input signals. In this case, the machine learning system will have the *sparse connection property* (see [54], Chapter 9), which roughly means that a pixel in the input only affects a few pixels in the output. If we have a deep structure, then we expect that in shallow layers each pixel in the output relates to only the neighboring pixels in the input, while in deep layers each pixel is affected by all the pixels in the input. It is therefore believed that CNNs are able to catch different levels of features in different layers.

Another property that makes convolution important is its *equivariance to translations* (in 1D case, it is said to be time-invariant), which says that the operation of convolution commutes with translation. Intuitively, if we move the object by certain pixels, the output will move by the same number of pixels.

1.2.1 The AlexNet and GoogleNet

The AlexNet (see [68]) and GoogleNet ([95]) are typical CNNs used in image classification. While they have different structures and the latter is much deeper than the former, they do have in common that they both contain convolution layers, detection layers and pooling layers. Their structures are illustrated in Figure 1.2 and 1.3, respectively.

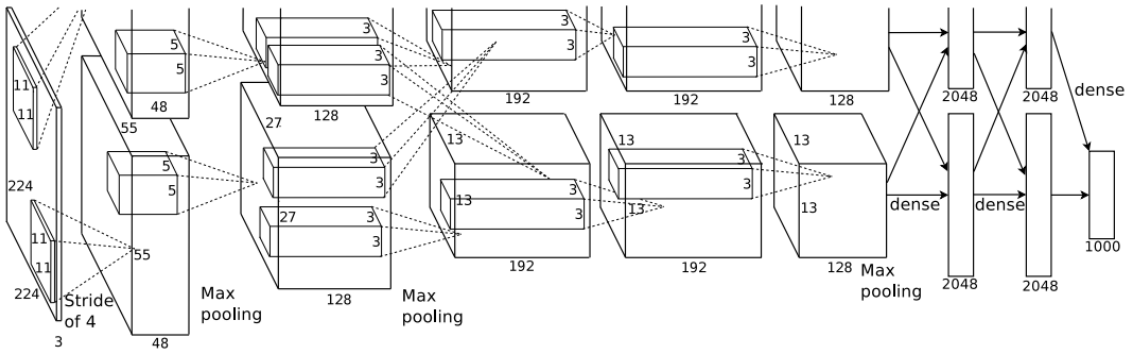

Figure 1.2: The AlexNet (the structure is the CNN used in [68]).

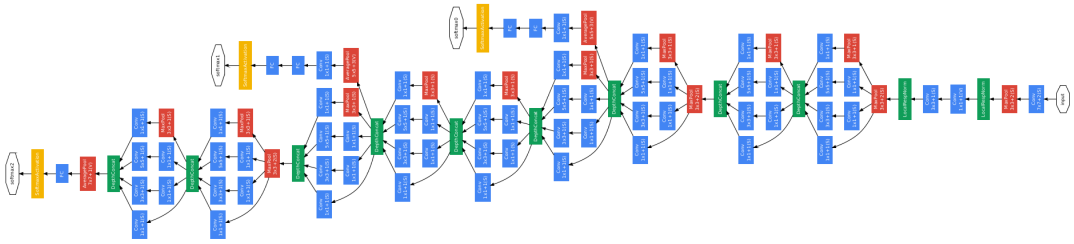

Figure 1.3: The GoogleNet (the struture is the CNN used in [95]).

In a detector layer, a nonlinear function called the *activation* function is applied entriwise to the signal. This function in some cases has biological motivations

(for instance the *rectifier*, which is defined to be $f(x) = \max\{0, x\}$) and in some cases (such as softmax function) has a probabilistic explanation.

In a pooling layer, the output of the network is replaced with a summary statistic of nearby outputs ([54], Chapter 9). In both the AlexNet and the GoogleNet, the maximum of pixels in a certain neighborhood is taken. On one hand, pooling reduces the computing complexity; on the other hand, it makes the representation to be approximately invariant to small translations.

In AlexNet and GoogleNet, the coefficients in the network are trained using certain training data. The representation of images are sent through fully connected layers and a softmax layer to output the classification results.

1.2.2 The Scattering Transform

The scattering network is a deep neural network introduced in [78]. It has been successfully applied in several image classification tasks [32, 63].

The scattering network is a CNN because it has convolutional layers, detection layers as well as pooling layers. It is different from the traditional CNNs for the following reasons: first, the convolutional layers are prescribed as certain class of wavelets; second, the nonlinearity is the absolute value instead of the rectifier; third, in the pooling layer the scattering network uses a lower frequency filter to take local averages.

An input of a scattering network propagates along the paths from the first layer of the network to an output. A path is defined as a sequence of filters. For

instance, in Figure 1.4, $(g_{1,1}, g_{2,1}, g_{3,1}, g_{4,1})$ is a path, while $(g_{1,2}, g_{2,1}, g_{3,1}, g_{4,1})$ is not a path. If the paths are labeled in this way we can denote the path to be the indexes, for instance, $((1,1),(2,1),(3,1),(4,1))$. Since we have a tree-structured network the paths are well-defined.

The scattering network is also called a *scattering transform*. There is a surjective map from $\mathbb{R}^{d+}$ to the set of all paths (see [78], Chapter 3) and thus the scattering network is comparable to the traditional Fourier transform.

A typical scattering transform is illustrated in Figure 1.4. We see that it is a tree-structured CNN and there is an output from every layer. In terms of CNN, we have only presented the feature extraction of the scattering network, and in practice, we put a classifier such as a PCA layer or an SVM layer (see [32]) at the bottom.

As can be seen in Figure 1.4, an input signal propagates through the paths of the network to generate outputs in all the layers. For instance, $y_{4,1}$ on the top right corner reads

$$y_{4,1}(t) = |||f * g_{1,1}| * g_{2,1}| * g_{3,1}| * \phi_4(t) \ .$$

In scattering networks, the filters $g_{i,j}$'s are taken to be different scales of a mother wavelet ψ, that is, we take dilations of ψ:

$$g_{i,j}(t) = \psi_{\lambda_{i,j}}(t) := \lambda_{i,j}^d \psi(\lambda_{i,j} t) \ , \tag{1.8}$$

and ϕ_i is taken to be a fixed scale of the scaling function ϕ associated with ψ, that is,

$$\phi_i(t) = 2^{-Jd}\phi(2^{-J}t), \qquad \forall i = 1, 2, 3, \cdots \ . \tag{1.9}$$

A scattering network is promising because some good properties can be demon-

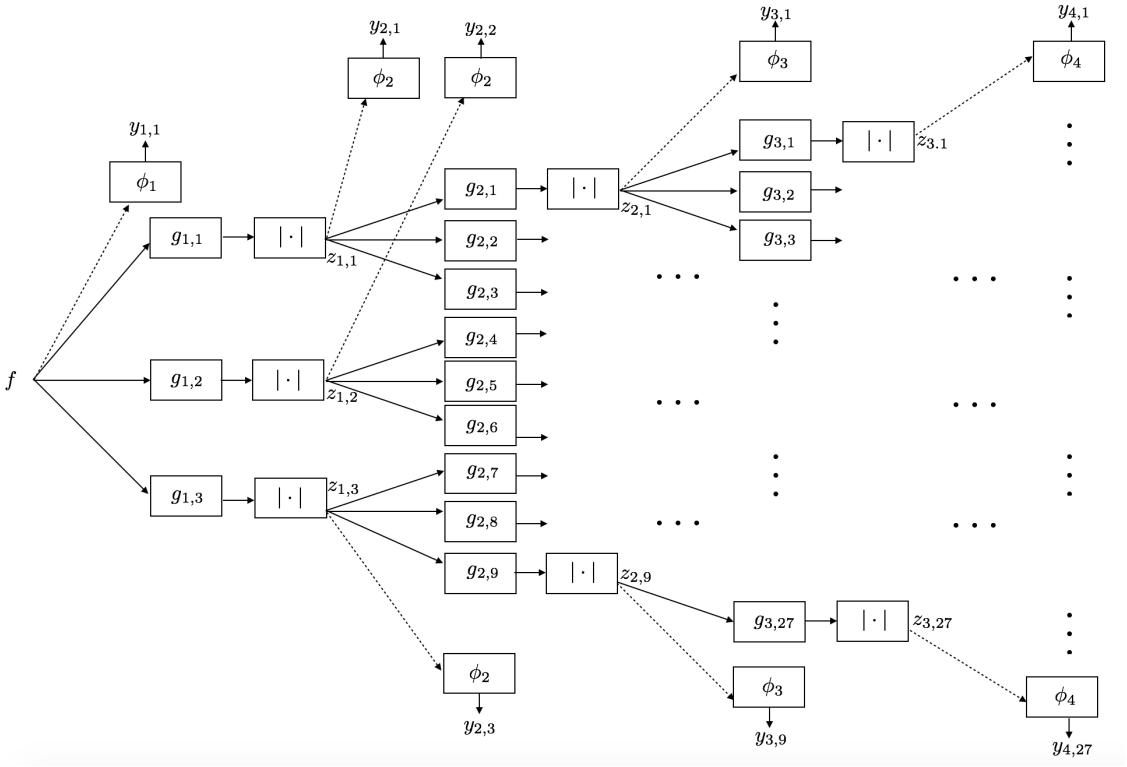

Figure 1.4: The scattering network.

strated. In particular, a scattering transform is approximately invariant to translations, and stable in correspondence with small deformations. These properties are important when we want to do image classification because we hope that translations and deformations will not change the class of an image. However, these properties hold to be true only if we use specifically designed wavelets. For instance, in 1D one possibility is to use the cubic spline Battle-Lemarié wavelets (see [44] for an instruction on how to construct this family of wavelets), with the scaling function given by

$$\hat{\phi}(\xi) = \frac{16\sqrt{315}\sin^4(\frac{\xi}{2})}{\sqrt{2\pi}\xi^4\sqrt{4\cos^6(\frac{\xi}{2}) + 114\cos^4((\frac{\xi}{2}) + 180\cos^2(\frac{\xi}{2}) + 17}} \tag{1.10}$$

and the mother wavelet given by

$$\hat{\psi}(\xi) = \frac{256\sqrt{315}}{\sqrt{2\pi}} e^{\frac{i\xi}{2}} \frac{\sin^8(\frac{\xi}{4})}{\xi^4} \left(\frac{F(\frac{\xi}{4})}{G(\frac{\xi}{4})G(\frac{\xi}{2})} \right)^{\frac{1}{2}} , \tag{1.11}$$

where

$$F(\xi) = 4\sin^6(\xi) + 114\sin^4(\xi) + 180\sin^2(\xi) + 17$$

and

$$G(\xi) = 4\cos^6(\xi) + 114\cos^4(\xi) + 180\cos^2(\xi) + 17 \ .$$

In practice, specifically in image classification, it is possible to take a Morlet wavelet and and achieve state-of-the-art result (see [32]).

In a scattering network, the translation invariance is achieved when the pooling layer contains only extremely low-pass filters. This is not practical and the resulting feature would be useless because almost all information is lost. Therefore, the invariance property that makes it work is still mainly the equivariance property due to the use of convolution.

In [105,106], the authors give a more general structure for scattering networks. Instead of families of wavelets, the authors consider the filters that compose a semi-discrete frame. They show the equivariance and stability of a general scattering network. We discuss the semi-discrete frames in Section 2.1 and compare their structure with our general framework in Chapter 4.

Chapter 2: Mathematical Preliminaries

2.1 Frame theory

In [16], the authors study the phase retrieval problem using the frames, which is a useful tool in applied harmonic analysis and signal processing. An extended study of frame theory can be found in [37,38,41]. We take the definition for a finite frame in a Hilbert space.

Definition 2.1.1. *Let $\mathcal{H}$ be a n-dimensional Hilbert space. $\mathcal{F} = \{f_1, f_2, \ldots, f_m\} \subset \mathcal{H}$ is a frame for $\mathcal{H}$ if there exist constants $A, B > 0$ such that*

$$A \|x\|^2 \leq \sum_{k=1}^{m} |\langle x, f_k \rangle|^2 \leq B \|x\|^2 \tag{2.1}$$

for any $x \in \mathcal{H}$.

The constants A and B in the above is called *lower* and *upper frame bounds*, respectively. A frame $\mathcal{F}$ is *tight* if it is possible to choose $A = B$ in (2.1). Further, if the optimal frame bounds are $A = B = 1$, then the frame is said to be a *Parseval frame*. If we only have the second inequality in (2.1), then we call $\mathcal{F}$ a Bessel sequence.

Since we have a finite dimensional space and a finite frame, we have the following equivalent definition given as a lemma.

Lemma 2.1.2. (see [41], Chapter 1) *$\mathcal{F}$ is a frame for $\mathcal{H}$ if only if it spans $\mathcal{H}$.*

Given a frame $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$, if we consider $x \in \mathcal{H}$ as a signal, we can take linear measurements with the frames as $\langle x, f_k \rangle$ for $k = 1, \cdots, m$. We can perfectly reconstruct x in this case.

Definition 2.1.3. *Let $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$ be a frame in $\mathcal{H}$. The analysis operator T associated with $\mathcal{F}$ is defined by*

$$T : \mathcal{H} \to \mathbb{C}^m \qquad T(x) = \{\langle x, f_k \rangle\}_{k=1}^m \; ; \tag{2.2}$$

whereas the synthesis operator T^* is defined by

$$T^* : \mathbb{C}^m \to \mathcal{H} \qquad T^*(c) = \sum_{k=1}^m c_k f_k \; . \tag{2.3}$$

The frame operator $S : \mathcal{H} \to \mathcal{H}$ is defined to be $S = T^*T$.

In this case S is always invertible by the definition of frames, and the perfect reconstruction formula is given by

$$x = S^{-1}Sx = \sum_{k=1}^m \langle x, f_k \rangle \, S^{-1} f_k \; . \tag{2.4}$$

Therefore, we can see that in the case of linear measurements, it is rather easy to reconstruct the signal. We never lose information (and even have redundant information) in the process of measurements, and therefore it does not take much effort to get back. However, if we make nonlinear measurements, the process of reconstruction will be much more difficult, as we will discuss in later chapters.

For convolutional neural networks, the linear measurements are taken using convolutions. We introduce the semi-discrete frames as defined in [105, 106]. In this case, we specify the signals to be taken from $L^2(\mathbb{R}^d)$.

Definition 2.1.4. *Let $\mathcal{F} = \{f_k\}_{k\in I} \subset L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ where I is an index set that is at most countable. $\mathcal{F}$ is said to form the atoms of a semi-discrete Bessel sequence if there exists a constant $B > 0$ such that*

$$\sum_{k\in I} \|x * f_k\|_2^2 \leq B \|x\|_2^2 \tag{2.5}$$

for any $x \in L^2(\mathcal{H})$. Further, we say $\mathcal{F}$ form the atoms of a semi-discrete frame if there is also a constant $A > 0$ such that

$$A \|x\|_2^2 \leq \sum_{k\in I} \|x * f_k\|_2^2 \leq B \|x\|_2^2 \tag{2.6}$$

for any $x \in L^2(\mathcal{H})$.

If we define the Fourier transform of a function $f \in L^1(\mathbb{R}^d)$ to be

$$\hat{f}(\omega) = \int_{\mathbb{R}^d} f(t) e^{-2\pi i \omega \cdot t} dt \ , \tag{2.7}$$

we have the Parseval's relation

$$\|x * f_k\|_2 = \left\| \hat{x} \cdot \hat{f}_k \right\|_2 \tag{2.8}$$

for each k.

According to the above, it makes sense to consider filters that lie in some larger space (for instance, that contains "the delta function") when we talk about the application in CNNs. We are going to define that space in Chapter 4. We introduce some preliminary definitions that will be used there.

Definition 2.1.5. (Algebra over a field) *Let K be a field. Let V be a vector space over K equipped with an additional binary operation (called the product)*

$$\cdot : V \times V \to V \ . \tag{2.9}$$

Then V is said to be an algebra *over K if for all $x, y, z \in V$ and all $a, b \in V$,*

1. $(x + y) \cdot z = x \cdot z + y \cdot z$;

2. $x \cdot (y + z) = x \cdot y + x \cdot z$;

3. $(ax) \cdot (by) = (ab)x \cdot y$.

Definition 2.1.6. (see [50]) *A* Banach algebra *is an algebra $\mathcal{B}$ over $\mathbb{C}$ equipped with a norm with respect to which it is a Banach space and which satisfies*

$$\|xy\| \leq \|x\| \, \|y\| \tag{2.10}$$

for all $x, y \in \mathcal{B}$.

2.2 Holomorphic functional calculus

Let T be a bounded linear operator on a Banach space, the holomorphic functional calculus relates a function of T with its spectrum $\rho(T)$. A detailed study can be found in [85], Chapter IX. We will mainly focus on the case where A is a Hermitian matrix.

Definition 2.2.1. *Let $\mathcal{B}$ be a Banach space and A be a linear operator. The resolvent transformation of A is defined to be*

$$R_A(z) = (A - zI)^{-1} \, . \tag{2.11}$$

The following theorem states a decomposition result for a general linear transformation.

Theorem 2.2.2. (see [85], Chapter XI.148) *Let A be a bounded linear map of the Banach space $\mathcal{B}$ and let ρ and ρ' be two complementary isolated parts of its spectrum. Then $\mathcal{B} = \mathcal{B}_\rho \oplus \mathcal{B}_{\rho'}$ each of which is transformed by A to itself. The projection P_ρ on $\mathcal{B}_\rho$ is equal to*

$$P_\rho = -\frac{1}{2\pi i}\int_{\partial D} R_A(z)dz \ , \tag{2.12}$$

taken along the boundary of an arbitrary domain D which is admissible with respect to A and contains all elements in ρ and no element in ρ'.

In the case where A is a Hermitian matrix, the above theorem reads

Theorem 2.2.3. *Let A be an n-by-n Hermitian matrix with spectral decomposition*

$$A = \sum_{j=1}^{n} \lambda_j P_j \tag{2.13}$$

where λ_j's are the eigenvalues of A and P_j's are the corresponding orthogonal projections. Then we have for each $j = 1, \cdots, n$ that

$$P_j = -\frac{1}{2\pi i}\int_{\partial \Gamma_j} R_A(z)dz \ , \tag{2.14}$$

where Γ_j is a Jordan curve that encloses only one eigenvalue λ_j.

A Hermitian matrix A has n eigenvalues (counting multiplicities), which we denote as λ_j or $\lambda_j(A)$ for $j = 1, \cdots, n$. In the following chapters, unless otherwise specified, we shall always assume that $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$.

2.3 Lipschitz continuity and extension theorems

The Lipschitz property (see [26]) is closely related to the stability of both the phaseless reconstruction and the CNN.

Definition 2.3.1. *Let (X, d_X) and (Y, d_Y) be two metric spaces where d_X and d_Y are the distance functions respectively. A continuous map $f : X \to Y$ is said to be Lipschitz continuous, or Lipschitz, if*

$$\sup_{x_1, x_2 \in X} \frac{d_Y\left(f(x_1), f(x_2)\right)}{d_X\left(x_1, x_2\right)} < \infty \ . \tag{2.15}$$

In this case, we denote

$$\mathrm{Lip}(f) := \sup_{x_1, x_2 \in X} \frac{d_Y\left(f(x_1), f(x_2)\right)}{d_X\left(x_1, x_2\right)} \ . \tag{2.16}$$

The Lipschitz continuity implies an upper bound. In some cases it makes sense to introduce a lower bound as well. We define the bi-Lipschitz property as follows.

Definition 2.3.2. *Let (X, d_X) and (Y, d_Y) be two metric spaces where d_X and d_Y are the distance functions respectively. A continuous map $f : X \to Y$ is said to be bi-Lipschitz, if there exist constants A and B, with $0 < A \leq B < \infty$, such that*

$$A d_X\left(x_1, x_2\right) \leq d_Y\left(f(x_1), f(x_2)\right) \leq B d_X\left(x_1, x_2\right) \tag{2.17}$$

Obviously, if a function f is bi-Lipschitz, then it is injective. However, in general the injectivity of a Lipschitz function does not imply the bi-Lipschitz property.

If a Lipschitz continuous function is defined only on a subspace of a metric space, under some conditions it is possible to extend the function to the whole space while keeping the Lipschitz constant. We state the Kirszbraun Theorem (see [104]).

Definition 2.3.3. (The Kirszbraun Property (K)) *Let X and Y be two metric spaces with metric d_X and d_Y respectively. (X, Y) is said to have Property (K) if for any pair of families of closed balls $\{B(x_i, r_i) : i \in I\}$, $\{B(y_i, r_i) : i \in I\}$, such*

that $d_Y(y_i, y_j) \leq d_X(x_i, x_j)$ *for each* $i, j \in I$*, it holds that*

$$\bigcap_{i \in I} B(x_i, r_i) \neq \emptyset \;\Rightarrow\; \bigcap_{i \in I} B(y_i, r_i) \neq \emptyset \; . \tag{2.18}$$

In general it may not be obvious whether a given pair of metric spaces satisfies the Property (K). Nevertheless, if X and Y are both Hilbert spaces, then the Property (K) is guaranteed. We state it as a theorem.

Theorem 2.3.4. ([104], Chapter 10) *Suppose* X *and* Y *are Hilbert spaces and* d_X *and* d_Y *are the metrics induced by the inner products in each space respectively. Then* (X, Y) *has Property (K).*

The Kirszbraun Theorem states the following:

Theorem 2.3.5. ([104], Chapter 10) *Let* X *and* Y *be two metric spaces and* (X, Y) *has Property (K). Suppose* U *is a subset of* X *and* $f : U \to Y$ *is a Lipschitz map. Then there exists a Lipschitz map* $F : X \to Y$ *which extends* f *to* X *such that* $F|_U = f$ *and* $\mathrm{Lip}(F) = \mathrm{Lip}(f)$.

2.4 Random processes

Random processes are useful models in signal processing. In this section, we define the basic concepts that will be used later. We refer to [67] for a detailed discussion.

Definition 2.4.1. *A family of random variables* $X_t : t \in \mathfrak{T}$ *defined on a common probability space* $(\Omega, \mathfrak{F}, \mathbb{P})$ *is called a random process if* $\mathfrak{T}$ *is a subset of the real line and it is called a random field if* $\mathfrak{T}$ *is multi-dimensional.*

When there is no misunderstanding, we shall call X_t a random process regardless of $\mathfrak{T}$. A trajectory of X_t is the realization $X_\omega(t)$ for some fixed $\omega \in \Omega$.

Now we introduce the notion of strict and wide sense stationary (SSS and WSS) processes.

Definition 2.4.2. *A random process X_t is called strict sense stationary (SSS) if for any $t_1, \cdots, t_n \in \mathfrak{T}$ and $A_1, \cdots, A_n \in \mathfrak{F}$ the probabilities*

$$\mathbb{P}\Big\{X_{t_1+\tau} \in A_1, \cdots, X_{t_n+\tau} \in A_n\Big\} \tag{2.19}$$

does not depend on τ, where $\tau \in \mathfrak{T}$.

Definition 2.4.3. *A random process X_t is called wide sense stationary (WSS) if there exist a constant (the expectation) μ and a function (the auto-correlation) $R(t), t \in \mathfrak{T}$ such that $\mathbb{E}(X_t) = \mu$ and $\mathbb{E}(X_t \bar{X}_s) = R(t-s)$ for all $t, s \in \mathfrak{T}$.*

Obviously, if the second moment is finite, then SSS implies WSS. For a WSS process, the power spectral density (PSD) is the Fourier transform of the autocorrelation. This is the Wiener-Khinchin theorem (see [82], Chapter 10):

Theorem 2.4.4. (Wiener-Khinchin) *Let $X = X(t)$ be a WSS process. Let $S_X(\omega) := \lim_{T\to\infty} \mathbb{E}\left[\left|(2T)^{-1}\int_{-T}^{T} X(t)e^{-2\pi i\omega t}dt\right|^2\right]$ be the spectral power density and let $R_X(t) = \mathbb{E}\left[X(t)\bar{X}(t-\tau)\right]$ be the auto-correlation function. Then*

$$S_X(\omega) = \hat{R}_X(\omega)\ . \tag{2.20}$$

Chapter 3: Stable reconstruction for the phase retrieval problem

3.1 Frame settings of the phase retrieval problem

We first give a rigorous setting of the phase retrieval problem, in the language of frame theory, introduced in Chapter 2.1.

3.1.1 The measurement maps

Let $\mathcal{H}$ be a n-dimensional real or complex Hilbert space, in practice $\mathcal{H} = \mathbb{R}^d$ or $\mathbb{C}^n$. Assume that $\mathcal{F} = \{f_1, \cdots, f_m\}$ is a frame for $\mathcal{H}$. We denote the nonlinear measurement maps α and β to be

$$\alpha : \mathcal{H} \to \mathbb{R}^m, \qquad \alpha(x) = \left(|\langle x, f_k\rangle|\right)_{1\le k\le m} \ , \tag{3.1}$$

and

$$\beta : \mathcal{H} \to \mathbb{R}^m, \qquad \beta(x) = \left(|\langle x, f_k\rangle|^2\right)_{1\le k\le m} \ . \tag{3.2}$$

We call α the *magnitude measurement map* and β the *square measurement map*.

Note that β is the entriwise square of α. Also, there is an ambiguity of a universal phase, that is,

$$\alpha(x) = \alpha(e^{i\phi}x), \qquad \forall \phi \in [0, 2\pi) \ . \tag{3.3}$$

Therefore, we consider an equivalence relation $\sim$ defined by $x \sim y$ if and only if there is a scalar a with $|a| = 1$ such that $y = ax$. Let $\hat{\mathcal{H}}$ denote the collection of the equivalence classes and $\hat{x}$ denote the elements in $\hat{\mathcal{H}}$. Now we read $\alpha : \hat{\mathcal{H}} \to \mathbb{R}^m$ and $\beta : \hat{\mathcal{H}} \to \mathbb{R}^m$.

The *phase retrieval problem*, or the *phaseless reconstruction problem*, refers to analyzing when α (or equivalently, β) is an injective map, and in this case to finding "good" left inverses. The frame $\mathcal{F}$ is said to be *phase retrievable* if the nonlinear map α (or β) is injective. We also say α (or β) is phase retrievable in the case.

3.1.2 Distance function of the quotient space

In general, a quotient space induced by an equivalence relation is not necessarily metrizable. Nevertheless, in the setting described above, there are natural distance functions associated with α and β.

We consider two classes of metrics (distances), respectively:

(1) the class of *natural metrics*. For every $1 \leq p \leq \infty$ and $x, y \in \mathcal{H}$, we define

$$D_p(\hat{x}, \hat{y}) = \min_{|a|=1} \|x - ay\|_p \ . \tag{3.4}$$

When no subscript is used, $\|\cdot\|$ denotes the Euclidean norm, $\|\cdot\| = \|\cdot\|_2$.

(2) the class of *matrix norm induced metrics*. For every $1 \leq p \leq \infty$ and $x, y \in \mathcal{H}$, we define

$$d_p(\hat{x}, \hat{y}) = \|xx^* - yy^*\|_p = \begin{cases} \left(\sum_{k=1}^{n} (\sigma_k)^p \right)^{1/p} & for \quad 1 \leq p < \infty \\ \max_{1 \leq k \leq n} \ \sigma_k & for \quad p = \infty \end{cases} , \tag{3.5}$$

where $(\sigma_k)_{1 \leq k \leq n}$ are the singular values of the operator $xx^* - yy^*$, which is of rank

at most 2. Here x^* denotes the adjoint of x (see [12] for a detailed discussion), which is the transpose conjugate of x if $H = \mathbb{R}^n$ or $\mathbb{C}^n$.

Our choice in (3.5) corresponds to the class of Schatten norms. In particular, d_∞ corresponds to the operator norm $\|\cdot\|_{op}$ in $\mathrm{Sym}(\mathcal{H}) = \{T : \mathcal{H} \to \mathcal{H}\ ,\ T = T^*\}$; d_2 corresponds to the Frobenius norm $\|\cdot\|_{Fr}$ in $\mathrm{Sym}(\mathcal{H})$; d_1 corresponds to the nuclear norm $\|\cdot\|_*$ in $\mathrm{Sym}(H)$. Specifically, we have

$$d_\infty(x,y) = \|xx^* - yy^*\|_{op}\ ,\ d_2(x,y) = \|xx^* - yy^*\|_{Fr}\ ,$$

$$d_1(x,y) = \|xx^* - yy^*\|_*\ .$$

Note that the Frobenius norm $\|T\|_{Fr} = \sqrt{\mathrm{tr}(TT^*)}$ induces the Euclidean distance on $\mathrm{Sym}(\mathcal{H})$. As a consequence of Lemma 3.8 in [12], we have:

$$d_\infty(x,y) = \frac{1}{2}|\,\|x\|^2 - \|y\|^2\,| + \frac{1}{2}\sqrt{(\|x\|^2 + \|y\|^2)^2 - 4|\,\langle x,y\rangle\,|^2}\ ,$$

$$d_2(x,y) = \sqrt{\|x\|^4 + \|y\|^4 - 2|\,\langle x,y\rangle\,|^2}\ ,$$

$$d_1(x,y) = \sqrt{(\|x\|^2 + \|y\|^2)^2 - 4|\,\langle x,y\rangle\,|^2}\ .$$

To study the above distances it is important to study eigenvalues of symmetric matrices. Let $S^{p,q}(\mathcal{H})$ denote the set of symmetric operators that have at most p strictly positive eigenvalues and q strictly negative eigenvalues. In particular, $S^{1,0}(\mathcal{H})$ is the set of non-negative symmetric operators of rank at most one:

$$S^{1,0}(\mathcal{H}) = \{\, xx^*,\quad x \in \mathcal{H}\,\}\ . \tag{3.6}$$

If $\mathcal{H} = \mathbb{R}^n$ or $\mathbb{C}^n$, then $\mathrm{Sym}(\mathcal{H})$ is the set of n-dimensional Hermitian matrices.

Theorem 3.1.1. *We have the following statements regarding D_p and d_p:*

1. *For each* $1 \leq p \leq \infty$, D_p *and* d_p *are well-defined metrics (distances) on* $\hat{\mathcal{H}}$.

2. $(D_p)_{1\leq p\leq\infty}$ *are equivalent metrics, that is, each* D_p *induces the same topology on* $\hat{\mathcal{H}}$ *as* D_1. *Additionally, for every* $1 \leq p, q \leq \infty$ *the embedding* $i : (\hat{\mathcal{H}}, D_p) \to (\hat{\mathcal{H}}, D_q)$, $i(x) = x$, *is Lipschitz with Lipschitz constant*

$$L^D_{p,q,n} = \max(1, n^{\frac{1}{q}-\frac{1}{p}}). \tag{3.7}$$

3. *For* $1 \leq p \leq \infty$, $(d_p)_{1\leq p\leq\infty}$ *are equivalent metrics, that is each* d_p *induces the same topology on* $\hat{\mathcal{H}}$ *as* d_1. *Additionally, for every* $1 \leq p, q \leq \infty$ *the embedding* $i : (\hat{\mathcal{H}}, d_p) \to (\hat{\mathcal{H}}, d_q)$, $i(x) = x$, *is Lipschitz with Lipschitz constant*

$$L^d_{p,q,n} = \max(1, 2^{\frac{1}{q}-\frac{1}{p}}). \tag{3.8}$$

4. *The identity map* $i : (\hat{\mathcal{H}}, D_p) \to (\hat{\mathcal{H}}, d_p)$, $i(x) = x$, *is continuous with continuous inverse. However it is not Lipschitz, nor is its inverse.*

5. *The metric space* $(\hat{\mathcal{H}}, D_p)$ *is Lipschitz isomorphic to* $S^{1,0}(\mathcal{H})$ *endowed with Schatten norm* $\|\cdot\|_p$. *The isomorphism is given by the map*

$$\kappa_\alpha : \hat{\mathcal{H}} \to S^{1,0}(\mathcal{H}) \quad , \quad \kappa_\alpha(x) = \begin{cases} \frac{1}{\|x\|} x x^* & if \quad x \neq 0 \\ 0 & if \quad x = 0 \end{cases} . \tag{3.9}$$

The embedding κ_α *is bi-Lipschitz with the lower Lipschitz constant*

$$\min(2^{\frac{1}{2}-\frac{1}{p}}, n^{\frac{1}{p}-\frac{1}{2}})$$

and the upper Lipschitz constant

$$\sqrt{2}\max(n^{\frac{1}{2}-\frac{1}{p}}, 2^{\frac{1}{p}-\frac{1}{2}}) \ .$$

In particular, for $p = 2$, the lower Lipschitz constant is 1 and the upper Lipschitz constant is $\sqrt{2}$.

6. *The metric space $(\hat{\mathcal{H}}, d_p)$ is isometrically isomorphic to $S^{1,0}(\mathcal{H})$ endowed with Schatten norm $\|\cdot\|_p$. The isomorphism is given by the map*

$$\kappa_\beta : \hat{\mathcal{H}} \to S^{1,0}(\mathcal{H}) \quad , \quad \kappa_\beta(x) = xx^*. \tag{3.10}$$

In particular the metric space $(\hat{\mathcal{H}}, d_1)$ is isometrically isomorphic to $S^{1,0}(\mathcal{H})$ endowed with the nuclear norm $\|\cdot\|_1$.

7. *The nonlinear map $\iota : (\hat{\mathcal{H}}, D_p) \to (\hat{\mathcal{H}}, d_p)$ defined by*

$$\iota(x) = \begin{cases} \frac{x}{\sqrt{\|x\|}} & if \quad x \neq 0 \\ 0 & if \quad x = 0 \end{cases}$$

is bi-Lipschitz with the lower Lipschitz constant $\min(2^{\frac{1}{2}-\frac{1}{p}}, n^{\frac{1}{p}-\frac{1}{2}})$ and the upper Lipschitz constant $\sqrt{2}\max(n^{\frac{1}{2}-\frac{1}{p}}, 2^{\frac{1}{p}-\frac{1}{2}})$.

Proof. 1. The well-defined-ness of the metrics means that the metrics are the same for different choices inside the equivalence classes and is obvious from the definition of the metrics. Also immediately seen from the definition, $D_p(\hat{x}, \hat{y}) \geq 0$ for any $\hat{x}$, $\hat{y} \in \hat{\mathcal{H}}$ and $D_p(\hat{x}, \hat{y}) = 0$ if and only if $\hat{x} = \hat{y}$. Also $D_p(\hat{x}, \hat{y}) = D_p(\hat{y}, \hat{x})$ since $\|x - ay\|_p = \|y - a^{-1}x\|_p$ for any x, $y \in \mathcal{H}$, $|a| = 1$. Moreover, for any $\hat{x}$, $\hat{y}$, $\hat{z} \in \hat{\mathcal{H}}$, fix $D_p(\hat{x}, \hat{y}) = \|x - ay\|_p$, $D_p(\hat{y}, \hat{z}) = \|z - by\|$, then

$$\begin{aligned} D_p(\hat{x}, \hat{z}) &\leq \left\|x - ab^{-1}z\right\|_p = \|bx - az\|_p \\ &\leq \|bx - aby\|_p + \|aby - az\|_p = D_p(\hat{x}, \hat{y}) + D_p(\hat{y}, \hat{z}) \ . \end{aligned}$$

Therefore D_p is a metric. d_p is also a metric since $\|\cdot\|_p$ in the definition of d_p is the standard Schatten p-norm of a matrix.

2. For $p \leq q$, by Hölder's inequality we have for any $x = (x_1, x_2, ..., x_n) \in \mathcal{H}$ that $\sum_{i=1}^n |x_i|^p \leq n^{(1-\frac{p}{q})}(\sum_{i=1}^n |x_i|^q)^{\frac{p}{q}}$. Thus $\|x\|_p \leq n^{(\frac{1}{p}-\frac{1}{q})} \|x\|_q$. Also, since $\|\cdot\|_p$ is homogeneous, we can assume $\|x\|_p = 1$. Then $\sum_{i=1}^n |x_i|^q \leq \sum_{i=1}^n |x_i|^p = 1$. Thus $\|x\|_q \leq \|x\|_p$. Therefore, we have

$$D_q(\hat{x}, \hat{y}) = \|x - a_1 y\|_q \geq n^{(\frac{1}{q}-\frac{1}{p})} \|x - a_1 y\|_p \geq n^{(\frac{1}{q}-\frac{1}{p})} D_p(\hat{x}, \hat{y})$$

and $D_p(\hat{x}, \hat{y}) = \|x - a_2 y\|_p \geq \|x - a_2 y\|_q \geq D_q(\hat{x}, \hat{y})$ for some a_1, a_2 with magnitude 1. Hence

$$D_q(\hat{x}, \hat{y}) \leq D_p(\hat{x}, \hat{y}) \leq n^{(\frac{1}{p}-\frac{1}{q})} D_q(\hat{x}, \hat{y}) \ .$$

We see that $(D_p)_{1 \leq p \leq \infty}$ are equivalent. The second part follows then immediately.

3. The proof is similar to Part 2. Note that there are at most two σ_i's that are nonzero, so we have $2^{(\frac{1}{p}-\frac{1}{q})}$ instead of $n^{(\frac{1}{p}-\frac{1}{q})}$.

4. To prove that D_p and d_p are equivalent, we need only to show that each open ball with respect to D_p contains an open ball with respect to d_p, and vice versa. By (ii) and (iii), it is sufficient to consider the case when $p = 2$.

   First, we fix $x \in \mathcal{H} = \mathbb{C}^n$, $r > 0$. Let $R = \min(1, rn^{-2}(2\|x\|_\infty + 1)^{-1})$. Then for any $\hat{y}$ such that $D_2(\hat{x}, \hat{y}) < R$, we take y such that $\|x - y\| < R$, then

$\forall 1 \le i, j \le n$, $|x_i\overline{x_j} - y_i\overline{y_j}| = |x_i(\overline{x_j} - \overline{y_j}) + (x_i - y_i)\overline{y_j}| < |x_i|R + R(|x_i| + R) = R(2|x_i| + R) \le R(2|x_i| + 1) \le rn^{-2}$. Hence $d_2(\hat{x}, \hat{y}) = \|xx^* - yy^*\|_2 < n^2 \cdot rn^{-2} = r$.

On the other hand, we fix $x \in \mathcal{H} = \mathbb{C}^n$, $R > 0$. Let $r = R^2/\sqrt{2}$. Then for any $\hat{y}$ such that $d_2(\hat{x}, \hat{y}) < r$, we have

$$(d_2(\hat{x}, \hat{y}))^2 = \|x\|^4 + \|y\|^4 - 2|\langle x, y\rangle|^2 < r^2 = \frac{R^4}{2} \ .$$

But we also have

$$(D_2(\hat{x}, \hat{y}))^2 = \min_{|a|=1} \|x - ay\|^2 = \left\|x - \frac{\langle x, y\rangle}{|\langle x, y\rangle|}y\right\|^2 = \|x\|^2 + \|y\|^2 - 2|\langle x, y\rangle| \ ,$$

so

$$(D_2(\hat{x}, \hat{y}))^4 = \|x\|^4 + \|y\|^4 + 2\|x\|^2\|y\|^2 - 4(\|x\|^2 + \|y\|^2)|\langle x, y\rangle| + 4|\langle x, y\rangle|^2 \ .$$

Since $|\langle x, y\rangle| \le \|x\|\,\|y\| \le (\|x\|^2 + \|y\|^2)/2$, we can easily check that $(D_2(\hat{x}, \hat{y}))^4 \le 2(d_2(\hat{x}, \hat{y}))^2 < R^4$. Hence $D_2(\hat{x}, \hat{y}) < R$.

Thus D_2 and d_2 are indeed equivalent metrics. Therefore D_p and d_q are equivalent. Also, the imbedding i is not Lipschitz: if we take $x = (x_1, 0, \ldots, 0) \in \mathbb{C}^n$, then $D_2(\hat{x}, 0) = |x_1|$, $d_2(\hat{x}, 0) = |x_1|^2$.

5. First, for $p = 2$, for $\hat{x} \neq \hat{y}$ in $\hat{\mathcal{H}} - \{0\}$, we compute the quotient

$$\begin{aligned}
\rho(x,y) &= \frac{\|\kappa_\alpha(x) - \kappa_\alpha(y)\|^2}{D_2(x,y)^2} \\
&= \frac{\left\| \|x\|^{-1} xx^* - \|y\|^{-1} yy^* \right\|^2}{\|x\|^2 + \|y\|^2 - 2\,|\langle x, y\rangle|} \\
&= \frac{\|xx^*\|^2 \|y\|^2 + \|x\|^2 \|yy^*\|^2 - 2\,\|x\|\,\|y\|\,\mathrm{tr}(xx^*yy^*)}{\|x\|^4 \|y\|^2 + \|x\|^2 \|y\|^4 - 2\,\|x\|^2 \|y\|^2\,|x^*y|} \\
&= 1 + \frac{2\,\|x\|\,\|y\|\,(\|x\|\,\|y\|\,|x^*y| - \mathrm{tr}(xx^*yy^*))}{\|x\|^4 \|y\|^2 + \|x\|^2 \|y\|^4 - 2\,\|x\|^2 \|y\|^2\,|x^*y|} \\
&= 1 + \frac{2\,(\|x\|\,\|y\|\,|x^*y| - \mathrm{tr}(xx^*yy^*))}{\|x\|^3 \|y\| + \|x\| \|y\|^3 - 2\,\|x\|\,\|y\|\,|x^*y|}\ ,
\end{aligned}$$

where we used $\|xx^*\| = \|x\|^2$. For simplicity write $a = \|x\|$, $b = \|y\|$ and $t = |\langle x, y\rangle| \cdot (\|x\|\,\|y\|)^{-1}$. We have $a > 0$, $b > 0$ and $0 \leq t \leq 1$.

Now

$$\rho(x,y) = 1 + \frac{2(abt - abt^2)}{a^2 + b^2 - 2abt}\ .$$

Obviously $\rho(x,y) \geq 1$. Now we prove that $\rho(x,y) \leq 2$. Note that

$$1 + \frac{2(abt - abt^2)}{a^2 + b^2 - 2abt} \leq 2 \ \Leftrightarrow\ a^2 + b^2 - 4abt + 2abt^2 \geq 0\ ,$$

but

$$a^2 + b^2 - 4abt + 2abt^2 \geq 2ab - 4abt + 2abt^2 = 2ab(t-1)^2 \geq 0,$$

so we are done. Note that take any x, y with $\langle x, y\rangle = 0$ we would have $\rho(x,y) = 1$. On the other hand, taking $\|x\| = \|y\|$ and let $t \to 1$ we see that $\rho(x,y) = 2 - \epsilon$ is achievable for any small $\epsilon > 0$. Therefore the constants are optimal. The case where one of x and y is zero would not break the constraint of these two constants. Therefore after taking the square root, we get lower Lipschitz constant 1 and upper Lipschitz constant $\sqrt{2}$.

For other p, we use the results in (ii) and (iii) to get that the lower Lipschitz constant for κ_α is $\min(2^{\frac{1}{2}-\frac{1}{p}}, n^{\frac{1}{p}-\frac{1}{2}})$ and the upper Lipschitz constant is $\sqrt{2}\max(n^{\frac{1}{2}-\frac{1}{p}}, 2^{\frac{1}{p}-\frac{1}{2}})$.

6. This follows directly from the construction of the map.

7. This follows directly from Part 5 and Part 6.

□

The distance functions that we choose for the phase retrieval setting are natural. In some other settings, for instance, when we can fix a component of the original signal (see [52]), we cannot quotient out a global phase and will be forced to use the Euclidean distance. In this case we do not have the bi-Lipschitz property (see [52]) that is crucial for a stable reconstruction. We discuss the bi-Lipschitz property in the following sections.

3.1.3 The noisy measurement model

While the choice of p is not important here, we are particularly interested in D_2 (corresponding to the Euclidean distance) and d_1 (corresponding to the nuclear norm) for their importance in various settings in other problems.

In the following sections we are going to establish two important results: first, a phase retrievable frame always induces a bi-Lipschitz measurement map; second, the inverse of the measurement map can be extended to the entire Euclidean space $\mathbb{R}^m$ while the Lipschitz constant is increased by only a small number. Specifically,

suppose a is the lower Lipschitz bound of the measurement map, we get a reconstruction map $\omega : \mathbb{R}^m \to \hat{\mathcal{H}}$ such that

$$\mathrm{Lip}(\omega) \leq \frac{8.25}{a} \,. \tag{3.11}$$

Consider the map α (a similar discussion works for β). Assume an additive noise model $y = \alpha(x) + \nu$, where $\nu \in \mathbb{R}^m$ is the noise. For a signal $x_0 \in \hat{\mathcal{H}}$, and noise $\nu_1 \in \mathbb{R}^m$, let $y_1 = \alpha(x_0) + \nu_1 \in \mathbb{R}^m$ be the measurement vector, and let $x_1 = \omega(y_1)$ be the reconstructed signal. We have

$$D_2(x_0, x_1) = D_2\left(\omega(\alpha(x_0)), \omega(y_1)\right) \leq \mathrm{Lip}(\omega) \cdot \|\alpha(x_0) - y_1\| = \mathrm{Lip}(\omega) \cdot \|\nu_1\| \,.$$

Figure 3.1 is an illustration of this model. In fact, we have stability in a stronger sense. If we have two noisy measurements $y_1 = \alpha(x_0) + \nu_1$ and $y_2 = \alpha(x_0) + \nu_2$ of the signal x_0, then

$$D_2(x_1, x_2) = D_2\left(\omega(y_1), \omega(y_2)\right) \leq \mathrm{Lip}(\omega) \cdot \|y_1 - y_2\| = \mathrm{Lip}(\omega) \cdot \|\nu_1 - \nu_2\| \,.$$

3.2 Phase retrievability implies bi-Lipschitz property

In this section we study the relation between Phase retrievability and bi-Lipschitz property. It is obvious that if α and β, as defined in (3.1) and (3.2), are bi-Lipschitz with respect to the corresponding metrics, then they are phase retrievable. We focus on the converse. In this section, we shall assume that α and β are phase retrievable.

We first define three types of Lipschitz bounds for α and β respectively.

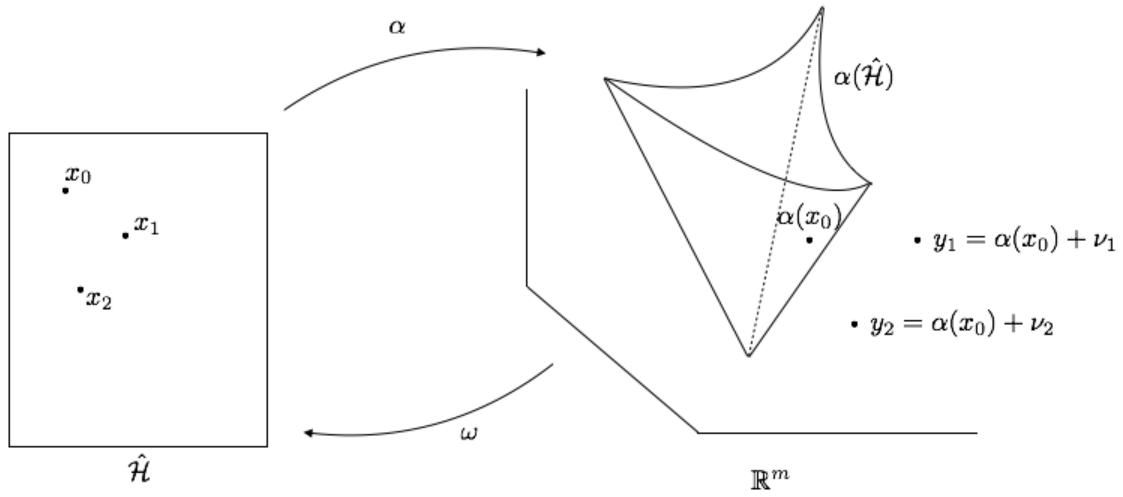

Figure 3.1: Illustration of the noisy measurement model

Definition 3.2.1. (Lipschitz bounds for α) *The following Lipschitz bounds are defined for the measurement α. The square roots of those bounds are called the Lipschitz constants.*

1. *The* global lower *and* upper Lipschitz bounds*, respectively:*

$$A_0 = \inf_{x,y \in \hat{H}} \frac{\|\alpha(x) - \alpha(y)\|_2^2}{D_2(x,y)^2} \ ,$$

$$B_0 = \sup_{x,y \in \hat{H}} \frac{\|\alpha(x) - \alpha(y)\|_2^2}{D_2(x,y)^2} \ ;$$

2. *The* type I local lower *and* upper Lipschitz bounds *at $z \in \hat{H}$, respectively:*

$$A(z) = \lim_{r \to 0} \inf_{\substack{x,y \in \hat{H} \\ D_2(x,z) < r \\ D_2(y,z) < r}} \frac{\|\alpha(x) - \alpha(y)\|_2^2}{D_2(x,y)^2} \ ,$$

$$B(z) = \lim_{r \to 0} \sup_{\substack{x,y \in \hat{H} \\ D_2(x,z) < r \\ D_2(y,z) < r}} \frac{\|\alpha(x) - \alpha(y)\|_2^2}{D_2(x,y)^2} \ ;$$

3. The type II local lower *and* upper Lipschitz bounds *at $z \in \hat{H}$, respectively:*

$$\tilde{A}(z) = \lim_{r \to 0} \inf_{\substack{x \in \hat{H} \\ D_2(x,z) < r}} \frac{\|\alpha(x) - \alpha(z)\|_2^2}{D_2(x,z)^2} \ ,$$

$$\tilde{B}(z) = \lim_{r \to 0} \sup_{\substack{x \in \hat{H} \\ D_2(x,z) < r}} \frac{\|\alpha(x) - \alpha(z)\|_2^2}{D_2(x,z)^2} \ .$$

Definition 3.2.2. (Lipschitz bounds for β) *The following Lipschitz bounds are defined for the measurement β. The square roots of those bounds are called the Lipschitz constants.*

1. The global lower *and* upper Lipschitz bounds*, respectively:*

$$a_0 = \inf_{x,y \in \hat{H}} \frac{\|\beta(x) - \beta(y)\|_2^2}{d_1(x,y)^2} \ ,$$

$$b_0 = \sup_{x,y \in \hat{H}} \frac{\|\beta(x) - \beta(y)\|_2^2}{d_1(x,y)^2} \ ;$$

2. The type I local lower *and* upper Lipschitz bounds *at $z \in \hat{H}$, respectively:*

$$a(z) = \lim_{r \to 0} \inf_{\substack{x,y \in \hat{H} \\ d_1(x,z) < r \\ d_1(y,z) < r}} \frac{\|\beta(x) - \beta(y)\|_2^2}{d_1(x,y)^2} \ ,$$

$$b(z) = \lim_{r \to 0} \sup_{\substack{x,y \in \hat{H} \\ d_1(x,z) < r \\ d_1(y,z) < r}} \frac{\|\beta(x) - \beta(y)\|_2^2}{d_1(x,y)^2} \ ;$$

3. The type II local lower *and* upper Lipschitz bounds *at $z \in \hat{H}$, respectively:*

$$\tilde{a}(z) = \lim_{r \to 0} \inf_{\substack{x \in \hat{H} \\ d_1(x,z) < r}} \frac{\|\beta(x) - \beta(z)\|_2^2}{d_1(x,z)^2} \ ,$$

$$\tilde{b}(z) = \lim_{r \to 0} \sup_{\substack{x \in \hat{H} \\ d_1(x,z) < r}} \frac{\|\beta(x) - \beta(z)\|_2^2}{d_1(x,z)^2} \ .$$

From the definitions, we have the following lemma due to homogeneity.

Lemma 3.2.3. *The Lipschitz bounds defined above satisfy the following relations:*

1. $A_0 = A(0)$, $B_0 = B(0)$, $a_0 = a(0)$, $b_0 = b(0)$.

2. *For* $z \neq 0$, $A(z) = A(z/\|z\|)$, $B(z) = B(z/\|z\|)$, $a(z) = a(z/\|z\|)$, $b(z) = b(z/\|z\|)$.

In the following sections we are going to establish the bi-Lipschitz properties for both α and β, given that they are phase retrievable.

3.2.1 The bi-Lipschitz property for the magnitude measurement map

3.2.1.1 The case $\mathcal{H} = \mathbb{R}^n$

For an index set $I \subset \{1, 2, \cdots, m\}$, let $\mathcal{F}[I] = \{f_k, k \in I\}$ denote the frame subset indexed by I. Also, let $\sigma_1^2[I]$ and $\sigma_n^2[I]$ denote the upper and lower frame bound of the set $\mathcal{F}[I]$, respectively. It is straightforward to see that they respectively correspond to the largest and smallest eigenvalues of $\sum_{k\in I} f_k f_k^*$, that is,

$$\sigma_1^2[I] = \lambda_1 \left(\sum_{k\in I} f_k f_k^* \right) \tag{3.12}$$

and

$$\sigma_n^2[I] = \lambda_n \left(\sum_{k\in I} f_k f_k^* \right) . \tag{3.13}$$

The following theorem summarizes some of the main results in [19].

Theorem 3.2.4. (see [19]) *Let $\mathcal{F} \subset \mathbb{R}^n$ be a phase retrievable frame for $\mathbb{R}^n$. Let A and B denote its optimal lower and upper frame bound, respectively. Then*

1. *For every* $0 \neq x \in \mathbb{R}^n$, $A(x) = \sigma_n^2[\text{supp}(\alpha(x))]$ *where* $\text{supp}(\alpha(x)) = \{k, \langle x, f_k \rangle \neq 0\}$;

2. *For every* $x \in \mathbb{R}^n$, $\tilde{A}(x) = A(x)$;

3. $A_0 = A(0) = \min_{I \subset \{1,2,\cdots,m\}}(\sigma_n^2[I] + \sigma_n^2[I^c])$;

4. *For every* $x \in \mathbb{R}^n$, $B(x) = \tilde{B}(x) = B$;

5. $B_0 = B(0) = \tilde{B}(0) = B$.

3.2.1.2 The case $\mathcal{H} = \mathbb{C}^n$

We analyze the complex case by doing a realification first. Consider the $\mathbb{R}$-linear map $\mathbf{j} : \mathbb{C}^n \to \mathbb{R}^{2n}$ defined by

$$\mathbf{j}(z) = \begin{bmatrix} \text{real}(z) \\ \\ \text{imag}(z) \end{bmatrix}.$$

This realification is studied in detail in [12]. We call $\mathbf{j}(z)$ the realification of z. For simplicity, in this paper we will denote $\xi = \mathbf{j}(x)$, $\eta = \mathbf{j}(y)$, $\zeta = \mathbf{j}(z)$, $\varphi = \mathbf{j}(f)$, $\delta = \mathbf{j}(d)$, respectively.

For a frame set $\mathcal{F} = \{f_1, f_2, \cdots, f_m\}$, define the symmetric operator

$$\Phi_k = \varphi_k \varphi_k^T + J\varphi_k \varphi_k^T J^T, \quad k = 1, 2, \cdots, m.$$

where

$$J = \begin{bmatrix} 0 & -I \\ \\ I & 0 \end{bmatrix} \tag{3.14}$$

is a matrix in $\mathbb{R}^{2n \times 2n}$.

Also, define $\mathcal{S} : \mathbb{R}^{2n} \to \mathrm{Sym}(\mathbb{R}^{2n})$ by

$$\mathcal{S}(\xi) = \begin{cases} 0 & , \text{ if } \xi = 0 \\ \sum_{k:\Phi_k\xi\neq 0} \frac{1}{\langle \Phi_k\xi, \xi\rangle} \Phi_k \xi\xi^T \Phi_k & , \text{ if } \xi \neq 0 \end{cases} .$$

We have the following result:

Theorem 3.2.5. *Let $\mathcal{F} \subset \mathbb{C}^n$ be a phase retrievable frame for $\mathbb{C}^n$. Let A and B denote its optimal lower and upper frame bound, respectively. For any $z \in \mathbb{C}^n$, let $\zeta = \mathbf{j}(z)$ be its realification. Then*

1. *For every $0 \neq z \in \mathbb{C}^n$, $A(z) = \lambda_{2n-1}(\mathcal{S}(\zeta))$;*
2. *$A_0 = A(0) > 0$;*
3. *For every $z \in \mathbb{C}^n$, $\tilde{A}(z) = \lambda_{2n-1}\left(\mathcal{S}(\zeta) + \sum_{k:\langle z, f_k\rangle = 0} \Phi_k\right)$;*
4. *$\tilde{A}(0) = A$;*
5. *For every $z \in \mathbb{C}^n$, $B(z) = \tilde{B}(z) = \lambda_1\left(\mathcal{S}(\zeta) + \sum_{k:\langle z, f_k\rangle = 0} \Phi_k\right)$;*
6. *$B_0 = B(0) = \tilde{B}(0) = B$.*

To prove the theorem we need the following lemma.

Lemma 3.2.6. *Fix $x \in \mathbb{C}^n$ and $z \in \mathbb{C}^n$. Let $\xi = \mathbf{j}(x)$ and $\zeta = \mathbf{j}(z)$ be their realifications, respectively. Let $\xi_0 \in \hat{\xi} := \{\mathbf{j}(\tilde{x}) \in \mathbb{R}^{2n} : \tilde{x} \in \hat{x}\}$ be a point in the equivalence class that satisfies $D_2(x, z) = \|\xi_0 - \zeta\|$. Then it is necessary that*

$$\langle \xi_0, J\zeta\rangle = 0 \tag{3.15}$$

and

$$\langle \xi_0, \zeta \rangle \geq 0 \ , \tag{3.16}$$

where J is defined as in (3.14).

Proof. For $\theta \in [0, 2\pi)$ define

$$U(\theta) := \cos(\theta) I + \sin(\theta) J \ .$$

Then it is easy to compute that

$$\mathbf{j}(e^{i\theta} x) = U(\theta)\xi \ .$$

Therefore,

$$D_2(x, z) = \min_{\theta \in [0,2\pi)} \|U(\theta)\xi - \zeta\|^2 = \|\xi\|^2 + \|\zeta\|^2 - 2 \max_{\theta \in [0,2\pi)} \langle U(\theta)\xi, \zeta \rangle \ .$$

If $\langle U(\theta)\xi, \zeta \rangle$ is constantly zero, then we are done. Otherwise, note that

$$\max_{\theta \in [0,2\pi)} \langle U(\theta)\xi, \zeta \rangle = \left(\langle \xi, \zeta \rangle^2 + \langle J\xi, \zeta \rangle^2 \right)^{\frac{1}{2}}$$

and the maximum is achieved at $\theta = \theta_0$ if and only if

$$\cos(\theta_0) = \frac{\langle \xi, \zeta \rangle}{\left(\langle \xi, \zeta \rangle^2 + \langle J\xi, \zeta \rangle^2 \right)^{\frac{1}{2}}}$$

and

$$\sin(\theta_0) = \frac{\langle J\xi, \zeta \rangle}{\left(\langle \xi, \zeta \rangle^2 + \langle J\xi, \zeta \rangle^2 \right)^{\frac{1}{2}}} \ .$$

Now we can compute

$$\begin{aligned}
\langle \xi_0, J\zeta \rangle &= \langle U(\theta_0)\xi, J\zeta \rangle \\
&= \cos(\theta_0) \langle \xi, J\zeta \rangle + \sin(\theta_0) \langle J\xi, J\zeta \rangle \\
&= \frac{\langle \xi, \zeta \rangle}{\left(\langle \xi, \zeta \rangle^2 + \langle J\xi, \zeta \rangle^2 \right)^{\frac{1}{2}}} \langle \xi, J\zeta \rangle + \frac{\langle J\xi, \zeta \rangle}{\left(\langle \xi, \zeta \rangle^2 + \langle J\xi, \zeta \rangle^2 \right)^{\frac{1}{2}}} \langle J\xi, J\zeta \rangle \\
&= \frac{\langle \xi, \zeta \rangle}{\left(\langle \xi, \zeta \rangle^2 + \langle J\xi, \zeta \rangle^2 \right)^{\frac{1}{2}}} \langle -J\xi, \zeta \rangle + \frac{\langle J\xi, \zeta \rangle}{\left(\langle \xi, \zeta \rangle^2 + \langle J\xi, \zeta \rangle^2 \right)^{\frac{1}{2}}} \langle \xi, \zeta \rangle \\
&= 0 \ .
\end{aligned}$$

So we get (3.15). (3.16) is obvious. □

Now we are ready to prove the theorem.

Proof. (of Theorem 3.2.5)

1. Denote

$$p(x,y) := \frac{\|\alpha(x) - \alpha(y)\|^2}{D_2(x,y)^2}, \qquad x, y \in \mathbb{C}^n, \ \hat{x} \neq \hat{y}. \tag{3.17}$$

We can represent this quotient in terms of ξ and η. It is easy to compute that

$$\begin{aligned}
p(x,y) &= P(\xi, \eta) \\
&:= \frac{\sum_{k=1}^m \langle \Phi_k \xi, \xi \rangle + \langle \Phi_k \eta, \eta \rangle - 2\sqrt{\langle \Phi_k \xi, \xi \rangle \langle \Phi_k \eta, \eta \rangle}}{\|\xi\|^2 + \|\eta\|^2 - 2\sqrt{\langle \xi, \eta \rangle^2 + \langle \xi, J\eta \rangle^2}} \ .
\end{aligned} \tag{3.18}$$

Fix $r > 0$. Take ξ, $\eta \in \mathbb{R}^{2n}$ that satisfy

$$D_2(x,z) = \|\xi - \zeta\| < r$$

and

$$D_2(y,z) = \|\eta - \zeta\| < r \ .$$

Let

$$\mu = \frac{\xi + \eta}{2}$$

and

$$\nu = \frac{\xi - \eta}{2} \ .$$

Then $\|\nu\| < r$.

Note that for r small enough we have that $\|\mu\| > \|\nu\|$ and that

$$\Phi_k \zeta \neq 0 \Rightarrow \Phi_k \mu \neq 0 \ .$$

Thus

$$
\begin{aligned}
P(\xi,\eta) &= \Bigg(\sum_{k=1}^{m} \langle \Phi_k(\mu+\nu),\mu+\nu\rangle + \langle \Phi_k(\mu-\nu),\mu-\nu\rangle - \\
&\qquad 2\sqrt{\langle \Phi_k(\mu+\nu),\mu+\nu\rangle \langle \Phi_k(\mu-\nu),\mu-\nu\rangle}\Bigg)\cdot \\
&\qquad \left(\|\mu+\nu\|^2 + \|\mu-\nu\|^2 - 2\sqrt{\langle \mu+\nu,\mu-\nu\rangle^2 + \langle \mu+\nu, J(\mu-\nu)\rangle^2}\right)^{-1} \\
&= \Bigg(\sum_{k=1}^{m} \langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle - \sqrt{(\langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle)^2 - 4\langle \Phi_k\mu,\nu\rangle^2}\Bigg)\cdot \\
&\qquad \left(\|\mu\|^2 + \|\nu\|^2 - \sqrt{\|\mu\|^4 + \|\nu\|^4 - 2\|\mu\|^2\|\nu\|^2 + 4\langle \mu, J\nu\rangle^2}\right)^{-1} \\
&\geq \Bigg(\sum_{k:\Phi_k\zeta\neq 0} \langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle - \sqrt{(\langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle)^2 - 4\langle \Phi_k\mu,\nu\rangle^2}\Bigg)\cdot \\
&\qquad \left(\|\mu\|^2 + \|\nu\|^2 - \sqrt{\|\mu\|^4 + \|\nu\|^4 - 2\|\mu\|^2\|\nu\|^2}\right)^{-1} \\
&= \frac{1}{2\|\nu\|^2} \sum_{k:\Phi_k\zeta\neq 0} \langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle - \sqrt{(\langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle)^2 - 4\langle \Phi_k\mu,\nu\rangle^2} \\
&= \frac{1}{2\|\nu\|^2} \sum_{k:\Phi_k\zeta\neq 0} \langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle - \\
&\qquad \langle \Phi_k\mu,\mu\rangle \sqrt{\left(1+\frac{\langle \Phi_k\nu,\nu\rangle}{\langle \Phi_k\mu,\mu\rangle}\right)^2 - 4\frac{\langle \Phi_k\mu,\nu\rangle^2}{\langle \Phi_k\mu,\mu\rangle^2}} \\
&= \frac{1}{2\|\nu\|^2} \sum_{k:\Phi_k\zeta\neq 0} \langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle - \\
&\qquad \langle \Phi_k\mu,\mu\rangle \sqrt{1+2\frac{\langle \Phi_k\nu,\nu\rangle}{\langle \Phi_k\mu,\mu\rangle} + \frac{\langle \Phi_k\nu,\nu\rangle^2}{\langle \Phi_k\mu,\mu\rangle^2} - 4\frac{\langle \Phi_k\mu,\nu\rangle^2}{\langle \Phi_k\mu,\mu\rangle^2}} \\
&= \frac{1}{2\|\nu\|^2} \sum_{k:\Phi_k\zeta\neq 0} \langle \Phi_k\mu,\mu\rangle + \langle \Phi_k\nu,\nu\rangle - \\
&\qquad \langle \Phi_k\mu,\mu\rangle \left(1+\frac{\langle \Phi_k\nu,\nu\rangle}{\langle \Phi_k\mu,\mu\rangle} - 2\frac{\langle \Phi_k\mu,\nu\rangle^2}{\langle \Phi_k\mu,\mu\rangle^2}\right) + O(\|\nu\|^4) \\
&= \sum_{k:\Phi_k\zeta\neq 0} \frac{\langle \Phi_k\mu,\nu\rangle^2}{\langle \Phi_k\mu,\mu\rangle\|\nu\|^2} + O(\|\nu\|^2) \\
&= \frac{1}{\|\nu\|^2}\langle \mathcal{S}(\mu)\nu,\nu\rangle + O(\|\nu\|^2)\ .
\end{aligned}
$$

Note that

$$|\langle J\mu, \nu\rangle| = |\langle J\mu, \nu\rangle - \langle J\zeta, \nu\rangle| \leq \|J\mu - J\zeta\| \, \|\nu\| = \|\mu - \zeta\| \, \|\nu\| \tag{3.19}$$

since $\langle J\zeta, \nu\rangle = 0$ by Lemma 3.2.6. Together with $\|\mu - \zeta\| < r$ we have

$$\|P_{J\mu}\nu\| = \frac{|\langle J\mu, \nu\rangle|}{\|J\mu\|} = \frac{|\langle J\mu, \nu\rangle|}{\|\mu\|} \leq \frac{r \, \|\nu\|}{\|\mu\|} \ ,$$

and thus

$$\left\|P_{J\mu}^{\perp}\nu\right\|^2 \geq \left(1 - \frac{r^2}{\|\mu\|^2}\right) \|\nu\|^2 \ .$$

As a consequence, we have

$$\begin{aligned} P(\xi, \eta) &= \frac{1}{\|\nu\|^2} \left\langle \mathcal{S}(\mu) P_{J\mu}^{\perp}\nu, P_{J\mu}^{\perp}\nu \right\rangle + O(\|\nu\|^2) \\ &\geq \frac{1}{\left\|P_{J\mu}^{\perp}\nu\right\|^2} \left\langle \mathcal{S}(\mu) P_{J\mu}^{\perp}\nu, P_{J\mu}^{\perp}\nu \right\rangle \left(1 - \frac{r^2}{\|\mu\|^2}\right) + O(r^2) \\ &\geq \left(1 - \frac{r^2}{\|\mu\|^2}\right) \lambda_{2n-1}\left(\mathcal{S}(\mu)\right) + O(r^2) \ . \end{aligned}$$

Take $r \to 0$, by the continuity of eigenvalues with respect to matrix entries we have that

$$A(z) \geq \lambda_{2n-1}(\mathcal{S}(\zeta)) \ . \tag{3.20}$$

On the other hand, take E_{2n-1} to be the unit-norm eigenvector correspondent to $\lambda_{2n-1}(\mathcal{S}(\zeta))$. For each $r > 0$, take $\xi = \zeta + \frac{r}{2}E_{2n-1}$ and $\eta = \zeta - \frac{r}{2}E_{2n-1}$. Then

$$p(x, y) = P(\xi, \eta) = \lambda_{2n-1}(\mathcal{S}(\zeta)) \ .$$

Hence

$$A(z) \leq \lambda_{2n-1}(\mathcal{S}(\zeta)) \ .$$

Together with (3.20) we have

$$A(z) = \lambda_{2n-1}(\mathcal{S}(\zeta)) \ .$$

2. Assume on the contrary that $A_0 = 0$, then for any $N \in \mathbb{N}$, there exist x_N, $y_N \in H$ for which

$$p(x_N, y_N) = \frac{\|\alpha(x_N) - \alpha(y_N)\|^2}{D_2(x_N, y_N)^2} \leq \frac{1}{N}. \tag{3.21}$$

Without loss of generality we assume that $\|x_N\| \geq \|y_N\|$ for each N, for otherwise we can just swap the role of x_N and y_N. Also due to homogeneity we assume $\|x_N\| = 1$. By compactness of the closed ball $\mathcal{B}_1(0) = \{x \in H : \|x\| \leq 1\}$ in $H = \mathbb{C}^n$, there exist convergent subsequences of $\{x_N\}_{N\in\mathbb{N}}$ and $\{y_N\}_{N\in\mathbb{N}}$, which to avoid overuse of notations we still denote as $\{x_N\}_{N\in\mathbb{N}} \to x_0 \in H$ and $\{y_N\}_{N\in\mathbb{N}} \to y_0 \in H$.

Since $\|x_0\| = 1$ we have from (i) that $A(x_0) > 0$. Note that $D_2(x_N, y_N) \leq \|x_N\| + \|y_N\| \leq 2$, so by (3.21) we have $\|\alpha(x_N) - \alpha(y_N)\| \to 0$. That is, $\|\alpha(x_0) - \alpha(y_0)\| = 0$. By injectivity we have $x_0 = y_0$ in $\hat{H}$. By Theorem 3.2.5 Part 1,

$$p(x_N, y_N) \geq A(x_0) - 1/N > 1/N$$

for N large enough. This is a contradiction with (3.21).

3. The case $z = 0$ is an easy computation. We now present the proof for $z \neq 0$. First we consider $p(x, z) = P(\xi, \zeta)$ as defined in (3.18). Fix $r > 0$. Take $\xi \in \mathbb{R}^{2n}$ that satisfy $D_2(x, z) = \|\xi - \zeta\| < r$. Let

$$d = x - z$$

and

$$\delta = \mathbf{j}(d) = \xi - \zeta \ .$$

Note that

$$P(\xi,\zeta) = \frac{\sum_{k=1}^{m} \langle \Phi_k \xi, \xi \rangle + \langle \Phi_k \zeta, \zeta \rangle - 2\sqrt{\langle \Phi_k \xi, \xi \rangle \langle \Phi_k \zeta, \zeta \rangle}}{\|\xi\|^2 + \|\zeta\|^2 - 2\sqrt{\langle \xi, \zeta \rangle^2 + \langle \xi, J\zeta \rangle^2}} \ .$$

We can compute its numerator

$$\begin{aligned}
&\sum_{k=1}^{m} \langle \Phi_k \xi, \xi \rangle + \langle \Phi_k \zeta, \zeta \rangle - 2\sqrt{\langle \Phi_k \xi, \xi \rangle \langle \Phi_k \zeta, \zeta \rangle} \\
= \ & \sum_{k=1}^{m} \langle \Phi_k \zeta, \zeta \rangle + 2 \langle \Phi_k \zeta, \delta \rangle + \langle \Phi_k \delta, \delta \rangle + \langle \Phi_k \zeta, \zeta \rangle - \\
& \qquad 2\sqrt{(\langle \Phi_k \zeta, \zeta \rangle + 2 \langle \Phi_k \zeta, \delta \rangle + \langle \Phi_k \delta, \delta \rangle) \cdot \langle \Phi_k \zeta, \zeta \rangle} \\
= \ & \sum_{k:\Phi_k \zeta \neq 0} 2 \langle \Phi_k \zeta, \zeta \rangle + 2 \langle \Phi_k \zeta, \delta \rangle + \langle \Phi_k \delta, \delta \rangle - \\
& \qquad 2 \langle \Phi_k \zeta, \zeta \rangle \left(1 + \frac{\langle \Phi_k \zeta, \zeta \rangle \langle \Phi_k \zeta, \delta \rangle + \frac{1}{2} \langle \Phi_k \zeta, \zeta \rangle \langle \Phi_k \delta, \delta \rangle}{\langle \Phi_k \zeta, \zeta \rangle^2} - \right. \\
& \qquad \left. \frac{1}{8} \cdot \frac{4 \langle \Phi_k \zeta, \zeta \rangle^2 \langle \Phi_k \zeta, \delta \rangle^2}{\langle \Phi_k \zeta, \zeta \rangle^4} + O\left(\|\delta\|^3 \right) \right) + \sum_{k:\Phi_k \zeta = 0} \langle \Phi_k \delta, \delta \rangle \\
= \ & \sum_{k:\Phi_k \zeta \neq 0} \frac{\langle \Phi_k \zeta, \delta \rangle^2}{\langle \Phi_k \zeta, \zeta \rangle} + \sum_{k:\Phi_k \zeta = 0} \langle \Phi_k \delta, \delta \rangle + O\left(\|\delta\|^3 \right) \ ;
\end{aligned}$$

and its denominator

$$\begin{aligned}
&\|\xi\|^2 + \|\zeta\|^2 - 2\sqrt{\langle \xi, \zeta \rangle^2 + \langle \xi, J\zeta \rangle^2} \\
= \ & 2\|\zeta\|^2 + \|\delta\|^2 + 2\langle \zeta, \delta \rangle - 2\|\zeta\|^2 \left(1 + \right. \\
& \qquad \left. \frac{\|\zeta\|^2 \langle \zeta, \delta \rangle + \frac{1}{2} \langle \zeta, \delta \rangle + \frac{1}{2} \langle J\zeta, \delta \rangle^2}{\|\zeta\|^4} - \frac{4\|\zeta\|^4 \langle \zeta, \delta \rangle^2}{8\|\zeta\|^8} + O\left(\|\delta\|^3 \right) \right) \\
= \ & \|\delta\|^2 + O\left(\|\delta\|^3 \right) \ .
\end{aligned}$$

We used Lemma 3.2.6 to get $\langle J\zeta, \delta \rangle = 0$ in the above.

Take $r \to 0$, we see that

$$\tilde{A}(z) \geq \lambda_{2n-1}\left(\mathcal{S}(\zeta) + \sum_{k:\langle z, f_k\rangle = 0} \Phi_k\right) .$$

Let $\tilde{E}_{2n-1}$ be the unit-norm eigenvector correspondening to

$$\lambda_{2n-1}\left(\mathcal{S}(\zeta) + \sum_{k:\langle z, f_k\rangle = 0} \Phi_k\right) .$$

Note that $\left\langle J\zeta, \tilde{E}_{2n-1}\right\rangle = 0$ since $\mathcal{S}(\zeta)J\zeta = 0$ and $\Phi_k J\zeta = J\Phi_k\zeta = 0$ for each k with $\langle z, f_k\rangle = 0$. Take $\xi = \zeta + \frac{r}{2}\tilde{E}_{2n-1}$ for each r, we again also have

$$\tilde{A}(z) \leq \lambda_{2n-1}\left(\mathcal{S}(\zeta) + \sum_{k:\langle z, f_k\rangle = 0} \Phi_k\right) .$$

Therefore

$$\tilde{A}(z) = \lambda_{2n-1}\left(\mathcal{S}(\zeta) + \sum_{k:\langle z, f_k\rangle = 0} \Phi_k\right) .$$

4. Take $z = 0$ in Part 3.

5. $\tilde{B}(z)$ can be computed in a similar way as in (iii) (in particular, the expansion for $P(\xi, \zeta)$ is exactly the same). We compute $B(z)$. $B(0)$ is computed in [22], Lemma 16. Now we consider $z \neq 0$. Use the same notations as in (3.18). Fix $r > 0$. Again, take ξ, $\eta \in \mathbb{R}^{2n}$ that satisfy $D_2(x, z) = \|\xi - \zeta\| < r$ and $D_2(y, z) = \|\eta - \zeta\| < r$. Let $\mu = (\xi + \eta)/2$ and $\nu = (\xi - \eta)/2$. Also let $\delta_1 = \xi - \zeta$ and $\delta_2 = \eta - \zeta$. Recall that

$$\begin{aligned} P(\xi, \eta) &= \frac{\sum_{k=1}^{m} \langle \Phi_k\xi, \xi\rangle + \langle \Phi_k\eta, \eta\rangle - 2\sqrt{\langle \Phi_k\xi, \xi\rangle \langle \Phi_k\eta, \eta\rangle}}{\|\xi\|^2 + \|\eta\|^2 - 2\sqrt{\langle \xi, \eta\rangle^2 + \langle \xi, J\eta\rangle^2}} \\ &= \sum_{k=1}^{m} \frac{\langle \Phi_k\xi, \xi\rangle + \langle \Phi_k\eta, \eta\rangle - 2\sqrt{\langle \Phi_k\xi, \xi\rangle \langle \Phi_k\eta, \eta\rangle}}{\|\xi\|^2 + \|\eta\|^2 - 2\sqrt{\langle \xi, \eta\rangle^2 + \langle \xi, J\eta\rangle^2}} . \end{aligned}$$

Now we compute it as $\sum_{k=1}^{m} = \sum_{k:\Phi_k\zeta\neq 0} + \sum_{k:\Phi_k\zeta=0}$. Again,

$$\begin{aligned}
&\sum_{k:\Phi_k\zeta\neq 0} \frac{\langle\Phi_k\xi,\xi\rangle + \langle\Phi_k\eta,\eta\rangle - 2\sqrt{\langle\Phi_k\xi,\xi\rangle\langle\Phi_k\eta,\eta\rangle}}{\|\xi\|^2 + \|\eta\|^2 - 2\sqrt{\langle\xi,\eta\rangle^2 + \langle\xi,J\eta\rangle^2}} \\
= &\sum_{k:\Phi_k\zeta\neq 0} \frac{\langle\Phi_k\mu,\mu\rangle + \langle\Phi_k\nu,\nu\rangle - \sqrt{(\langle\Phi_k\mu,\mu\rangle + \langle\Phi_k\nu,\nu\rangle)^2 - 4\langle\Phi_k\mu,\nu\rangle^2}}{\|\mu\|^2 + \|\nu\|^2 - \sqrt{\|\mu\|^4 + \|\nu\|^4 - 2\|\mu\|^2\|\nu\|^2 + 4\langle\mu,J\nu\rangle^2}} .
\end{aligned} \tag{3.22}$$

Using the same computation as in (i), we get that the numerator is

$$\begin{aligned}
&\sum_{k:\Phi_k\zeta\neq 0} \langle\Phi_k\mu,\mu\rangle + \langle\Phi_k\nu,\nu\rangle - \sqrt{(\langle\Phi_k\mu,\mu\rangle + \langle\Phi_k\nu,\nu\rangle)^2 - 4\langle\Phi_k\mu,\nu\rangle^2} \\
= \; &2\langle\mathcal{S}(\mu)\nu,\nu\rangle + O(\|\nu\|^4) .
\end{aligned}$$

Since $\mu \neq 0$, the denominator is

$$\begin{aligned}
&\|\mu\|^2 + \|\nu\|^2 - \sqrt{\|\mu\|^4 + \|\nu\|^4 - 2\|\mu\|^2\|\nu\|^2 + 4\langle\mu,J\nu\rangle^2} \\
= \; &\|\mu\|^2 + \|\nu\|^2 - \|\mu\|^2\sqrt{1 + \frac{\|\nu\|^4}{\|\mu\|^4} - \frac{2\|\nu\|^2}{\|\mu\|^2} + \frac{4\langle\mu,J\nu\rangle^2}{\|\mu\|^4}} \\
= \; &\|\mu\|^2 + \|\nu\|^2 - \|\mu\|^2\left(1 - \frac{\|\nu\|^2}{\|\mu\|^2} + \frac{2\langle\mu,J\nu\rangle^2}{\|\mu\|^4}\right) + O(\|\nu\|^4) \\
= \; &2\|\nu\|^2 - \frac{2\langle J\mu,\nu\rangle^2}{\|\mu\|^2} + O(\|\nu\|^4) \\
= \; &2\|\nu\|^2 + O(\|\nu\|^4) \quad \text{by (3.19)}.
\end{aligned} \tag{3.23}$$

Also we can compute using the denominator as above (note that $\nu = (\delta_1 - \delta_2)/2$) that

$$\begin{aligned}
&\sum_{k:\Phi_k\zeta=0} \frac{\langle\Phi_k\xi,\xi\rangle + \langle\Phi_k\eta,\eta\rangle - 2\sqrt{\langle\Phi_k\xi,\xi\rangle\langle\Phi_k\eta,\eta\rangle}}{\|\xi\|^2 + \|\eta\|^2 - 2\sqrt{\langle\xi,\eta\rangle^2 + \langle\xi,J\eta\rangle^2}} \\
= &\sum_{k:\Phi_k\zeta=0} \frac{\left(\left\|\Phi_k^{1/2}\delta_1\right\| - \left\|\Phi_k^{1/2}\delta_2\right\|\right)^2}{\|\delta_1 - \delta_2\|^2 + O(\|\nu\|^4)} .
\end{aligned} \tag{3.24}$$

Now put together (3.22), (3.23) and (3.24), we get

$$P(\xi,\eta) \;=\; \frac{\langle \mathcal{S}(\mu)\nu,\nu\rangle + O(\|\nu\|^4)}{\|\nu\|^2 + O(\|\nu\|^4)} + \sum_{k:\Phi_k\zeta=0} \frac{\left(\left\|\Phi_k^{1/2}\delta_1\right\| - \left\|\Phi_k^{1/2}\delta_2\right\|\right)^2}{\|\delta_1-\delta_2\|^2 + O(\|\nu\|^4)} \ .$$

Note that

$$\left(\left\|\Phi_k^{1/2}\delta_1\right\| - \left\|\Phi_k^{1/2}\delta_2\right\|\right)^2 \le \langle \Phi_k(\delta_1-\delta_2),\delta_1-\delta_2\rangle$$

since it is equivalent to

$$\langle \Phi_k\delta_1,\delta_1\rangle \langle \Phi_k\delta_2,\delta_2\rangle \ge \left(\langle \Phi_k\delta_1,\delta_2\rangle\right)^2 \ , \tag{3.25}$$

which is the Cauchy-Schwarz inequality. Therefore, we have that

$$\begin{aligned} P(\xi,\eta) \;&\le\; \frac{\left\langle (\mathcal{S}(\mu) + \sum_{k:\Phi_k\zeta=0}\Phi_k)\nu,\nu\right\rangle + O(\|\nu\|^4)}{\|\nu\|^2 + O(\|\nu\|^4)} \\ &\le\; \lambda_1\left(\mathcal{S}(\mu) + \sum_{k:\Phi_k\zeta=0}\Phi_k\right) + O(r^2) \ . \end{aligned}$$

Take $r \to 0$ we have

$$B(z) \le \lambda_1\left(\mathcal{S}(\zeta) + \sum_{k:\Phi_k\zeta=0}\Phi_k\right) \ .$$

Again we get the other direction of the above inequality by taking $\xi = \zeta + \frac{r}{2}E_1$ and $\eta = \zeta - \frac{r}{2}E_1$ for each $r > 0$ where E_1 is the unit-norm eigenvector correspondent to $\lambda_1\left(\mathcal{S}(\zeta) + \sum_{k:\langle z,f_k\rangle=0}\Phi_k\right)$. Note that for each r, the equality in (3.25) holds for this pair of ξ and η.

6. Take $z = 0$ in Part 5.

□

3.2.2 The bi-Lipschitz property for the square measurement map

The nonlinear map β as defined in (3.2) naturally induces a linear map between $\mathrm{Sym}(\mathcal{H})$ and $\mathbb{R}^m$:

$$\mathcal{A} : \mathrm{Sym}(\mathcal{H}) \to \mathbb{R}^m \qquad \mathcal{A}(T) = (\langle Tf_k, f_k\rangle)_{1\le k\le m} \ , \tag{3.26}$$

where $\mathrm{Sym}(\mathcal{H})$ is the space of symmetric operator on $\mathcal{H}$. Note that the map β is injective if and only if $\mathcal{A}$ restricted to $\mathcal{S}^{1,0}(\mathcal{H})$ is injective.

The following theorem summarizes the results on the bi-Lipschitz properties for β:

Theorem 3.2.7. (see [12,19]) *Let $\mathcal{F}$ be a phase retrievable frame for $H = \mathbb{C}^n$. Then*

1. *the global lower Lipschitz bound $a_0 > 0$;*

2. *the global upper Lipschitz bound $b_0 < \infty$, and*

$$\begin{aligned} b_0 &= \max_{\|x\|=\|y\|=1} \sum_{k=1}^{m} \left(\mathrm{real}\left(\langle x, f_k\rangle \langle f_k, y\rangle\right)\right)^2 \\ &= \max_{\|x\|=1} \sum_{k=1}^{m} |\langle x, f_k\rangle|^4 \\ &= \|T\|^4_{B(H,l^4_m)} \ , \end{aligned}$$

where $T : H \to \mathbb{C}^m$ is the analysis operator defined by $x \mapsto (\langle x, f_k\rangle)_{k=1}^m$, and $l^4_m := (\mathbb{C}^m, \|\cdot\|_4)$.

Remark 3.2.8. *An upper bound of b_0 is given by*

$$b_0 \le B \left(\max_{1\le k\le m} \|f_k\|\right)^2 \le B^2 \ ,$$

where B is the upper frame bound of $\mathcal{F}$.

We give an expression of the local Lipschitz bounds as well. Define $\mathcal{R} : \mathbb{R}^{2n} \to \mathrm{Sym}(\mathbb{R}^{2n})$ by

$$\mathcal{R}(\xi) = \sum_{k=1}^{m} \Phi_k \xi \xi^T \Phi_k \ .$$

Theorem 3.2.9. *Let $\mathcal{F}$ be a phase retrievable frame for $H = \mathbb{C}^n$. For every $0 \neq z \in H$, let $\zeta = \mathbf{j}(z)$ denote the realification of z. Then*

1. $a(z) = \tilde{a}(z) = \lambda_{2n-1}(\mathcal{R}(\zeta))/\left\|\zeta\right\|^2$;
2. $b(z) = \tilde{b}(z) = \lambda_1(\mathcal{R}(\zeta))/\left\|\zeta\right\|^2$;
3. (see [12]) $a(0) = a_0 = \min_{\|\zeta\|=1} \lambda_{2n-1}\left(\mathcal{R}(\zeta)\right)$;
4. $\tilde{a}(0) = \min_{\|x\|=1} \sum_{k=1}^{m} \left|\langle x, f_k\rangle\right|^4$;
5. $b(0) = \tilde{b}(0) = b_0$.

Proof. Only the first two parts are nontrivial. We prove them as follows.

Fix $z \in \mathbb{C}^n$. Take $x = z + d_1$ and $y = z + d_2$ with $\|d_1\| < r$ and $\|d_2\| < r$ for r small. Let $u = x+y = 2z+d_1+d_2$ and $v = x-y = d_1-d_2$. Let $\mu = 2\zeta+\delta_1+\delta_2 \in \mathbb{R}^{2n}$ and $\nu = \delta_1 - \delta_2 \in \mathbb{R}^{2n}$ be the realification of u and v, respectively. Define

$$\rho(x,y) = \frac{\left\|\beta(x) - \beta(y)\right\|^2}{d_1(x,y)^2} \ . \tag{3.27}$$

By the same computation as in [12], Section 4.1, we get

$$\begin{aligned} \rho(x,y) &= Q(\zeta; \delta_1, \delta_2) \\ &:= \frac{\langle \mathcal{R}(2\zeta + \delta_1 + \delta_2)(\delta_1 - \delta_2), \delta_1 - \delta_2 \rangle}{\left\|2\zeta + \delta_1 + \delta_2\right\|^2 \left\langle P^{\perp}_{J(2\zeta+\delta_1+\delta_2)}(\delta_1 - \delta_2), \delta_1 - \delta_2 \right\rangle} \ . \end{aligned}$$

Since

$$J(2\zeta + \delta_1 + \delta_2) \in \ker \mathcal{R}(2\zeta + \delta_1 + \delta_2)$$

, we have

$$
\begin{aligned}
& Q(\zeta;\delta_1,\delta_2) \\
= & \frac{\left\langle \mathcal{R}(2\zeta+\delta_1+\delta_2)P^{\perp}_{J(2\zeta+\delta_1+\delta_2)}(\delta_1-\delta_2), P^{\perp}_{J(2\zeta+\delta_1+\delta_2)}(\delta_1-\delta_2)\right\rangle}{\|2\zeta+\delta_1+\delta_2\|^2\left\langle P^{\perp}_{J(2\zeta+\delta_1+\delta_2)}(\delta_1-\delta_2),\delta_1-\delta_2\right\rangle} .
\end{aligned}
$$

Now let $\delta=\delta_1+\delta_2$ and $\nu=\delta_1-\delta_2$. Note the set inclusion relation

$$
\begin{aligned}
& \left\{\delta_1,\delta_2\in\mathbb{R}^{2n}: \quad \|\delta\|<\frac{r}{2},\ \ \|\nu\|<\frac{r}{2},\ \nu\perp J(2\zeta+\delta)\right\} \\
\subset & \left\{\delta_1,\delta_2\in\mathbb{R}^{2n}: \quad \|\delta_1\|<r,\ \ \|\delta_2\|<r,\ \nu\perp J(2\zeta+\delta)\right\} \\
\subset & \left\{\delta_1,\delta_2\in\mathbb{R}^{2n}: \quad \|\delta\|<2r,\ \ \|\nu\|<2r,\ \nu\perp J(2\zeta+\delta)\right\} .
\end{aligned}
$$

Thus we have

$$
\inf_{\substack{\|\delta\|<2r\\ \|\nu\|<2r\\ \nu\perp J(2\zeta+\delta)}} Q(\zeta;\delta_1,\delta_2) \quad \le \quad \inf_{\substack{\|\delta_1\|<r\\ \|\delta_2\|<r\\ \nu\perp J(2\zeta+\delta)}} Q(\zeta;\delta_1,\delta_2) \quad \le \quad \inf_{\substack{\|\delta\|<r/2\\ \|\nu\|<r/2\\ \nu\perp J(2\zeta+\delta)}} Q(\zeta;\delta_1,\delta_2) .
$$

That is,

$$
\inf_{\|\delta\|<2r}\frac{\lambda_{2n-1}(\mathcal{R}(2\zeta+\delta))}{\|2\zeta+\delta\|^2} \quad \le \quad \inf_{\substack{\|\delta_1\|<r\\ \|\delta_2\|<r\\ \nu\perp J(2\zeta+\delta)}} Q(\zeta;\delta_1,\delta_2) \quad \le \inf_{\|\delta\|<r/2}\frac{\lambda_{2n-1}(\mathcal{R}(2\zeta+\delta))}{\|2\zeta+\delta\|^2} .
$$

Take $r\to 0$, by the continuity of eigenvalues with respect to the matrix entries, we have

$$
\lambda_{2n-1}(\mathcal{R}(\zeta))/\|\zeta\|^2 \le a(z) \le \lambda_{2n-1}(\mathcal{R}(\zeta))/\|\zeta\|^2 .
$$

That is,

$$
a(z)=\lambda_{2n-1}(\mathcal{R}(\zeta))/\|\zeta\|^2 .
$$

Now consider

$$
\rho(x,z)=\frac{\|\beta(x)-\beta(z)\|^2}{d_1(x,z)^2} .
$$

For simplicity write $\delta = \delta_1$. We can compute that

$$\begin{aligned}\rho(x,z) &= Q(\zeta;\delta) \\ &= \frac{\langle \mathcal{R}(2\zeta+\delta)\delta, \delta\rangle}{\|2\zeta+\delta\|^2 \left\langle P^{\perp}_{J(2\zeta+\delta)}\delta, \delta\right\rangle} \\ &= \frac{\left\langle \mathcal{R}(2\zeta+\delta)P^{\perp}_{J(2\zeta+\delta)}\delta, P^{\perp}_{J(2\zeta+\delta)}\delta\right\rangle}{\|2\zeta+\delta\|^2 \left\langle P^{\perp}_{J(2\zeta+\delta)}\delta, \delta\right\rangle} \ .\end{aligned}$$

Note that

$$\inf_{\substack{\|\delta\|<r \\ \delta\perp J(2\zeta+\delta)}} Q(\zeta;\delta) \quad \geq \quad \inf_{\|\sigma\|<r} \inf_{\substack{\|\delta\|<r \\ \delta\perp J(2\zeta+\delta)}} Q(\zeta;\delta) \quad = \quad \inf_{\|\sigma\|<r} \lambda_{2n-1}(\mathcal{R}(2\zeta+\delta)) \ .$$

Take $r \to 0$ we have that

$$\tilde{a}(z) \geq \lambda_{2n-1}(\mathcal{R}(2\zeta))/\left\|2\zeta\right\|^2 = \lambda_{2n-1}(\mathcal{R}(\zeta))/\left\|\zeta\right\|^2 \ .$$

On the other hand, take $\tilde{e}_{2n-1}$ to be a unit-norm eigenvector correspondent to $\lambda_{2n-1}(\mathcal{R}(2\zeta))$. Then by the continuity of eigenvalues with respect to the matrix entries, for any $\epsilon > 0$, there exists $t > 0$ so that $\delta = t\tilde{e}_{2n-1}$ satisfy

$$\frac{\langle \mathcal{R}(2\zeta+\delta)\delta, \delta\rangle}{\left\langle P^{\perp}_{J(2\zeta+\delta)}\delta, \delta\right\rangle} \leq \lambda_{2n-1}(\mathcal{R}(2\zeta)) + \epsilon$$

and from there we have

$$\tilde{a}(z) \leq \lambda_{2n-1}(\mathcal{R}(2\zeta))/\left\|2\zeta\right\|^2 = \lambda_{2n-1}(\mathcal{R}(\zeta))/\left\|\zeta\right\|^2 \ .$$

Therefore,

$$\tilde{a}(z) = \lambda_{2n-1}(\mathcal{R}(\zeta))/\left\|\zeta\right\|^2 \ .$$

In a similar way (replacing infimum by supremum) we also get $b(z)$ and $\tilde{b}(z)$ as stated in the theorem. □

3.3 Global stable reconstruction

The results above show that if the frame $\mathcal{F}$ is phase retrievable, then the nonlinear map α (resp., β) is bi-Lipschitz between the metric spaces $(\hat{H}, D_p)$ (resp., $(\hat{H}, d_p)$) and $(\mathbb{R}^m, \|\cdot\|_q)$. Recall that the Lipschitz constants between $(\hat{\mathcal{H}}, D_2)$ (resp., $(\hat{\mathcal{H}}, d_1)$) and $(\mathbb{R}^m, \|\cdot\| = \|\cdot\|_2)$ are given by $\sqrt{A_0}$ (resp., $\sqrt{a_0}$) and $\sqrt{B_0}$ (resp., $\sqrt{b_0}$):

$$\sqrt{A_0} D_2(x, y) \leq \|\alpha(x) - \alpha(y)\| \leq \sqrt{B_0} D_2(x, y) \ , \tag{3.28}$$

$$\sqrt{a_0} d_1(x, y) \leq \|\beta(x) - \beta(y)\| \leq \sqrt{b_0} d_1(x, y) \ . \tag{3.29}$$

Clearly the inverse map defined on the range of α (resp., β) from metric space $(\alpha(\hat{\mathcal{H}}), \|\cdot\|)$ (resp., $(\beta(\hat{\mathcal{H}}), \|\cdot\|)$) to $(\hat{\mathcal{H}}, D_2)$ (resp., $(\hat{\mathcal{H}}, d_1)$):

$$\tilde{\omega} : \alpha(\hat{\mathcal{H}}) \subset \mathbb{R}^m \to \hat{\mathcal{H}} \ \ , \ \ \tilde{\omega}(c) = x \ \text{ if } \alpha(x) = c \ ; \tag{3.30}$$

$$\tilde{\psi} : \beta(\hat{\mathcal{H}}) \subset \mathbb{R}^m \to \hat{\mathcal{H}} \ \ , \ \ \tilde{\psi}(c) = x \ \text{ if } \beta(x) = c \ . \tag{3.31}$$

is Lipschitz with Lipschitz constant $1/\sqrt{A_0}$ (resp., $1/\sqrt{a_0}$). We prove that both $\tilde{\omega}$ and $\tilde{\psi}$ can be extended to the entire $\mathbb{R}^m$ as a Lipschitz map, and its Lipschitz constant is increased by a small factor.

The proof should be easy to establish if we can establish that $(\mathbb{R}^m, \mathcal{H})$ has the Property (K) as defined in Section 2.3. However, the following examples show that it is not the case.

Example 3.3.1. Property (K) does not hold for $\hat{\mathcal{H}}$ with norm D_p. Specifically, $(\mathbb{R}^m, \mathbb{R}^n/\sim)$ does not have Property (K). We give a counterexample for $m = n = 2, p = 2$: Let $\tilde{y}_1 = (3, 1)$, $\tilde{y}_2 = (-1, 1)$, $\tilde{y}_3 = (0, 1)$ be the representatives of three

points y_1, y_2, y_3 in $\mathbb{R}^2/\sim$. Then $D_2(y_1, y_2) = 2\sqrt{2}$, $D_2(y_2, y_3) = 1$ and $D_2(y_1, y_3) = 3$. Consider $x_1 = (0, 0)$, $x_2 = (0, -2\sqrt{2})$, $x_3 = (-1, -2\sqrt{2})$ in $\mathbb{R}^2$ with the Euclidean distance, then we have $\|x_1 - x_2\| = 2\sqrt{2}$, $\|x_2 - x_3\| = 1$ and $\|x_1 - x_3\| = 3$. For $r_1 = \sqrt{6}$, $r_2 = 2 - \sqrt{2}$, $r_3 = \sqrt{6} - \sqrt{3}$, we see that $(1 - \sqrt{2}, 1 + \sqrt{2}) \in \bigcap_{i=1}^{3} B(x_i, r_i)$ but $\bigcap_{i=1}^{3} B(y_i, r_i) = \emptyset$. To see $\bigcap_{i=1}^{3} B(y_i, r_i) = \emptyset$, it suffices to look at the upper half plane in $\mathbb{R}^2$. If we look at the upper half plane H, then $B(y_1, r_1)$ becomes the union of two parts, namely $B(\tilde{y}_1, r_1) \cap H$ and $B(-\tilde{y}_1, r_1) \cap H$, and $B(y_i, r_i)$ becomes $B(\tilde{y}_i, r_i)$ for $i = 2, 3$. But $(B(\tilde{y}_1, r_1) \cap H) \cap B(\tilde{y}_2, r_2) = \emptyset$ and $(B(-\tilde{y}_1, r_1) \cap H) \cap B(\tilde{y}_3, r_3) = \emptyset$. So we obtain that $\bigcap_{i=1}^{3} B(y_i, r_i) = \emptyset$.

Example 3.3.2. Property (K) does not hold for $\hat{\mathcal{H}}$ with norm d_p. Specifically, $(\mathbb{R}^m, \mathbb{C}^n/\sim)$ does not have Property (K). Let m be any positive integer and $n = 2$, $p = 2$. We want to show that $(X, Y) = (\mathbb{R}^m, \mathbb{C}^n/\sim)$ does not have Property (K). Let $\tilde{y}_1 = (1, 0)$ and $\tilde{y}_2 = (0, \sqrt{3})$ be representatives of y_1, $y_2 \in Y$, respectively. Then $d_1(y_1, y_2) = 4$. Pick any two points x_1, x_2 in X with $\|x_1 - x_2\| = 4$. Then $B(x_1, 2)$ and $B(x_2, 2)$ intersect at $x_3 = (x_1 + x_2)/2 \in X$. It suffices to show that the closed balls $B(y_1, 2)$ and $B(y_2, 2)$ have no intersection in H. Assume on the contrary that the two balls intersect at y_3, then pick a representative of y_3, say $\tilde{y}_3 = (a, b)$ where a, $b \in \mathbb{C}$. It can be computed that

$$d_1(y_1, y_3) = |a|^4 + |b|^4 - 2|a|^2 + 2|b|^2 + 2|a|^2|b|^2 + 1 \ , \tag{3.32}$$

and

$$d_1(y_2, y_3) = |a|^4 + |b|^4 + 6|a|^2 - 6|b|^2 + 2|a|^2|b|^2 + 9 \ . \tag{3.33}$$

Set $d_1(y_1, y_3) = d_1(y_2, y_3) = 2$. Take the difference of the right hand side of (3.32)

and (3.33), we have $|b|^2 - |a|^2 = 1$ and thus $|b|^2 \geq 1$. However, the right hand side of (3.32) can be rewritten as $(|a|^2 + |b|^2 - 1)^2 + 4|b|^2$, so $d_1(y_1, y_3) = 2$ would imply that $|b|^2 \leq 1/2$. This is a contradiction.

Nevertheless, we can still use the Kirszbraun theorem. We need to circumvent by first constructing a Lipschitz map from the symmetric matrices to the rank-one's. This is stated as the following lemma.

Lemma 3.3.3. *Consider the spectral decomposition of any self-adjoint operator A in $\mathrm{Sym}(\mathcal{H})$, say $A = \sum_{k=1}^{d} \lambda_{m(k)} P_k$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the n eigenvalues including multiplicities, and P_1,...,P_d are the orthogonal projections associated to the d distinct eigenvalues. Additionally, $m(1) = 1$ and $m(k+1) = m(k) + r(k)$, where $r(k) = rank(P_k)$ is the multiplicity of eigenvalue $\lambda_{m(k)}$. Then the map*

$$\pi : \mathrm{Sym}(\mathcal{H}) \to S^{1,0}(\mathcal{H}) \quad , \quad \pi(A) = (\lambda_1 - \lambda_2) P_1 \tag{3.34}$$

satisfies the following two properties:

1. *for $1 \leq p \leq \infty$, π is Lipschitz continuous from $(\mathrm{Sym}(\mathcal{H}), \|\cdot\|_p)$ to $(S^{1,0}(\mathcal{H}), \|\cdot\|_p)$ with Lipschitz constant $\mathrm{Lip}(\pi) \leq 3 + 2^{1+\frac{1}{p}}$;*

2. *$\pi(A) = A$ for all $A \in S^{1,0}(\mathcal{H})$.*

Proof. Part 2 follows directly from the expression of π. We prove Part 1 as follows.

Let A, $B \in \mathrm{Sym}(\mathcal{H})$ where

$$A = \sum_{k=1}^{d} \lambda_{m(k)} P_k$$

and

$$B = \sum_{k'=1}^{d'} \mu_{m(k')} Q_{k'} \ .$$

We now show that

$$\|\pi(A) - \pi(B)\|_p \leq (3 + 2^{1+\frac{1}{p}}) \|A - B\|_p \ . \tag{3.35}$$

Assume $\lambda_1 - \lambda_2 \leq \mu_1 - \mu_2$. Otherwise switch the notations for A and B. If $\mu_1 - \mu_2 = 0$ then $\pi(A) = \pi(B) = 0$ and the inequality (3.35) is satisfied. Assume now $\mu_1 - \mu_2 > 0$. Thus Q_1 is of rank 1 and $\|Q_1\|_p = 1$ for all p. First we consider the case $\lambda_1 - \lambda_2 > 0$. In this case P_1 is of rank 1, and we have

$$\begin{aligned} \pi(A) - \pi(B) &= (\lambda_1 - \lambda_2)P_1 - (\mu_1 - \mu_2)Q_1 \\ &= (\lambda_1 - \lambda_2)(P_1 - Q_1) + (\lambda_1 - \mu_1 - (\lambda_2 - \mu_2))Q_1 \ . \end{aligned} \tag{3.36}$$

Here $\|P_1\|_\infty = \|Q_1\|_\infty = 1$. Therefore we have $\|P_1 - Q_1\|_\infty \leq 1$ since P_1, $Q_1 \geq 0$. From that we have $\|P_1 - Q_1\|_p \leq 2^{\frac{1}{p}}$. Also, by Weyl's inequality we have $|\lambda_i - \mu_i| \leq \|A - B\|_\infty$ for each i. Apply this to $i = 1$, 2 we get $|\lambda_1 - \mu_1 - (\lambda_2 - \mu_2)| \leq |\lambda_1 - \mu_1| + |\lambda_2 - \mu_2| \leq 2\|A - B\|_\infty$. Thus $|\lambda_1 - \mu_1| + |\lambda_2 - \mu_2| \leq 2\|A - B\|_\infty \leq 2\|A - B\|_p$. Let $g := \lambda_1 - \lambda_2$, $\delta := \|A - B\|_p$, then apply the above inequality to (3.36) we get

$$\|\pi(A) - \pi(B)\|_p \leq g\|P_1 - Q_1\|_p + 2\delta \leq 2^{\frac{1}{p}} g + 2\delta \ . \tag{3.37}$$

If $0 < g \leq (2 + 2^{-\frac{1}{p}})\delta$, then $\|\pi(A) - \pi(B)\|_p \leq (2^{1+\frac{1}{p}} + 3)\delta$ and we are done.

Now we consider the case where $g > (2 + 2^{-\frac{1}{p}})\delta$. Note that in this case we have $\delta < g/2$. Thus we have $|\lambda_1 - \mu_1| < g/2$ and $|\lambda_2 - \mu_2| < g/2$. That means $\mu_1 > (\lambda_1 + \lambda_2)/2$ and $\mu_2 < (\lambda_1 + \lambda_2)/2$. Therefore, we can use holomorphic functional

calculus as introduced in Section 2.2 to put

$$P_1 = -\frac{1}{2\pi i} \oint_\gamma R_A dz$$

and

$$Q_1 = -\frac{1}{2\pi i} \oint_\gamma R_B dz$$

where $R_A = (A - zI)^{-1}$, $R_B = (B - zI)^{-1}$, and $\gamma = \gamma(t)$ is the contour given in Figure 3.2 (note that γ encloses μ_1 but not μ_2) and used also by [108]. Therefore

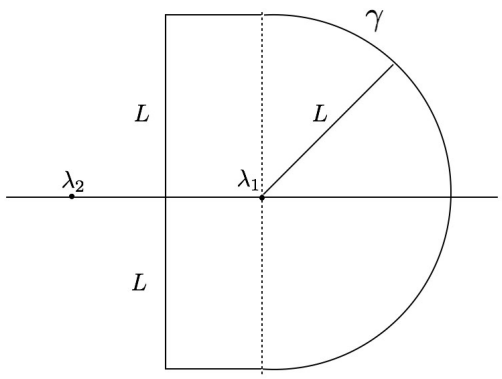

Figure 3.2: Contour for the integrals

we have

$$\|P_1 - Q_1\|_p \leq \frac{1}{2\pi} \int_I \|(R_A - R_B)(\gamma(t))\|_p \, |\gamma'(t)| dt \; . \tag{3.38}$$

Now we have

$$\begin{aligned} (R_A - R_B)(z) &= R_A(z) - (I + R_A(z)(B - A))^{-1} R_A(z) \\ &= \sum_{n \geq 1} (-1)^n (R_A(z)(B - A))^n R_A(z) \; , \end{aligned} \tag{3.39}$$

since for large L we have

$$\begin{aligned} \|R_A(z)(B-A)\|_\infty &\leq \|R_A(z)\|_\infty \|B-A\|_p \\ &\leq \frac{\delta}{\mathrm{dist}(z,\sigma(A))} \\ &\leq \frac{2\delta}{g} < \frac{2}{2+2^{-\frac{1}{p}}} < 1 \ , \end{aligned} \tag{3.40}$$

where $\sigma(A)$ denotes the spectrum of A. Therefore we have

$$\begin{aligned} \|(R_A - R_B)(\gamma(t))\|_p &\leq \sum_{n\geq 1} \|R_A(\gamma(t))\|_\infty^{n+1} \|A-B\|_p^n \\ &= \frac{\|R_A(\gamma(t))\|_\infty^2 \|A-B\|_p}{1-\|R_A(\gamma(t))\|_\infty \|A-B\|_p} \\ &< \frac{\|A-B\|_p}{\mathrm{dist}^2(\gamma(t),\sigma(A))} \cdot (2^{1+\frac{1}{p}}+1) \ , \end{aligned} \tag{3.41}$$

since $\mathrm{dist}(\gamma(t),\sigma(A)) \geq g/2$ for each t for large L. Here we used the fact that if we order the singular values of any matrix X such that $\sigma_1(X) \geq \sigma_2(X) \geq \cdots$, then for any i we have $\sigma_i(XY) \leq \sigma_1(X)\sigma_i(Y)$, and thus for two operators X, $Y \in \mathrm{Sym}(\mathcal{H})$, we have $\|XY\|_p \leq \|X\|_\infty \|Y\|_p$. Hence by (3.38) and (3.41) we have

$$\|P_1 - Q_1\|_p \leq (2^{\frac{1}{p}}+2^{-1}) \frac{\|A-B\|_p}{\pi} \int_I \frac{1}{\mathrm{dist}^2(\gamma(t),\sigma(A))} |\gamma'(t)| dt \ . \tag{3.42}$$

By evaluating the integral and letting L approach infinity for the contour, we have as in [108]

$$\begin{aligned} \int_I \frac{1}{\mathrm{dist}^2(\gamma(t),\sigma(A))} |\gamma'(t)| dt &= 2\int_0^\infty \frac{1}{t^2+(\frac{g}{2})^2} dt \\ &= \left[\frac{4}{g} \arctan\left(\frac{2t}{g}\right)\right]_0^\infty \\ &= \frac{2\pi}{g} \ . \end{aligned} \tag{3.43}$$

Hence

$$\|P_1 - Q_1\|_p \leq (2^{\frac{1}{p}}+2^{-1}) \frac{\|A-B\|_p}{\pi} \cdot \frac{2\pi}{g} = (2^{1+\frac{1}{p}}+1)\frac{\delta}{g} \ . \tag{3.44}$$

Thus by the first inequality in (3.37) and (3.44) we have

$$\|\pi(A) - \pi(B)\|_p \leq (3 + 2^{1+\frac{1}{p}})\delta \ .$$

Now we are left with the case $\lambda_1 - \lambda_2 = 0 < \mu_1 - \mu_2$. Note that in this case we have that $\pi(A) - \pi(B) = -(\mu_1 - \mu_2)Q_1 = ((\lambda_1 - \mu_1) - (\lambda_2 - \mu_2))Q_1$, and therefore

$$\|\pi(A) - \pi(B)\|_p \leq 2\,\|A - B\|_p < (3 + 2^{1+\frac{1}{p}})\,\|A - B\|_p \ .$$

We have proved that $\|\pi(A) - \pi(B)\|_p \leq (3 + 2^{1+\frac{1}{p}})\,\|A - B\|_p$. That is to say,

$$\pi : (\mathrm{Sym}(H), \|\cdot\|_p) \to (S^{1,0}(H), \|\cdot\|_p)$$

is Lipschitz continuous with

$$\mathrm{Lip}(\pi) \leq 3 + 2^{1+\frac{1}{p}} \ .$$

□

Remark 3.3.4. Numerical experiments suggest that the Lipschitz constant of π is smaller than 5 for $p = \infty$. On the other hand, it cannot be smaller than 2 as the following example shows.

Example 3.3.5. If $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $B = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$, then $\pi(A) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ and $\pi(B) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$. Here we have $\|\pi(A) - \pi(B)\|_\infty = 2$ and $\|A - B\|_\infty = 1$. Thus for this example we have

$$\|\pi(A) - \pi(B)\|_\infty = 2\,\|A - B\|_\infty \ .$$

Now we are ready to establish the extension result. The precise statement is given in the following theorem:

Theorem 3.3.6. *Let $\mathcal{F} = \{f_1, \ldots, f_m\}$ be a phase retrievable frame for the n dimensional Hilbert space $\mathcal{H}$, and let $\alpha, \beta : \hat{\mathcal{H}} \to \mathbb{R}^m$ denote the injective nonlinear analysis maps as defined in (3.1) and (3.2). Let A_0 and a_0 denote the positive constants as in (3.28) and (3.29). Then*

1. *there exists a Lipschitz continuous function $\omega : \mathbb{R}^m \to \hat{\mathcal{H}}$ so that $\omega(\alpha(x)) = x$ for all $x \in \hat{\mathcal{H}}$. For any $1 \leq p, q \leq \infty$, ω has an upper Lipschitz constant $\mathrm{Lip}(\omega)_{p,q}$ between $(\mathbb{R}^m, \|\cdot\|_p)$ and $(\hat{\mathcal{H}}, D_q)$ bounded by:*

$$\mathrm{Lip}(\omega)_{p,q} \leq \begin{cases} \frac{3\sqrt{2}+4}{\sqrt{A_0}} \cdot 2^{\frac{1}{q}-\frac{1}{2}} \cdot \max(1, m^{\frac{1}{2}-\frac{1}{p}}) & \text{for } q \leq 2; \\ \frac{3\sqrt{2}+2^{\frac{3}{2}+\frac{1}{q}}}{\sqrt{A_0}} \cdot n^{\frac{1}{2}-\frac{1}{q}} \cdot \max(1, m^{\frac{1}{2}-\frac{1}{p}}) & \text{for } q > 2. \end{cases} \tag{3.45}$$

   Explicitly this means: for $q \leq 2$ and for all $c, d \in \mathbb{R}^m$:

$$D_q(\omega(c), \omega(d)) \leq \frac{3\sqrt{2}+4}{\sqrt{A_0}} \cdot 2^{\frac{1}{q}-\frac{1}{2}} \cdot \max(1, m^{\frac{1}{2}-\frac{1}{p}}) \|c-d\|_p \ , \tag{3.46}$$

   whereas for $q > 2$ and for all $c, d \in \mathbb{R}^m$:

$$D_q(\omega(c), \omega(d)) \leq \frac{3\sqrt{2}+2^{\frac{3}{2}+\frac{1}{q}}}{\sqrt{A_0}} \cdot n^{\frac{1}{2}-\frac{1}{q}} \cdot \max(1, m^{\frac{1}{2}-\frac{1}{p}}) \|c-d\|_p \ . \tag{3.47}$$

   In particular, for $p = 2$ and $q = 2$ its Lipschitz constant $\mathrm{Lip}(\omega)_{2,2}$ is bounded by $\frac{4+3\sqrt{2}}{\sqrt{A_0}}$:

$$D_2(\omega(c), \omega(d)) \leq \frac{4+3\sqrt{2}}{\sqrt{A_0}} \|c-d\| \ . \tag{3.48}$$

2. *there exists a Lipschitz continuous function $\psi : \mathbb{R}^m \to \hat{\mathcal{H}}$ so that $\psi(\beta(x)) = x$ for all $x \in \hat{\mathcal{H}}$. For any $1 \leq p, q \leq \infty$, ψ has an upper Lipschitz constant*

$\mathrm{Lip}(\psi)_{p,q}$ *between* $(\mathbb{R}^m, \|\cdot\|_p)$ *and* $(\hat{\mathcal{H}}, d_q)$ *bounded by:*

$$\mathrm{Lip}(\psi)_{p,q} \leq \begin{cases} \frac{3+2\sqrt{2}}{\sqrt{a_0}} \cdot 2^{\frac{1}{q}-\frac{1}{2}} \cdot \max(1, m^{\frac{1}{2}-\frac{1}{p}}) & \text{for } q \leq 2; \\ \frac{3+2^{1+\frac{1}{q}}}{\sqrt{a_0}} \max(1, m^{\frac{1}{2}-\frac{1}{p}}) & \text{for } q > 2. \end{cases} \tag{3.49}$$

Explicitly this means: for $q \leq 2$ *and for all* $c, d \in \mathbb{R}^m$*:*

$$d_q(\psi(c), \psi(d)) \leq \frac{3+2\sqrt{2}}{\sqrt{a_0}} \cdot 2^{\frac{1}{q}-\frac{1}{2}} \cdot \max(1, m^{\frac{1}{2}-\frac{1}{p}}) \|c-d\|_p \ , \tag{3.50}$$

whereas for $q > 2$ *and for all* $c, d \in \mathbb{R}^m$*:*

$$d_q(\psi(c), \psi(d)) \leq \frac{3+2^{1+\frac{1}{q}}}{\sqrt{a_0}} \max(1, m^{\frac{1}{2}-\frac{1}{p}}) \|c-d\|_p \ . \tag{3.51}$$

In particular, for $p = 2$ *and* $q = 1$ *its Lipschitz constant* $\mathrm{Lip}(\psi)_{2,1}$ *is bounded by* $\frac{4+3\sqrt{2}}{\sqrt{a_0}}$*:*

$$d_1(\psi(c), \psi(d)) \leq \frac{4+3\sqrt{2}}{\sqrt{a_0}} \|c-d\| \ . \tag{3.52}$$

Proof. The proof for α and β are the same in essence. For simplicity we do it for β first.

We need to construct a map $\psi : (\mathbb{R}^m, \|\cdot\|_p) \to (\hat{\mathcal{H}}, d_q)$ such that $\psi(\beta(x)) = x$ for all $x \in \hat{\mathcal{H}}$, and ψ is Lipschitz continuous. We prove the Lipschitz bound (3.49), which implies (3.52) for $p = 2$ and $q = 1$.

The following construction of ψ is summarized in Figure 3.3.

Set $M = \beta(\hat{\mathcal{H}}) \subset \mathbb{R}^m$. Due to the bi-Lipschitz property of β, there is a map $\tilde{\psi}_1 : M \to \hat{\mathcal{H}}$ that is Lipschitz continuous and satisfies $\tilde{\psi}_1(\beta(x)) = x$ for all $x \in \hat{\mathcal{H}}$. Additionally, the Lipschitz bound between $(M, \|\cdot\|_2)$ (that is, M with Euclidean distance) and $(\hat{\mathcal{H}}, d_1)$ is given by $1/\sqrt{a_0}$.

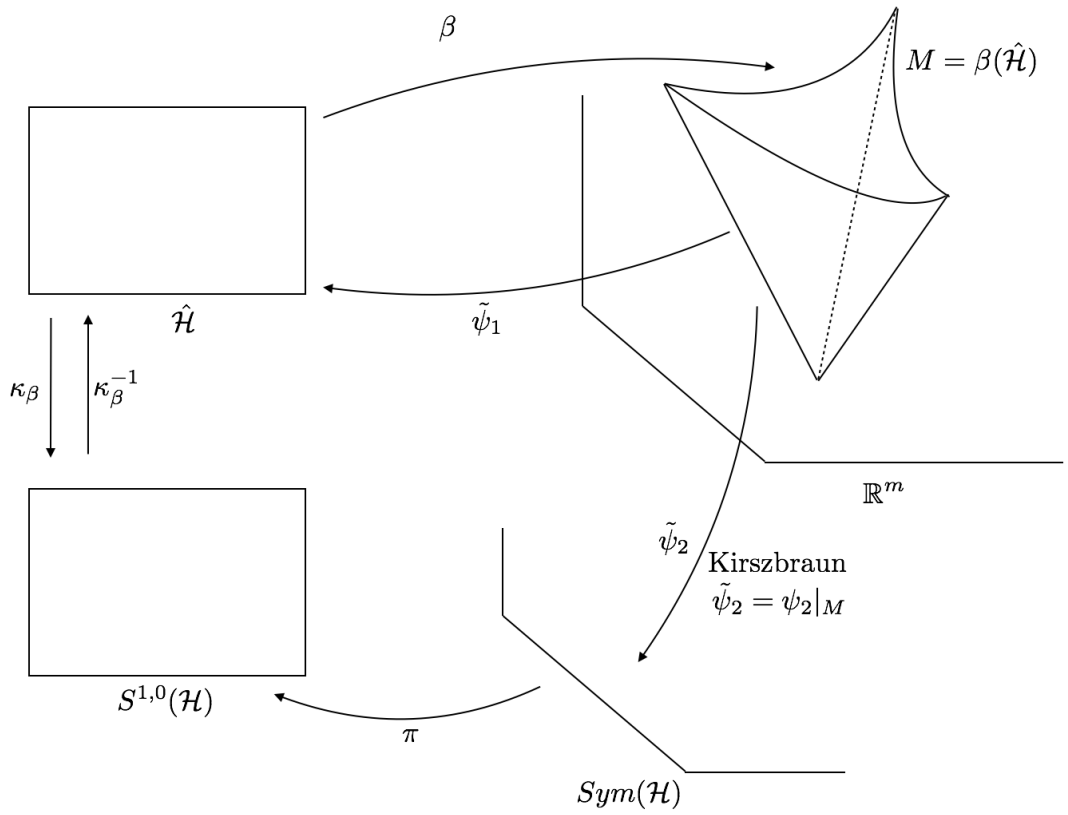

Figure 3.3: Illustration of the extended Lipschitz map. We omitted the change of norms. First we construct a map from the image of the nonlinear measurement to $\mathrm{Sym}(\mathcal{H})$, then we use the Kirszbraun Theorem to extend it to $\mathbb{R}^m$, we use π defined in Lemma 3.3.3 to map it to $S^{1,0}(\mathcal{H})$ and isometrically transform it back to $\hat{\mathcal{H}}$.

First we change the metric on $\hat{\mathcal{H}}$ from d_1 to d_2 and embed isometrically $\hat{\mathcal{H}}$ into $\mathrm{Sym}(\mathcal{H})$ with Frobenius norm (i.e. the Euclidean metric):

$$(M, \|\cdot\|_2) \xrightarrow{\tilde{\psi}_1} (\hat{\mathcal{H}}, d_1) \xrightarrow{i_{1,2}} (\hat{\mathcal{H}}, d_2) \xrightarrow{\kappa_\beta} (\mathrm{Sym}(\mathcal{H}), \|\cdot\|_{Fr}) \ , \tag{3.53}$$

where $i_{1,2}(x) = x$ is the identity of $\hat{\mathcal{H}}$ and κ_β is the isometry (3.10) . We obtain a map $\tilde{\psi}_2 : (M, \|\cdot\|_2) \to (\mathrm{Sym}(H), \|\cdot\|_{Fr})$ of Lipschitz constant

$$\mathrm{Lip}(\tilde{\psi}_2) \leq \mathrm{Lip}(\tilde{\psi}_1)\mathrm{Lip}(i_{1,2})\mathrm{Lip}(\kappa_\beta) = \frac{1}{\sqrt{a_0}} \ ,$$

where we used $\mathrm{Lip}(i_{1,2}) = L^d_{1,2,n} = 1$ by (3.8).

Kirszbraun Theorem extends isometrically $\tilde{\psi}_2$ from M to the entire $\mathbb{R}^m$ with Euclidean metric $\|\cdot\|$. Thus we obtain a Lipschitz map $\psi_2 : (\mathbb{R}^m, \|\cdot\|) \to (\mathrm{Sym}(\mathcal{H}), \|\cdot\|_{Fr})$ of Lipschitz constant $\mathrm{Lip}(\psi_2) = \mathrm{Lip}(\tilde{\psi}_2) \leq \frac{1}{\sqrt{a_0}}$ such that $\psi_2(\beta(x)) = xx^*$ for all $x \in \hat{\mathcal{H}}$.

The third step is to piece together ψ_2 with norm changing identities. For $q \leq 2$ we consider the following maps:

$$\begin{aligned}(\mathbb{R}^m, \|\cdot\|_p) \xrightarrow{j_{p,2}} (\mathbb{R}^m, \|\cdot\|_2) \xrightarrow{\psi_2} (\mathrm{Sym}(\mathcal{H}), \|\cdot\|_{Fr}) \\ \xrightarrow{\pi} (S^{1,0}(\mathcal{H}), \|\cdot\|_{Fr}) \xrightarrow{\kappa_\beta^{-1}} (\hat{\mathcal{H}}, d_2) \xrightarrow{i_{2,q}} (\hat{\mathcal{H}}, d_q) \ ,\end{aligned} \tag{3.54}$$

where $j_{p,2}$ and $i_{2,q}$ are identity maps on the respective spaces that change the metric and π is the map defined in Eq. (3.34). The map ψ claimed by Theorem 3.3.6 is obtained by composing:

$$\psi : (\mathbb{R}^m, \|\cdot\|_p) \to (\hat{\mathcal{H}}, d_q) \quad , \quad \psi = i_{2,q} \cdot \kappa_\beta^{-1} \cdot \pi \cdot \psi_2 \cdot j_{p,2} \ .$$

Its Lipschitz constant is bounded by

$$\begin{aligned}\mathrm{Lip}(\psi)_{p,q} &\leq \mathrm{Lip}(j_{p,2})\mathrm{Lip}(\psi_2)\mathrm{Lip}(\pi)\mathrm{Lip}(\kappa_\beta^{-1})\mathrm{Lip}(i_{2,q}) \\ &\leq \max(1, m^{\frac{1}{2}-\frac{1}{p}}) \frac{1}{\sqrt{a_0}} \cdot (3 + 2\sqrt{2}) \cdot 1 \cdot 2^{\frac{1}{q}-\frac{1}{2}} \ .\end{aligned}$$

Hence we obtained (3.50). The other equation (3.52) follows for $p = 2$ and $q = 1$.

For $q > 2$ we use:

$$\begin{aligned}(\mathbb{R}^m, \|\cdot\|_p) \xrightarrow{j_{p,2}} (\mathbb{R}^m, \|\cdot\|_2) \xrightarrow{\psi_2} (\mathrm{Sym}(\mathcal{H}), \|\cdot\|_{Fr}) \\ \xrightarrow{I_{2,q}} (\mathrm{Sym}(\mathcal{H}), \|\cdot\|_q) \xrightarrow{\pi} (S^{1,0}(\mathcal{H}), \|\cdot\|_q) \xrightarrow{\kappa_\beta^{-1}} (\hat{\mathcal{H}}, d_q) \ ,\end{aligned} \tag{3.55}$$

where $j_{p,2}$ and $I_{2,q}$ are identity maps on the respective spaces that change the metric. The map ψ claimed by Theorem 3.3.6 is obtained by composing:

$$\psi : (\mathbb{R}^m, \|\cdot\|_p) \to (\hat{\mathcal{H}}, d_q) \quad , \quad \psi = \kappa_\beta^{-1} \cdot \pi \cdot I_{2,q} \cdot \psi_2 \cdot j_{p,2} \ .$$

Its Lipschitz constant is bounded by

$$\begin{aligned} \mathrm{Lip}(\psi)_{p,q} &\leq \mathrm{Lip}(j_{p,2})\mathrm{Lip}(\psi_2)\mathrm{Lip}(I_{2,q})\mathrm{Lip}(\pi)\mathrm{Lip}(\kappa_\beta^{-1}) \\ &\leq \max(1, m^{\frac{1}{2}-\frac{1}{p}}) \frac{1}{\sqrt{a_0}} \cdot 1 \cdot (3 + 2^{1+\frac{1}{q}}) \cdot 1 \ . \end{aligned}$$

Hence we obtained (3.51).

Replace β by α, ψ by ω, and κ_β by κ_α in the proof above, using the Lipschitz constants for κ_α in Proposition 3.1.1, we obtain (3.46) and (3.47). □

The theorem above guarantees a Lipschitz extension for both measurements α and β. In fact this works for more general measurements because we do not need to assume a rank-1 measurement for our theory to work. For instance, in quantum tomography as introduced in Section 1.1.2, suppose we have a pure state $\rho = |\psi\rangle\langle\psi|$, and the measurements are given as $\mathbf{A}(\rho) = (\mathrm{tr}\rho A_k)_{k=1}^m$. Naturally, we adapt the distance function d_p induced by the matrix norms for ρ. Then suppose the measurement is bi-Lipschitz, we can use the same way as in Theorem 3.3.6 to extend the inverse map to $\mathbb{R}^m$.

Chapter 4: Lipschitz properties of Convolutional neural networks

4.1 Motivations for studying the stability of deep networks

Although AlexNet and GoogleNet introduced in Section 1.2 achieve state-of-the-art classification accuracy, a small variation of an input image may easily cause classification errors. In [96], the authors found that for AlexNet, a randomly selected image can be slightly distorted and be classified wrong. We take their illustrating example as shown in Figure 4.1. In all those examples, a tiny distortion (that cannot be told by human eyes) on the input image causes the classification to be wrong.

Figure 4.1: The adversarial examples given in [96]. In each group (row) of pictures, the picture on the left is correctly labeled by AlexNet, the picture on the right is labeled wrong as "ostrich", and the picture in the middle show their difference.

The authors of [96] have studied the upper frame bounds for each layer of the AlexNet. The following table shows the frame bounds computed numerically.

Layer	Size	Stride	Upper bound
Conv. 1	$3 \times 11 \times 11 \times 96$	4	2.75
Conv. 2	$96 \times 5 \times 5 \times 256$	1	10
Conv. 3	$256 \times 3 \times 3 \times 384$	1	7
Conv. 4	$384 \times 3 \times 3 \times 384$	1	7.5
Conv. 5	$384 \times 3 \times 3 \times 256$	1	11
FC. 1	9216×4096	N/A	3.12
FC. 2	4096×4096	N/A	4
FC. 3	4096×1000	N/A	4

Figure 4.2: The frame bound for each layer of AlexNet. Taken from [96].

The scattering network has the property that it is approximately translation invariant and stable to deformation. However, that property depends on a careful selection of wavelets. Moreover, the feature generating process is fixed for different problems while the feature selection process is trained from data.

We are interested to see whether we can do feature selection inside the network. To do this, We do a case study in which we free the dilation factors in a Scattering Network and train it from data.

We seek to have a scattering network trained for the task of image classification. The training and testing data are taken from the MNIST dataset of handwritten digits. We take a two-layer Scattering Network and put at the bottom an SVM for classification. The structure is illustrated in Figure 4.3. We take a Morlet wavelet and train the convolutional filters as dilations of the wavelet. The training of the dilation factor is done iteratively with the training of the linear SVM, using both deterministic and stochastic gradient descent.

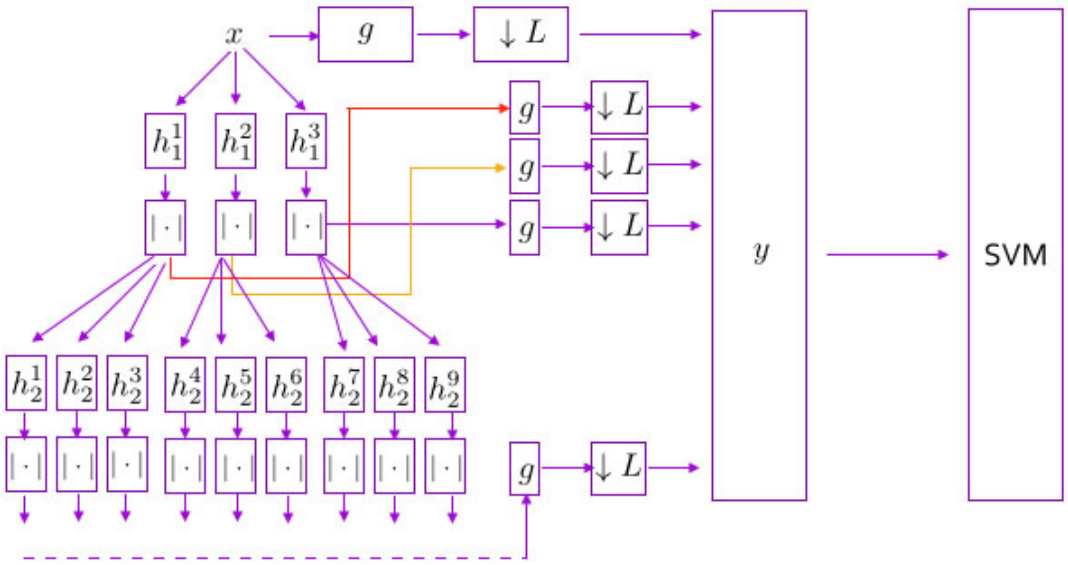

Figure 4.3: The structure of the scattering network for our case study. x is the input signal; h_k^j's are the convolutional filters taken to be the dilation of a Morlet wavelet with trained scales; g is the pooling function followed by a downsampling factor L; the feature y goes through a linear SVM to generate the classification result.

In our setting, the two-dimensional filters, h_k^j, are parametrized as dilations of the tensor products of two one-dimensional wavelets. We use the same pre-defined wavelet ψ for both. That is,

$$h_k^j(t_1, t_2) = \psi_{\lambda_{k,1}^j} \otimes \psi_{\lambda_{k,2}^j}(t_1, t_2) = \lambda_{k,1}^j \lambda_{k,2}^j \psi(\lambda_{k,1}^j t_1)\psi(\lambda_{k,2}^j t_2) \ . \tag{4.1}$$

The optimization problem associated with the linear SVM is

$$\min_{\boldsymbol{\lambda};w,b} \quad \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N} l(y_n, a_n; w, b) \ , \tag{4.2}$$

where

$$l(y, a; w, b) = \max(0, 1 - a(b + \langle w, y\rangle)) \ , \tag{4.3}$$

and y is the vector composed of the following vectors:

$$\begin{aligned} y_0 &= x * g \ ; \\ y_1^j &= \left|x * h_1^j\right| * g \ , 1 \leq j \leq 3 \ ; \\ y_2^j &= \left|\left|x * h_1^{\lceil j/3 \rceil}\right| * h_2^j\right| * g \ , 1 \leq j \leq 9 \ . \end{aligned}$$

error rate (%)	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
stochastic gradient descent	11.5	2.75	32.87	49.88	39.12	42.63	21.62	38.5	38.37	41
deterministic gradient descent	1.87	1.12	6	8.25	5.5	10.25	4.5	8	12	10.87
libSVM	3	1.62	6.25	7.5	4.87	9.37	5.25	7.75	10.87	10
Square (deterministic)	1.63	1.25	6.12	7.88	4.25	10.62	3.62	5.75	10.13	9.38

Figure 4.4: The classification results for MNIST. The error rate shows the percentage of data correctly labeled. The first row shows the results using the stochastic gradient descent method, the second row shows the results using the deterministic gradient descent method, the third row shows the results using libSVM, the fourth row shows the results where $|\cdot|$ is replaced by $|\cdot|^2$.

The testing results are shown in Figure 4.4. From the testing results, we see that training a variant version of scattering network is not successful since (1) the widely-used stochastic gradient descent method works bad (2) the result from deterministic method does not provide significant improvement from the linear SVM result. One reason is that the scattering network structure is too restrictive and learning the dilation factors does not seem to be a well-posed problem. We are

motivated to develop the theory for a more general model, preferably including all: AlexNet, GoogleNet and the scattering network.

4.2 A framework for a general convolutional neural network

We consider a CNN that maps the input signal to the representations (of the features of the signal). In most applications, a fully connected neural network is put at the bottom of the representations and outputs the classes for the input signal.a

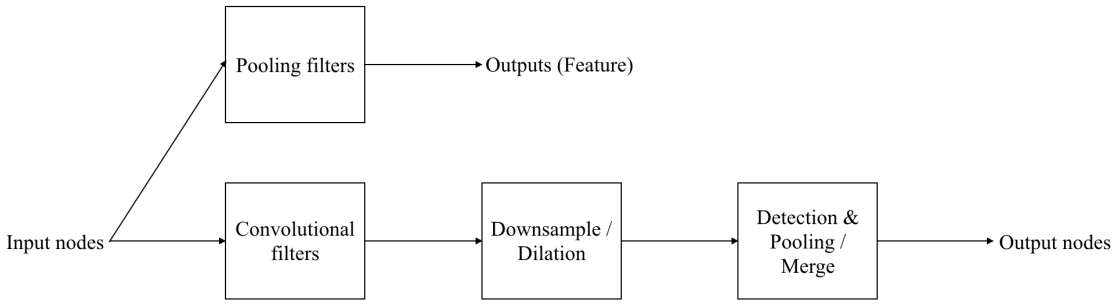

Figure 4.5: The structure of a layer of CNN. A CNN consists of a chain of layers, which makes the structure "deep".

The CNN that we consider has a feed-forward structure and the input propagates through several layers. In the CNN, each layer consists of input nodes, convolutional filters, detection / merge operations, pooling filters, outputs (feature) and output nodes. We understand each component as follows.

1. The *input nodes* are signals from the output nodes in the previous layer (it is the input of the whole network for the first layer).

2. The *convolutional filters* are the filters that perform convolution with the signal

from the input nodes. Suppose y is a signal in one of the input nodes, and g is the filter, the output is

$$z(t) = y * g(t) = \int y(t-s)g(s)ds = \int y(s)g(t-s)ds \ .$$

3. The *detection / merge operations* are nonlinear operations applied pointwise to the output of the convolutional filters. In this stage, several outputs may be merged by some pointwise operation to produce a single output.

4. The *pooling filters* lower the dimensionality/bandwidth to generate the outputs.

5. The *output nodes* are signals that are passed to the next layer. The signal on the output nodes is identical to that on the input nodes of the next layer.

The space of the filters are chosen to be the Banach Algebra of tempered distributions with an essentially bounded Fourier Transform, that is,

$$\mathcal{B} = \{f \in \mathcal{S}'(\mathbb{R}^d), \left\|\hat{f}\right\|_\infty < \infty\} \ , \tag{4.4}$$

with $\|f\|_{\mathcal{B}} := \left\|\hat{f}\right\|_\infty$ for each $f \in \mathcal{B}$.

Note that in the above definition we ask $\hat{f}$ to be an ordinary function so it makes sense to define its L^∞ norm and multiplication. We check that $\mathcal{B}$ is indeed a Banach algebra as follows.

Lemma 4.2.1. *$\mathcal{B}$ as defined in (4.4) is a Banach algebra, where the $+$ operation is pointwise addition, and the $\cdot$ operation is the convolution defined by*

$$f * g = \left(\hat{f}\hat{g}\right)^\vee \ , \tag{4.5}$$

where " $\check{}$ " denotes the inverse Fourier transform.

Proof. Note that $\mathcal{B}$ is closed under the convolution in the sense of (4.5) because $\hat{f}\hat{g} \in L^\infty(\mathbb{R}^d)$ and therefore is also in $\mathcal{S}'(\mathbb{R}^d)$. Since the Fourier transform is an isomorphism on $\mathcal{S}'(\mathbb{R}^d)$, the inverse Fourier transform of $\hat{f}\hat{g}$ also lies in $\mathcal{S}'(\mathbb{R}^d)$.

After the closedness is clear, it is trivial to check that $\mathcal{B}$ is indeed an algebra. The fact that $\mathcal{B}$ is a Banach algebra is due to the norm inequality

$$\left\|\hat{f}\hat{g}\right\|_\infty \leq \left\|\hat{f}\right\|_\infty \left\|\hat{g}\right\|_\infty . \tag{4.6}$$

□

We now return to the settings of CNN. In our framework, there can be three types of merging. Type I takes inputs $y_1, \cdots, y_k$ from k different channels, apply a nonlinearity function $\sigma_1, \cdots, \sigma_k$ respectively, and then take the sum. That is, the output is

$$z = \sum_{j=1}^{k} \sigma_j(y_j) . \tag{4.7}$$

Type II takes inputs $y_1, \cdots, y_k$ from k different channels, apply a nonlinearity, and then apply a pointwise p-norm aggregation. That is, the output is

$$z = \left(\sum_{j=1}^{k} |\sigma_j(y_j)|^p \right)^{1/p} . \tag{4.8}$$

Type III takes inputs $y_1, \cdots, y_k$ from k different channels, apply a nonlinearity with L^∞ norm bounded by 1, and then apply a pointwise multiplication. That is, the output is

$$z = \prod_{j=1}^{k} \sigma_j(y_j) , \tag{4.9}$$

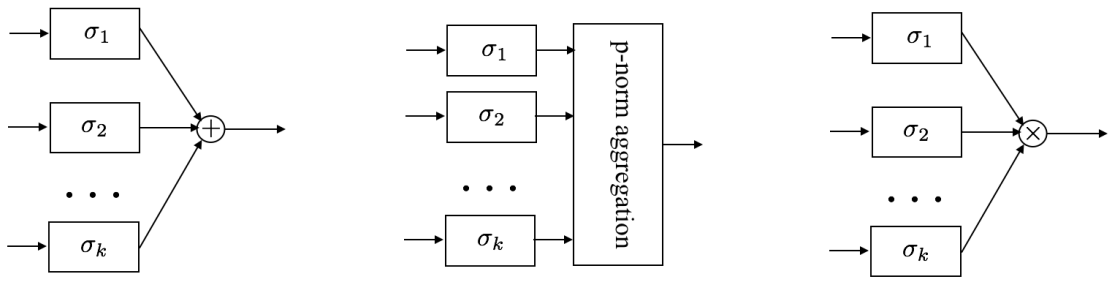

Figure 4.6: The three types of merge. Type I is taking sum of three inputs, Type II is taking p-norm aggregation, Type III is taking pointwise product.

with $\|\sigma_j\|_\infty \leq 1$ for each j.

We now stop to discuss two widely used operations that can be modeled by merging, namely, the max pooling and the average pooling. These operations are illustrated in Figure (4.7).

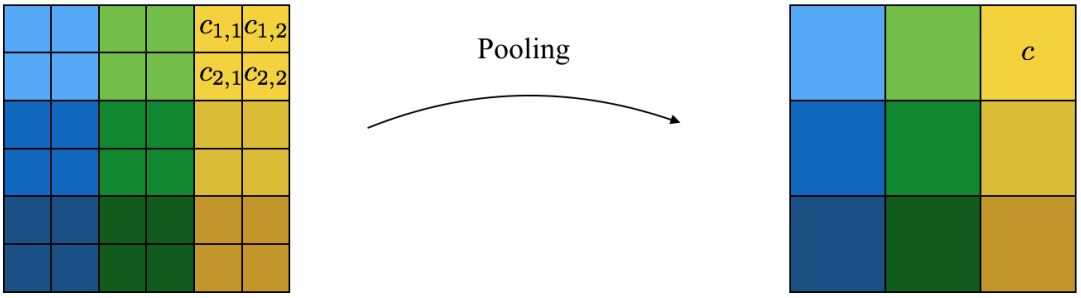

Figure 4.7: A toy example that shows how pooling works. The left image is subdivided into nine regions. The pooling operation outputs one value for each region. We take the top right corner for example. In the case of max pooling, we have $c = \max\{c_{1,1}, c_{1,2}, c_{2,1}, c_{2,2}\}$; in the case of average pooling, we have $c = (c_{1,1} + c_{1,2} + c_{2,1} + c_{2,2})/4$.

Max pooling is the operation of taking the maximal element among those in the same sub-regions. We can use translations and dilation to separate elements in a sub-region to distinct channels, as illustrated in Figure 4.8. Then a L^∞-aggregation select the largest element and does the max pooling.

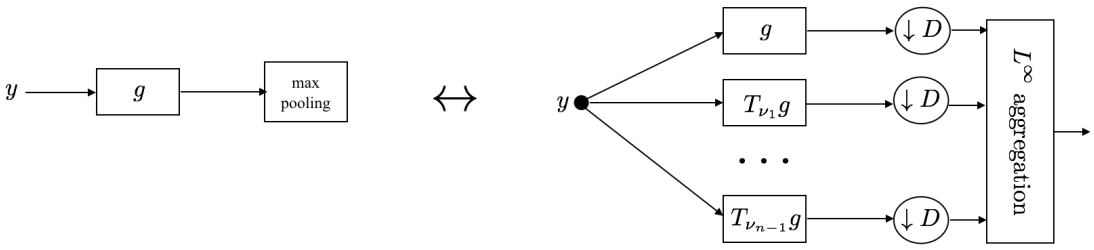

Figure 4.8: Max pooling modeled as Type II aggregation using L^∞ norm.

Average pooling replaces "taking the max" by "taking the average". Similarly to max pooling, it can be done by taking the sum as illustrated in Figure 4.9.

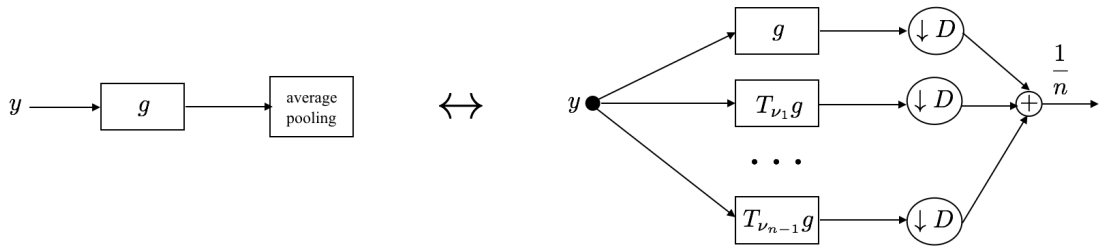

Figure 4.9: Average pooling modeled as Type I aggregation.

We illustrate the entire structure of an M-layer ConvNet as Figure 4.10.

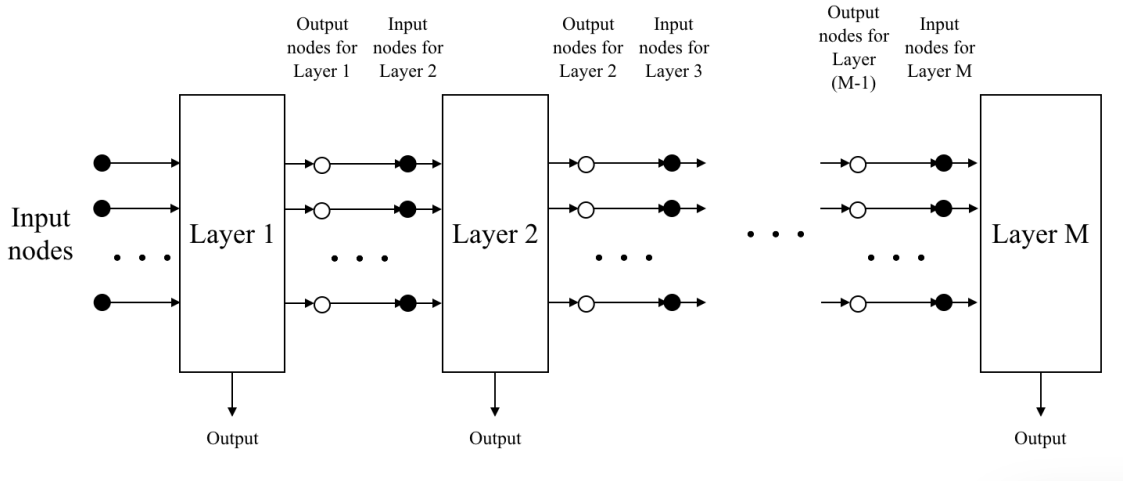

Figure 4.10: The detail of an M-layer ConvNet. The signals at output nodes are identical as at input nodes in the next layer. There may or may not be output in each layer.

Suppose there are n_m nodes in the m's layer (this works for $m < M$ but $m = M$ is a similar case in which there is no output node). We denote them by $\mathcal{I}_m = \{N_{m,1}, N_{m,2}, \cdots, N_{m,n_m}\}$. Then within the layer, each node is connected to several filters. The filter can be either a pooling filter, or a convolutional filter. Associated with $N_{m,n}$ for $1 \leq k \leq n_m$, we denote the pooling filter to be $\phi_{m,n}$, and

the convolutional filters to be $G_{m,n} = \{g_{m,n;1}, \cdots g_{m,n;k_{m,n}}\}$. Then the set of filters in the m-th layer is

$$G_m = \bigcup_{n=1}^{n_m} G_{m,n} \ . \tag{4.10}$$

The filters can be classified into three categories according to the three types of merging (if a filter is not merged with other filters, then we classify it as Type I). We denote the sets of all Type-I, II, III filters by τ_1, τ_2, τ_3, respectively.

Note that each filter is associated with one and only one output node. We use $\mathcal{O}_m = \{N'_{m,1}, N'_{m,2}, \cdots, N'_{m,n'_m}\}$ to denote the set of output nodes of the m-th layer. Note that $n'_m = n_{m+1}$ and there is a one-one correspondence between $\mathcal{O}_m$ and $\mathcal{I}_{m+1}$. The output nodes automatically divides G_m into n'_m disjoint subsets $G_m = \cup_{n'=1}^{n'_m} G'_{m,n'}$, where $G'_{m,n'}$ is the set of filters merged into $N'_{m,n'}$.

The detail of one layer is illustrated in Figure 4.11.

For each filter $g_{m,n;k}$, we define the associated multiplier $l_{m,n;k}$ in the following way: suppose $g_{m,n;k} \in G'_{m,n'}$, let $K = \left|G'_{m,n'}\right|$ denote the cardinality of $G'_{m,n'}$. Then

$$l_{m,n;k} = \begin{cases} K & \text{, if } g_{m,n;k} \in \tau_1 \cup \tau_3 \\ K^{\max\{0,2/p-1\}} & \text{, if } g_{m,n;k} \in \tau_2 \end{cases} \tag{4.11}$$

We define the 1st type Bessel bound for the node $N_{m,n}$ to be

$$B^{(1)}_{m,n} = \left\| \left|\hat{\phi}_{m,n}\right|^2 + \sum_{k=1}^{k_{m,n}} l_{m,n;k} D^{-d}_{m,n;k} \left|\hat{g}_{m,n;k}\right|^2 \right\|_\infty \ , \tag{4.12}$$

the 2nd type Bessel bound to be

$$B^{(2)}_{m,n} = \left\| \sum_{k=1}^{k_{m,n}} l_{m,n;k} D^{-d}_{m,n;k} \left|\hat{g}_{m,n;k}\right|^2 \right\|_\infty \ , \tag{4.13}$$

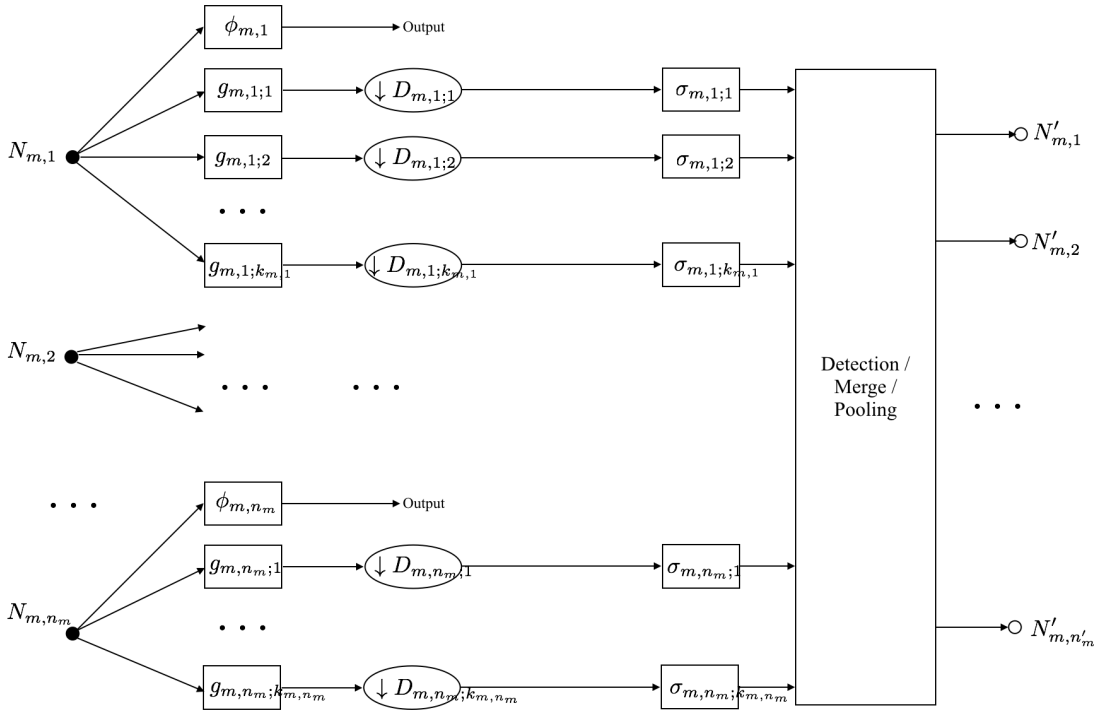

Figure 4.11: The detail of one layer. N's are the input nodes, N''s are the output nodes. ϕ's and g's are the filters, D's are the dilation factors. σ's are the nonlinearities.

and the generating bound to be

$$B_{m,n}^{(3)} = \left\| \hat{\phi}_{m,n} \right\|_{\infty}^{2} . \tag{4.14}$$

Then we define the 1st type Bessel bound for the m-th layer to be

$$B_{m}^{(1)} = \max_{1 \le n \le n_m} B_{m,n}^{(1)} , \tag{4.15}$$

the 2nd type Bessel bound to be

$$B_{m}^{(2)} = \max_{1 \le n \le n_m} B_{m,n}^{(2)} , \tag{4.16}$$

and the generating bound to be

$$B_m^{(3)} = \max_{1 \le n \le n_m} B_{m,n}^{(3)} \ . \tag{4.17}$$

4.3 Computation of the Lipschitz constant

Let a CNN defined in the previous section be given. For any input signal f and $\tilde{f}$. Let f_N be the output for f from the node N, and $\tilde{f}_N$ be the output for $\tilde{f}$ from the node N. Let V be the collection of all nodes. We say L is a Lipschitz bound for the CNN if

$$\sum_{N \in V} \left\| f_N - \tilde{f}_N \right\|_2^2 \le L \left\| f - \tilde{f} \right\|_2^2 \ . \tag{4.18}$$

Define the map $\Phi : L^2(\mathbb{R}^d) \to [L^2(\mathbb{R}^d)]^{|V|}$ by

$$\Phi(f) = (f_N)_{N \in V} \ . \tag{4.19}$$

Then a norm $|||\cdot|||$ defined on $[L^2(\mathbb{R}^d)]^{|V|}$ by

$$\left|\left|\left|(f_N)_{N \in V}\right|\right|\right| = \left(\sum_{N \in V} \|f_N\|_2^2 \right)^{1/2}$$

is well defined and $\sqrt{L}$ is a Lipschitz constant in the sense that

$$\left|\left|\left|\Phi(f) - \Phi(\tilde{f})\right|\right|\right| \le \sqrt{L} \left\| f - \tilde{f} \right\|_2 \ . \tag{4.20}$$

We have the following theorem for computing the Lipschitz bound.

Theorem 4.3.1. *Consider a ConvNet with M layers and in the m-th layer it has 1st type Bessel bound $B_m^{(1)}$, 2nd type Bessel bound $B_m^{(2)}$ and generating bound $B_m^{(3)}$.*

Then the ConvNet implies a nonlinear map that is Lipschitz continuous, and its Lipschitz bound is given by the optimal value of the following linear program:

$$\begin{aligned} \max \quad & \sum_{m=1}^{M} z_m \\ s.t. \quad & y_0 = 1 \\ & y_m + z_m \le B_m^{(1)} y_{m-1}, \quad 1 \le m \le M-1 \\ & y_m \le B_m^{(2)} y_{m-1}, \quad 1 \le m \le M-1 \\ & z_m \le B_m^{(3)} y_{m-1}, \quad 1 \le m \le M \\ & y_m, z_m \ge 0, \quad \textit{for all} \quad m \end{aligned} \tag{4.21}$$

Proof. We are going to show that the optimal value for the linear program (4.21) is a Lipschitz bound. In particular, we are going to study the sum $\sum_{N\in V} \left\| f_N - \tilde{f}_N \right\|_2^2$ as $\sum_{m=1}^{M} \sum_{N\in V_m} \left\| f_N - \tilde{f}_N \right\|_2^2$.

We take the m-th layer for analysis. With Figure 4.16, we mark the signals at the input nodes to be $h_{m,1}, \cdots, h_{m,n_m}$ and the signals at the output nodes to be $h'_{m,1}, \cdots, h'_{m,n'_m}$. We estimate the Lipschitz bound by comparing the output nodes and input nodes for each layer, and then derive a relation between the outputs and the input at the very first layer. Note that with our notation here, $h_{1,1} = f$ and $\tilde{h}_{1,1} = \tilde{f}$.

We have three types of merging. We study the relation between the output and input of the merging blocks. For Type I, see Figure 4.12.

We have

$$y_0 = \sum_{k=1}^{K} \sigma_k(y_k) \ , \tag{4.22}$$

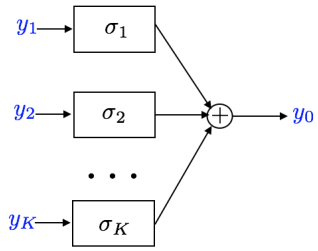

Figure 4.12: Type I merging. y_0 is the sum of $\sigma_1(y_1), \cdots, \sigma_K(y_K)$.

and

$$\tilde{y}_0 = \sum_{k=1}^{K} \sigma_k(\tilde{y}_k) \ . \tag{4.23}$$

Therefore

$$\begin{aligned} \|y_0 - \tilde{y}_0\|_2^2 &= \left\| \sum_{k=1}^{K} \sigma_k(y_k) - \sigma_k(\tilde{y}_k) \right\|_2^2 \\ &\leq K \sum_{k=1}^{K} \|\sigma_k(y_k) - \sigma_k(\tilde{y}_k)\|_2^2 \\ &\leq K \sum_{k=1}^{K} \|y_k - \tilde{y}_k\|_2^2 \ . \end{aligned} \tag{4.24}$$

For Type II, see Figure 4.13.

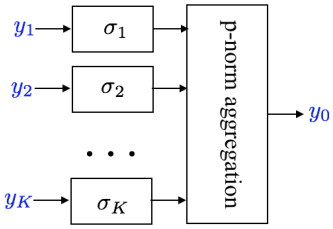

Figure 4.13: Type II merging. y_0 is the aggregate of $\sigma_1(y_1), \cdots, \sigma_K(y_K)$ using p-norm.

We have

$$y_0 = \left(\sum_{k=1}^{K} |\sigma_k(y_k)|^p \right)^{1/p} , \tag{4.25}$$

and

$$\tilde{y}_0 = \left(\sum_{k=1}^{K} |\sigma_k(\tilde{y}_k)|^p \right)^{1/p}, \tag{4.26}$$

Therefore if $p \leq 2$ we have

$$\begin{aligned}
& \|y_0 - \tilde{y}_0\|_2^2 \\
= & \left\| \left(\sum_{k=1}^{K} |\sigma_k(y_k)|^p \right)^{1/p} - \left(\sum_{k=1}^{K} |\sigma_k(\tilde{y}_k)|^p \right)^{1/p} \right\|_2^2 \\
\leq & \left\| \Big(\sum_{k=1}^{K} |\sigma_k(y_k) - \sigma_k(\tilde{y}_k)|^p \Big)^{1/p} \right\|_2^2 \\
\leq & K^{2/p-1} \cdot \left\| \Big(\sum_{k=1}^{K} |\sigma_k(y_k) - \sigma_k(\tilde{y}_k)|^2 \Big)^{1/2} \right\|_2^2 \\
= & K^{2/p-1} \cdot \sum_{k=1}^{K} \|\sigma_k(y_k) - \sigma_k(\tilde{y}_k)\|_2^2 \\
\leq & K^{2/p-1} \cdot \sum_{k=1}^{K} \|y_k - \tilde{y}_k\|_2^2 \ ;
\end{aligned} \tag{4.27}$$

and if $p > 2$ we have

$$\begin{aligned}
& \|y_0 - \tilde{y}_0\|_2^2 \\
= & \left\| \left(\sum_{k=1}^{K} |\sigma_k(y_k)|^p \right)^{1/p} - \left(\sum_{k=1}^{K} |\sigma_k(\tilde{y}_k)|^p \right)^{1/p} \right\|_2^2 \\
\leq & \left\| \Big(\sum_{k=1}^{K} |\sigma_k(y_k) - \sigma_k(\tilde{y}_k)|^p \Big)^{1/p} \right\|_2^2 \\
\leq & \left\| \Big(\sum_{k=1}^{K} |\sigma_k(y_k) - \sigma_k(\tilde{y}_k)|^2 \Big)^{1/2} \right\|_2^2 \\
= & \sum_{k=1}^{K} \|\sigma_k(y_k) - \sigma_k(\tilde{y}_k)\|_2^2 \\
\leq & \sum_{k=1}^{K} \|y_k - \tilde{y}_k\|_2^2 \ .
\end{aligned} \tag{4.28}$$

For Type III, see Figure 4.14.

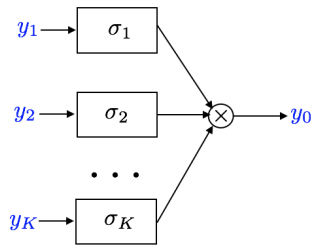

Figure 4.14: Type III merging. y_0 is the product of $\sigma_1(y_1), \cdots, \sigma_K(y_K)$, with $\|\sigma_j\|_\infty \leq 1$ for $j = 1, \cdots, K$.

We have $y_0 = \prod_{k=1}^{K} \sigma_k(y_k)$ and $\tilde{y}_0 = \prod_{k=1}^{K} \sigma_k(\tilde{y}_k)$. Therefore,

$$
\begin{aligned}
& \|y_0 - \tilde{y}_0\|_2 \\
= & \left\| \prod_{k=1}^{K} \sigma_k(y_k) - \prod_{k=1}^{K} \sigma_k(\tilde{y}_k) \right\|_2 \\
= & \left\| \prod_{k=1}^{K} \sigma_k(y_k) + \sum_{J=1}^{K-1} \Big[- \prod_{k=1}^{J} \sigma_k(y_k) \prod_{k=J+1}^{K} \sigma_k(\tilde{y}_k) + \prod_{k=1}^{J} \sigma_k(y_k) \prod_{k=J+1}^{K} \sigma_k(\tilde{y}_k) \Big] + \prod_{k=1}^{K} \sigma_k(\tilde{y}_k) \right\|_2 \\
= & \left\| \prod_{k=1}^{K-1} \sigma_k(y_k) \cdot (\sigma_K(y_K) - \sigma_K(\tilde{y}_K)) + \sum_{J=2}^{K-1} \prod_{k=1}^{J-1} \sigma_k(y_k) \cdot (\sigma_J(y_J) - \sigma_J(\tilde{y}_J)) \cdot \prod_{k=J+1}^{K} \sigma_k(\tilde{y}_k) + \right. \\
& \qquad \left. (\sigma_1(y_1) - \sigma_1(\tilde{y}_1)) \cdot \prod_{k=2}^{K} \sigma_k(\tilde{y}_k) \right\|_2 \\
\leq & \prod_{k=1}^{K-1} \|\sigma_k(y_k)\|_\infty \cdot \|\sigma_K(y_K) - \sigma_K(\tilde{y}_K)\|_2 + \\
& \qquad \sum_{J=2}^{K-1} \prod_{k=1}^{J-1} \|\sigma_k(y_k)\|_\infty \cdot \prod_{k=J+1}^{K} \|\sigma_k(\tilde{y}_k)\|_\infty \cdot \|\sigma_J(y_J) - \sigma_J(\tilde{y}_J)\|_2 + \\
& \qquad \prod_{k=2}^{K} \|\sigma_k(\tilde{y}_k)\|_\infty \cdot \|\sigma_1(y_1) - \sigma_1(\tilde{y}_1)\|_2 \\
\leq & \sum_{k=1}^{K} \|\sigma_k(y_k) - \sigma_k(\tilde{y}_k)\|_2 \\
\leq & \sum_{k=1}^{K} \|y_k - \tilde{y}_k\|_2 \ ,
\end{aligned}
\tag{4.29}
$$

and thus

$$\|y_0 - \tilde{y}_0\|_2^2 \leq K \sum_{k=1}^{K} \|y_k - \tilde{y}_k\|_2^2 \ . \tag{4.30}$$

For the downsampling / dilation operation, see Figure 4.15. We have

Figure 4.15: Downsampling / dilation. $f^{(2)}$ is the downsampled version of $f^{(1)}$.

$$\begin{aligned} \left\| f^{(2)} - \tilde{f}^{(2)} \right\|_2^2 &= \int \left| f^{(1)}(Dx) - \tilde{f}^{(1)}(Dx) \right|^2 dx \\ &= \frac{1}{D^d} \int \left| f^{(1)}(x) - \tilde{f}^{(1)}(x) \right|^2 dx \\ &= \frac{1}{D^d} \left\| f^{(1)} - \tilde{f}^{(1)} \right\|_2^2 \ . \end{aligned} \tag{4.31}$$

Therefore, when we compare the input nodes and output nodes of the m-th layer, we have

$$\begin{aligned} &\sum_{1}^{n'_m} \left\| h'_{m,n} - \tilde{h}'_{m,n} \right\|_2^2 + \sum_{n=1}^{n_m} \left\| f_{m,n} - f'_{m,n} \right\|_2^2 \\ \leq \ & B_m^{(1)} \left\| h_{m,n} - \tilde{h}_{m,n} \right\|_2^2 \ , \end{aligned} \tag{4.32}$$

where $B_m^{(1)}$ is as defined in Equation (4.15).

By the one-one correspondence of the output nodes in the $(m+1)$-th layer and the input nodes in the m-th layer, we know that

$$\sum_{n=1}^{n_{m+1}} \left\| h_{m+1,n} - \tilde{h}_{m+1,n} \right\|_2^2 = \sum_{n=1}^{n'_m} \left\| h'_{m,n} - \tilde{h}'_{m,n} \right\|_2^2 \ , \tag{4.33}$$

and therefore,

$$\begin{aligned} &\sum_{n=1}^{n_{m+1}} \left\| h_{m+1,n} - \tilde{h}_{m+1,n} \right\|_2^2 + \sum_{n=1}^{n_m} \left\| f_{m,n} - \tilde{f}_{m,n} \right\|_2^2 \\ \leq \ & B_m^{(1)} \sum_{n=1}^{n_m} \left\| h_{m,n} - \tilde{h}_{m,n} \right\|_2^2 \ , \end{aligned} \tag{4.34}$$

for $1 \leq m \leq M-1$.

If we do not consider the output generating, then the forward propagation relation is

$$\sum_{n=1}^{n_m} \left\| h_{m+1,n} - \tilde{h}_{m+1,n} \right\|_2^2 \leq B_m^{(2)} \sum_{n=1}^{n_m} \left\| h_{m,n} - \tilde{h}_{m,n} \right\|_2^2 , \tag{4.35}$$

for $1 \leq m \leq M-1$, and similarly, considering the output generating nodes alone gives

$$\sum_{n=1}^{n_m} \left\| f_{m,n} - \tilde{f}_{m,n} \right\|_2^2 \leq B_m^{(3)} \sum_{n=1}^{n_m} \left\| h_{m,n} - \tilde{h}_{m,n} \right\|_2^2 , \tag{4.36}$$

for $1 \leq m \leq M$.

Since we would like to compare $\sum_{m=1}^{M} \sum_{n=1}^{n_m} \left\| f_{m,n} - \tilde{f}_{m,n} \right\|_2^2$ with $\left\| h_{1,1} - \tilde{h}_{1,1} \right\|_2^2$, by (4.34)-(4.36), we see that the maximal value of the linear program (4.21) gives a Lipschitz bound.

□

We can also give a Lipschitz bound more explicit to compute.

Corollary 4.3.2. *Consider a CNN with M layers and in the m-th layer it has 1st type Bessel bound B_m. Then the CNN induces a nonlinear map that is Lipschitz continuous, and its Lipschitz bound is given by*

$$\prod_{m=1}^{M} \max\{1, B_m\} . \tag{4.37}$$

Proof. From the definitions of $B_{m,n}^{(1)}$, $B_{m,n}^{(2)}$ and $B_{m,n}^{(3)}$ (4.12)-(4.14) it is obvious that

$$B_{m,n}^{(1)} \leq B_{m,n}^{(2)} + B_{m,n}^{(3)} \tag{4.38}$$

and from (4.15)-(4.17) we have hence

$$B_m^{(1)} \leq B_m^{(2)} + B_m^{(3)} \tag{4.39}$$

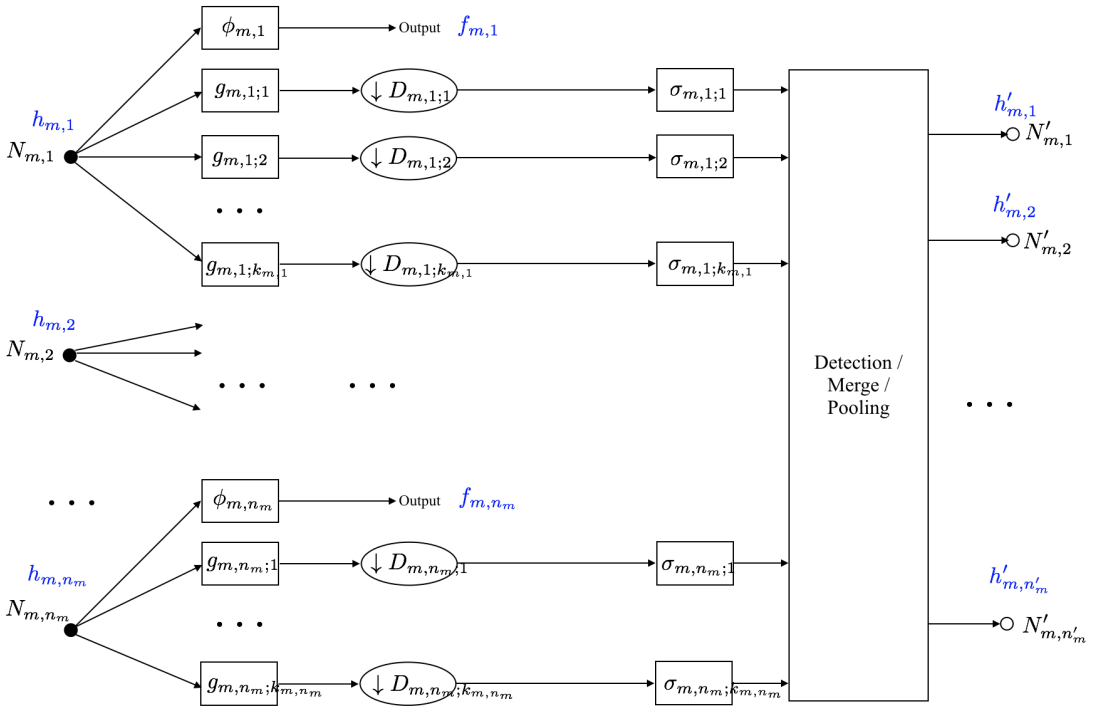

Figure 4.16: Details of one layer with signals marked as blue.

for each m. Then note that if $\{y_m\}_{m=0}^{M-1}$ and $\{z_m\}_{m=0}^{M-1}$ are the maximums of the linear program (4.21), then

$$z_m \leq B_m^{(1)} y_{m-1} - y_m, \qquad 1 \leq m \leq M-1, \tag{4.40}$$

and

$$z_M \leq B_M^{(1)} y_{M-1} \tag{4.41}$$

(note that $B_M^{(1)} = B_M^{(3)}$).

We take the sum over all m's to get (denote $y_M = 0$)

$$\begin{aligned}\sum_{m=1}^{M} z_m &\leq \sum_{m=1}^{M} B_m^{(1)} y_{m-1} - y_m \\ &= \sum_{m=0}^{M-1} B_{m+1}^{(1)} y_m - \sum_{m=1}^{M-1} y_m \\ &= B_1^{(1)} + \sum_{m=1}^{M-1} (B_{m+1}^{(1)} - 1) y_m \ .\end{aligned} \tag{4.42}$$

Also, $y_m \leq B_m^{(2)} y_{m-1}$ implies $y_m \leq B_m^{(1)} y_{m-1}$, so

$$\begin{aligned}\sum_{m=1}^{M} z_m &\leq B_1^{(1)} + \sum_{m=1}^{M-1} (\max\{1, B_{m+1}^{(1)}\} - 1) \cdot \\ &\qquad \prod_{m'=1}^{m} \max\{1, B_{m'}^{(1)}\} \\ &= \prod_{m=1}^{M} \max\{1, B_m^{(1)}\} \ .\end{aligned} \tag{4.43}$$

□

4.4 Examples

The Scattering Network, AlexNet and GoogleNet as introduced in Section 1.2 all fall in our general framework. In particular, *Scattering network* is a 1-Lipschitz map. In each layer the filters come from the wavelets used in multi-resolution analysis. A natural choice of wavelets has $B_{m,n}^{(1)} = B_{m,n}^{(2)} = B_{m,n}^{(3)} = 1$, for all m, n. Therefore, the optimal solution in the linear program (4.21) is 1. The *Alexnet* and *GoogleNet* do not generate outputs in each layer except for the last one. Therefore, $B_{m,n}^{(1)} = B_{m,n}^{(3)}$ for each $1 \leq m \leq M-1$. In this case, the result in Corollary 4.3.2 gives the optimal Lipschitz bound in the linear program (4.21).

4.4.1 A three-layer Scattering Network

Given a CNN, We can use three different approaches to estimate the Lipschitz constant. The first is by propagating backward from the outputs (either analytically or numerically) and collecting the Lipschitz constants of all the outputs. The second is by directly applying what we have discussed in Section 4.2. The third is by deriving a lower bound for L, either because of the specifies of the network or by numerical simulating.

We first give an example of a scattering network of three layers. The structure is the same as Figure 1.4. We consider the 1D case and the wavelet given by the Haar wavelets with the scaling function

$$\phi(t) = \begin{cases} 1, & \text{if } 0 \leq t < 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.44}$$

and the mother wavelet

$$\psi(t) = \begin{cases} 1, & \text{if } 0 \leq t < 1/2 \\ -1, & \text{if } 1/2 \leq t < 1 \\ 0, & \text{otherwise} \end{cases} . \tag{4.45}$$

In our convention, the sinc function is defined as

$$\text{sinc}(x) = \begin{cases} \frac{\sin(\pi x)}{(\pi x)}, & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{cases} . \tag{4.46}$$

We first look at real input functions. In this case the Haar wavelets ϕ and ψ readily satisfies the unitarity condition given as Equation (2.7) in [78]. Recall that

in the scattering network the convolutional filters are given by scales of the mother wavelet ψ:

$$\psi_{2^j}(x) = 2^{dj}\psi(2^j x) \ , \tag{4.47}$$

and the output generating filter is given by a scale of the scaling function ϕ:

$$\phi_{2^{-J}}(x) = 2^{-dJ}\phi(2^{-J}x) \ . \tag{4.48}$$

Take $J = 3$ in our example and consider all possible three-layer paths for $j = 0, -1, -2$. We have three branches from each node. Therefore we have outputs from $1 + 3 + 3^2 + 3^3 = 40$ nodes.

To convert the settings to our notations discussed in this chapter, we have a three-layer convolutional network (as in Section 2) for which the filters are given by $g_{1,l_1}, l_1 \in \{1, 2, 3\}$, $g_{2,l_2}, l_2 \in \{1, \cdots, 9\}$ and $g_{3,l_3}, l_3 \in \{1, \cdots, 27\}$, where

$$g_{m,l} = \begin{cases} \psi, & \text{if} \quad \text{mod } (l, 3) = 1; \\ \psi_{2^{-1}}, & \text{if} \quad \text{mod } (l, 3) = 2; \\ \psi_{2^{-2}}, & \text{if} \quad \text{mod } (l, 3) = 0. \end{cases}$$

$q = ((1, l_1), (2, l_2), (3, l_3))$ is a path if and only if $l_2 \in \{3l_1 - k, k = 1, 2, 3\}$ and $l_3 \in \{3l_2 - k, k = 1, 2, 3\}$. $q = ((1, l_1), (2, l_2))$ is a path if and only if $l_2 \in \{3l_1 - k, k =$

$1, 2, 3\}$. The set of all paths is

$$Q = \Big\{\emptyset, \{(1,1)\}, \{(1,2)\}, \{(1,3)\}, \{(1,1),(2,1)\}, \{(1,1),(2,2)\}, \{(1,1),(2,3)\},$$
$$\{(1,2),(2,4)\}, \{(1,2),(2,5)\}, \{(1,2),(2,6)\}, \{(1,3),(2,7)\},$$
$$\{(1,3),(2,8)\}, \{(1,3),(2,9)\}\Big\} \bigcup$$
$$\Big\{(1,l_1),(2,l_2),(3,l_3), 1 \le l_1 \le 3, \quad l_2 \in \{3l_1 - k, k = 1,2,3\},$$
$$l_3 \in \{3l_2 - k, k = 1,2,3\}\Big\} .$$

Also, for the output generation, $\phi_1 = \phi_2 = \phi_3 = \phi_4 = 2^{-J}\phi(2^{-J}\cdot)$. An illustration of the network is as in Figure 4.17, which appeared also as Figure 1.4.

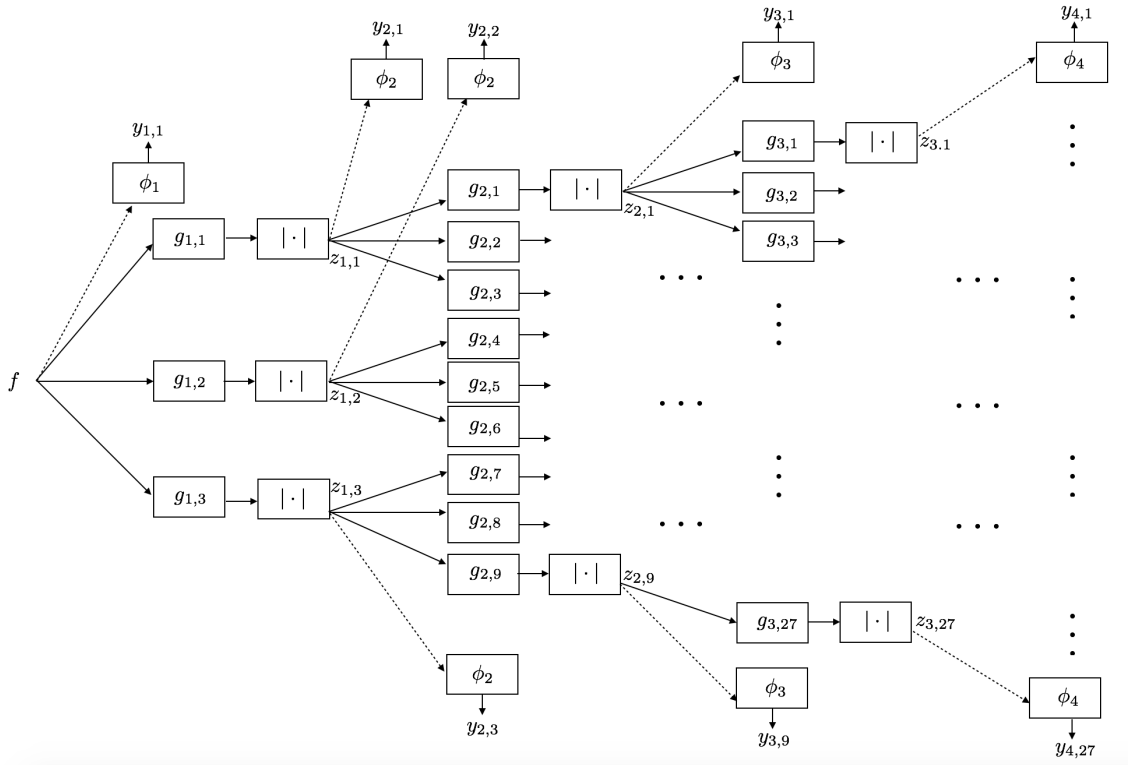

Figure 4.17: The three-layer scattering network in the example.

The first approach. We know that we have a set of 40 paths, each generating an output. We can track back from each output to the input and compute a Lipschitz

bound, and then collect all the bound to compute a total bound. To do so, we use backpropagation and the chain rule. Note that $\psi_{2^j}(t) = 2^j\psi(2^j t)$ and thus $\|\psi\|_1 = \|\psi_{2^j}\|_1 = 1$. Therefore $\|g_{m,l}\|_1 = 1$ for all m, l. Similarly, $\|\phi_j\|_1 = 1$ for all j. Let y's denote the outputs and z's denote the intermediate values, as marked in Figure 4.17. Note that each y is associated with a unique path. Consider two inputs f and $\tilde{f}$, and $r \geq 1$. Take a path $q = ((1, l_1), (2, l_2), (3, l_3))$ we have

$$\begin{aligned}
\|y_{4,l_3} - \tilde{y}_{4,l_3}\|_r &= \|(z_{3,l_3} - \tilde{z}_{3,l_3}) * \phi_4\|_r \\
&\leq \|z_{3,l_3} - \tilde{z}_{3,l_3}\|_r \|\phi_4\|_1 = \|z_{3,l_3} - \tilde{z}_{3,l_3}\|_r \ ; \\
\|z_{3,l_3} - \tilde{z}_{3,l_3}\|_r &= \||z_{2,l_2} * g_{3,l_3}| - |\tilde{z}_{2,l_2} * g_{3,l_3}|\|_r \\
&\leq \|z_{2,l_2} - \tilde{z}_{2,l_2}\|_r \|g_{3,l_3}\|_1 = \|z_{2,l_2} - \tilde{z}_{2,l_2}\|_r \ ; \\
\|z_{2,l_2} - \tilde{z}_{2,l_2}\|_r &= \||z_{1,l_1} * g_{2,l_2}| - |\tilde{z}_{1,l_1} * g_{2,l_2}|\|_r \\
&\leq \|z_{1,l_1} - \tilde{z}_{1,l_1}\|_r \|g_{2,l_2}\|_1 = \|z_{1,l_1} - \tilde{z}_{1,l_1}\|_r \ ; \\
\|z_{1,l_1} - \tilde{z}_{1,l_3}\|_r &= \left\| |f * g_{1,l_1}| - \left|\tilde{f} * g_{1,l_1}\right| \right\|_r \\
&\leq \left\|f - \tilde{f}\right\|_r \left\|g_{1,l_1}\right\|_1 = \left\|f - \tilde{f}\right\|_r \ .
\end{aligned}$$

and similarly for all output y_{m,l_m}'s. Therefore, we have

$$\left|\left|\left|\Phi(f) - \Phi(\tilde{f})\right|\right|\right|^2 = \sum_{m,l_m} \left|\left|y_{m,l_m} - \tilde{y}_{m,l_m}\right|\right|_2^2 \leq 40 \left\|f - \tilde{f}\right\|_2^2 \ .$$

The second approach. According to the result from multi-resolution analysis (see [24, 44, 77]), we have $\left|\hat{\phi}_{2^{-J}}(\omega)\right| + \sum_{j=-2}^{0} \left|\hat{\psi}_{2^j}(\omega)\right|^2 \leq 1$ (plotted in Figure 4.18), we have the first-type Bessel bounds for all layers are equal to 1. Indeed, we

can compute that

$$\left|\hat{\phi}_{2^{-J}}(\omega)\right| + \sum_{j=-2}^{0} \left|\hat{\psi}_{2^j}(\omega)\right|^2 = \operatorname{sinc}^2(8\omega) + \operatorname{sinc}^2(\omega/2)\sin^2(\pi\omega/2) + \operatorname{sinc}^2(\omega)\sin^2(\pi\omega) + \operatorname{sinc}^2(2\omega)\sin^2(2\pi\omega) \ .$$

Thus in this way, according to our discussion in Section 2, we have $\left|\left|\left|\Phi(f) - \Phi(\tilde{f})\right|\right|\right|^2 \leq \left\|f - \tilde{f}\right\|_2^2$.

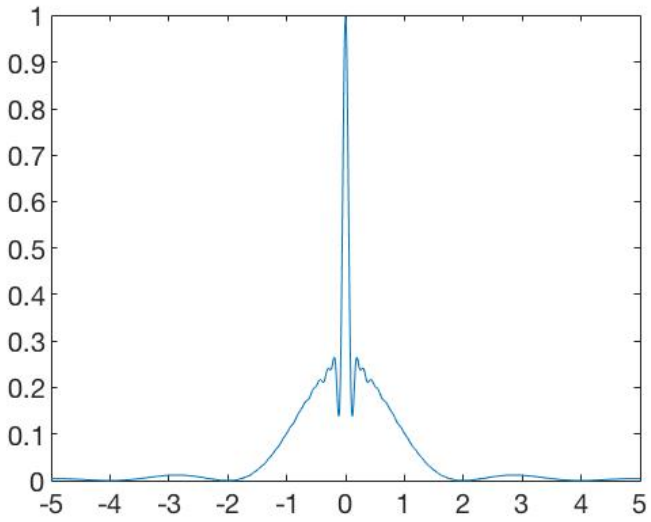

Figure 4.18: Plot of $\left|\hat{\phi}_{2^{-J}}(\omega)\right| + \sum_{j=-2}^{0} \left|\hat{\psi}_{2^j}(\omega)\right|^2$

The third approach. A lower bound for the Lipschitz constant is derived by considering only the output $y_{1,1}$ from the input layer. Obviously

$$\left|\left|\left|\Phi(f) - \Phi(\tilde{f})\right|\right|\right|^2 \geq \left\|(f - \tilde{f}) * \phi_1\right\|_1^2 \ .$$

Thus

$$\sup_{f \neq \tilde{f}} \frac{\left|\left|\left|\Phi(f) - \Phi(\tilde{f})\right|\right|\right|^2}{\left\|f - \tilde{f}\right\|_2^2} \geq \sup_{f \neq \tilde{f}} \frac{\left\|(f - \tilde{f}) * \phi_1\right\|_1^2}{\left\|f - \tilde{f}\right\|_2^2} = \left\|\hat{\phi}_1\right\|_\infty^2 = 1 \ .$$

Therefore, 1 is the exact Lipschitz bound (and Lipschitz constant) in our example.

4.4.2 A general three-layer CNN

We now give an example of how to compute the Lipschitz constant as in Figure 4.19. In Figure 4.19 f is the input, y's are the outputs and z's are the intermediate values within the network. We assume that $p \geq 2$.

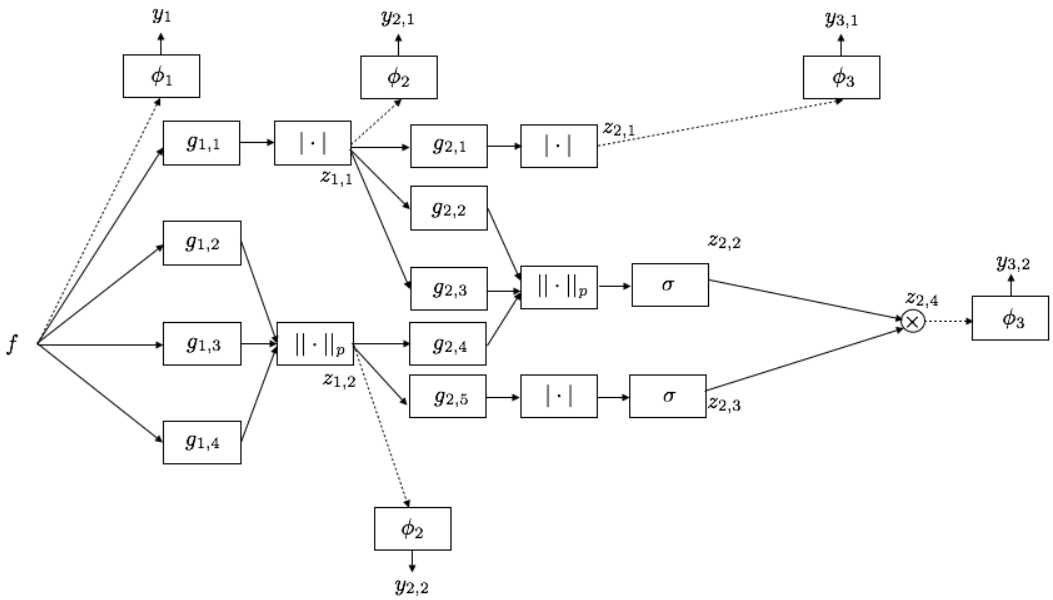

Figure 4.19: An example of a three-layer CNN. σ denotes the nonlinear pointwise function tanh.

Again we use three approaches to estimate the Lipschitz constant.

The first approach. In this approach we do not analyze the network by layers, but directly look at the outputs. We make use of the following rules: (1) backpropagation using the product rule and the chain rule; (2) each p-norm block is a multi-input-single-output nonlinear system with Lipschitz constant 1 for each channel.

Take two signals f and $\tilde{f}$. We use $\tilde{y}$'s and $\tilde{z}$'s to denote the outputs and intermediate values corresponding to $\tilde{f}$. Starting from the leftmost channels, we have for the first layer that

$$|y_1 - \tilde{y}_1| = \left|(f - \tilde{f}) * \phi_1\right| \ ,$$

and thus for any $1 \leq r \leq \infty$,

$$\|y_1 - \tilde{y}_1\|_r \leq \left\|f - \tilde{f}\right\|_r \|\phi_1\|_1 \ . \tag{4.49}$$

For the second layer we have

$$|y_{2,1} - \tilde{y}_{2,1}| = |(z_{1,1} - \tilde{z}_{1,1}) * \phi_{2,2}| \ ,$$

and thus

$$\|y_{2,1} - \tilde{y}_{2,1}\|_r \leq \|z_{1,1} - \tilde{z}_{1,1}\|_r \|\phi_2\|_1 \ .$$

With

$$\|z_{1,1} - \tilde{z}_{1,1}\|_r \leq \left\|f - \tilde{f}\right\|_r \|g_{1,1}\|_1 \ ,$$

we have

$$\|y_{2,1} - \tilde{y}_{2,1}\|_r \leq \left\|f - \tilde{f}\right\|_r \|g_{1,1}\|_1 \|\phi_2\|_1 \ . \tag{4.50}$$

Similarly,

$$\|y_{2,2} - \tilde{y}_{2,2}\|_r \leq \|z_{1,2} - \tilde{z}_{1,2}\|_r \|\phi_2\|_1 \ ,$$

and with

$$\begin{aligned}
|z_{1,2} - \tilde{z}_{1,2}| &= \Big| \left(|f * g_{1,2}|^p + |f * g_{1,3}|^p + |f * g_{1,4}|^p\right)^{1/p} - \\
&\qquad \left(\left|\tilde{f} * g_{1,2}\right|^p + \left|\tilde{f} * g_{1,3}\right|^p + \left|\tilde{f} * g_{1,4}\right|^p\right)^{1/p} \Big| \\
&\leq \left(\left|(f - \tilde{f}) * g_{1,2}\right|^p + \left|(f - \tilde{f}) * g_{1,3}\right|^p + \left|(f - \tilde{f}) * g_{1,4}\right|^p\right)^{1/p} \\
&\leq \left|(f - \tilde{f}) * g_{1,2}\right| + \left|(f - \tilde{f}) * g_{1,3}\right| + \left|(f - \tilde{f}) * g_{1,4}\right|
\end{aligned}$$

we have

$$\|z_{1,2} - \tilde{z}_{1,2}\|_r \leq \left\|f - \tilde{f}\right\| \left(\|g_{1,2}\|_1 + \|g_{1,3}\|_1 + \|g_{1,4}\|_1\right) .$$

Therefore

$$\|y_{2,2} - \tilde{y}_{2,2}\|_r \leq \left\|f - \tilde{f}\right\| \left(\|g_{1,2}\|_1 + \|g_{1,3}\|_1 + \|g_{1,4}\|_1\right) \|\phi_2\|_1 \ . \tag{4.51}$$

For the third layer we have

$$\|y_{3,1} - \tilde{y}_{3,1}\|_r \leq \|z_{2,1} - \tilde{z}_{2,1}\|_r \|\phi_3\|_1 \ .$$

With

$$\|z_{2,1} - \tilde{z}_{2,1}\|_r \leq \|z_{1,1} - \tilde{z}_{1,1}\|_r \|g_{2,1}\|_1 \ ,$$

we have

$$\|y_{3,1} - \tilde{y}_{3,1}\|_r \leq \left\|f - \tilde{f}\right\|_r \|g_{1,1}\|_1 \|g_{2,1}\|_1 \|\phi_3\|_1 \ . \tag{4.52}$$

Also,

$$\begin{aligned}
|z_{2,2} - \tilde{z}_{2,2}| &= \Big| \left(|z_{1,1} * g_{2,2}|^p + |z_{1,1} * g_{2,3}|^p + |z_{1,2} * g_{2,4}|^p\right)^{1/p} - \\
&\qquad \left(|\tilde{z}_{1,1} * g_{2,2}|^p + |\tilde{z}_{1,1} * g_{2,3}|^p + |\tilde{z}_{1,2} * g_{2,4}|^p\right)^{1/p} \Big| \\
&\leq \left(|(z_{1,1} - \tilde{z}_{1,1}) * g_{2,2}|^p + |(z_{1,1} - \tilde{z}_{1,1}) * g_{2,3}|^p + |(z_{1,2} - \tilde{z}_{1,2}) * g_{2,4}|^p\right)^{1/p} \\
&\leq |(z_{1,1} - \tilde{z}_{1,1}) * g_{2,2}| + |(z_{1,1} - \tilde{z}_{1,1}) * g_{2,3}| + |(z_{1,2} - \tilde{z}_{1,2}) * g_{2,4}| \ ,
\end{aligned}$$

which gives

$$\|z_{2,2} - \tilde{z}_{2,2}\|_r \leq \|z_{1,1} - \tilde{z}_{1,1}\|_r \left(\|g_{2,2}\|_1 + \|g_{2,3}\|_1\right) + \|z_{1,2} - \tilde{z}_{1,2}\|_r \|g_{2,4}\|_1 \ .$$

A more obvious relation is

$$\|z_{2,3} - \tilde{z}_{2,3}\|_r \leq \|z_{1,2} - \tilde{z}_{1,2}\|_r \|g_{2,5}\|_1 \ .$$

Since tanh has value bounded in $[-1,1]$, the L^∞ norm at $z_{2,2}$ and $z_{2,3}$ are bounded above by 1, and therefore we have

$$\begin{aligned}
\|z_{2,4}-\tilde{z}_{2,4}\|_r &= \|z_{2,3}z_{2,2}-\tilde{z}_{2,3}\tilde{z}_{2,2}\|_r \\
&= \|z_{2,3}z_{2,2}-\tilde{z}_{2,3}z_{2,2}+\tilde{z}_{2,3}z_{2,2}-\tilde{z}_{2,3}\tilde{z}_{2,2}\|_r \\
&\leq \|z_{2,3}-\tilde{z}_{2,3}\|_r \|z_{2,2}\|_\infty + \|\tilde{z}_{2,3}\|_\infty \|z_{2,2}-\tilde{z}_{2,2}\|_r \\
&\leq \|z_{2,2}-\tilde{z}_{2,2}\|_r + \|z_{2,3}-\tilde{z}_{2,3}\|_r \ ,
\end{aligned}$$

and consequently we have

$$\begin{aligned}
\|y_{3,2}-\tilde{y}_{3,2}\|_r &\leq \|z_{2,4}-\tilde{z}_{2,4}\|_r \|\phi_3\|_1 \\
&\leq (\|z_{2,2}-\tilde{z}_{2,2}\|_r + \|z_{2,3}-\tilde{z}_{2,3}\|_r)\|\phi_3\|_1 \\
&\leq \|z_{1,1}-\tilde{z}_{1,1}\|_r (\|g_{2,2}\|_1 + \|g_{2,3}\|_1)\|\phi_3\|_1 + \\
&\qquad \|z_{1,2}-\tilde{z}_{1,2}\|_r (\|g_{2,4}\|_1 + \|g_{2,5}\|_1)\|\phi_3\|_1 \\
&\leq \left\|f-\tilde{f}\right\|_r \Big(\|g_{1,1}\|_1 (\|g_{2,2}\|_1 + \|g_{2,3}\|_1) + \\
&\qquad (\|g_{1,2}\|_1 + \|g_{1,3}\|_1 + \|g_{1,4}\|_1)(\|g_{2,4}\|_1 + \|g_{2,5}\|_1)\Big)\|\phi_3\|_1 \ .
\end{aligned} \tag{4.53}$$

Collecting (4.49)-(4.53) we have

$$\begin{aligned}
\sum_{m,l} \|y_{m,l} - \tilde{y}_{m,l}\|_r \;\leq\; & \left\|f - \tilde{f}\right\|_r \Big(\|\phi_1\|_1 + \|g_{1,1}\|_1 \|\phi_2\|_1 + \\
& (\|g_{1,2}\|_1 + \|g_{1,3}\|_1 + \|g_{1,4}\|_1) \|\phi_2\|_1 + \\
& \|g_{1,1}\|_1 \|g_{2,1}\|_1 \|\phi_3\|_1 + \Big(\|g_{1,1}\|_1 (\|g_{2,2}\|_1 + \|g_{2,3}\|_1) + \\
& (\|g_{1,2}\|_1 + \|g_{1,3}\|_1 + \|g_{1,4}\|_1)(\|g_{2,4}\|_1 + \|g_{2,5}\|_1) \Big) \|\phi_3\|_1 \Big) \\
= \; & \left\|f - \tilde{f}\right\|_r \Big(\|\phi_1\|_1 + (\|g_{1,1}\|_1 + \|g_{1,2}\|_1 + \|g_{1,3}\|_1 + \|g_{1,4}\|_1) \|\phi_2\|_1 + \\
& \Big(\|g_{1,1}\|_1 (\|g_{2,1}\|_1 + \|g_{2,2}\|_1 + \|g_{2,3}\|_1) + \\
& (\|g_{1,2}\|_1 + \|g_{1,3}\|_1 + \|g_{1,4}\|_1)(\|g_{2,4}\|_1 + \|g_{2,5}\|_1) \Big) \|\phi_3\|_1 \Big) .
\end{aligned}$$

On the other hand we also have

$$\begin{aligned}
\left|\left|\left|\Phi(f) - \Phi(\tilde{f})\right|\right|\right|^2 \;=\; & \sum_{m,l} \|y_{m,l} - \tilde{y}_{m,l}\|_2^2 \\
\leq \; & \left\|f - \tilde{f}\right\|_2^2 \Big(\|\phi_1\|_1^2 + \|g_{1,1}\|_1^2 \|\phi_2\|_1^2 + \\
& (\|g_{1,2}\|_1 + \|g_{1,3}\|_1 + \|g_{1,4}\|_1)^2 \|\phi_2\|_1^2 + \\
& \|g_{1,1}\|_1^2 \|g_{2,1}\|_1^2 \|\phi_3\|_1^2 + \Big(\|g_{1,1}\|_1 (\|g_{2,2}\|_1 + \|g_{2,3}\|_1) + \\
& (\|g_{1,2}\|_1 + \|g_{1,3}\|_1 + \|g_{1,4}\|_1)(\|g_{2,4}\|_1 + \|g_{2,5}\|_1) \Big)^2 \|\phi_3\|_1^2 \Big) .
\end{aligned} \tag{4.54}$$

The second approach. To apply our formula, we first add δ's and form a network as in Figure 4.20. We have a four-layer network and as we have discussed,

we can compute, since $p \geq 2$, that

$$\begin{aligned}
\tilde{B}_1 &= \left\| |\hat{g}_{1,1}|^2 + |\hat{g}_{1,2}|^2 + |\hat{g}_{1,3}|^2 + |\hat{g}_{1,4}|^2 + \left|\hat{\phi}_1\right|^2 \right\|_\infty ; \\
\tilde{B}_2 &= \max\left\{1, \left\| |\hat{g}_{2,1}|^2 + |\hat{g}_{2,2}|^2 + |\hat{g}_{2,3}|^2 + \left|\hat{\phi}_2\right|^2 \right\|_\infty, \left\| |\hat{g}_{2,4}|^2 + |\hat{g}_{2,5}|^2 + \left|\hat{\phi}_2\right|^2 \right\|_\infty \right\}; \\
\tilde{B}_3 &= \max\left\{2, \left\|\hat{\phi}_3\right\|_\infty^2 \right\}; \\
\tilde{B}_4 &= \max\left\{1, \left\|\hat{\phi}_3\right\|_\infty^2 \right\}.
\end{aligned}$$

Then the Lipschitz constant is given by $(\tilde{B}_1\tilde{B}_2\tilde{B}_3\tilde{B}_4)^{1/2}$, that is,

$$\left|\left|\left|\Phi(f) - \Phi(\tilde{f})\right|\right|\right|^2 \leq (\tilde{B}_1\tilde{B}_2\tilde{B}_3\tilde{B}_4) \left\|f - \tilde{f}\right\|_2^2 . \tag{4.55}$$

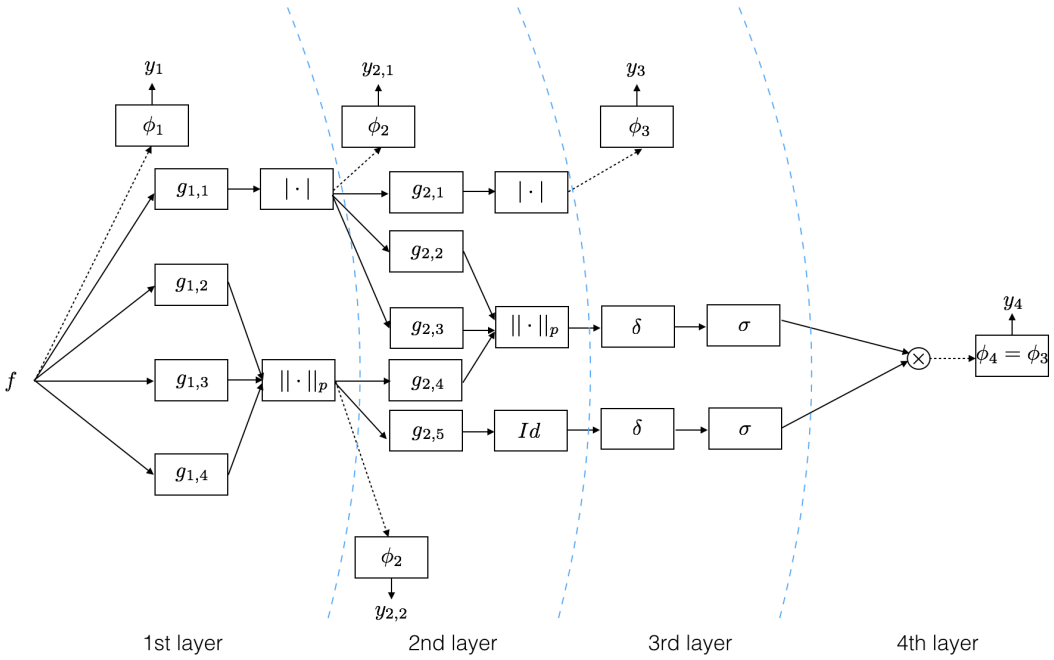

Figure 4.20: Equivalence of the example. We can clearly see four layers from this illustration.

The third approach. In general (4.55) provides a more optimal bound than (4.54) because the latter does not consider the intrinsic relations of the filters that

are grouped together in the same layer. The actual Lipschitz bound can depend on the actual design of filters, not only on the Bessel bounds. We do a numerical experiment in which the Fourier transform of the filters in the same layer are the (smoothed) characteristic functions supported disjointly in the frequency domain.

Define $F(\omega) = \exp(4\omega^2/(4\omega^2-1)) \cdot \chi_{(-1/2,0)}(\omega)$ (as illustrated in Figure 4.21), and $G(\omega) = F(-\omega)$. The filters are defined in the Fourier domain to be

$$\begin{aligned}
\hat{\phi}_1(\omega) &= F(\omega+1) + \chi_{(-1,1)}(\omega) + G(\omega-1) \\
\hat{g}_{1,1}(\omega) &= F(\omega+3) + \chi_{(-3,-2)}(\omega) + G(\omega+2) + F(\omega-2) + \chi_{(2,3)}(\omega) + G(\omega-3) \\
\hat{g}_{1,2}(\omega) &= F(\omega+5) + \chi_{(-5,-4)}(\omega) + G(\omega+4) + F(\omega-4) + \chi_{(4,5)}(\omega) + G(\omega-5) \\
\hat{g}_{1,3}(\omega) &= F(\omega+7) + \chi_{(-7,-6)}(\omega) + G(\omega+6) + F(\omega-6) + \chi_{(6,7)}(\omega) + G(\omega-7) \\
\hat{g}_{1,4}(\omega) &= F(\omega+9) + \chi_{(-9,-8)}(\omega) + G(\omega+8) + F(\omega-8) + \chi_{(8,9)}(\omega) + G(\omega-9) \\
\hat{\phi}_2(\omega) &= F(\omega+2) + \chi_{(-2,2)}(\omega) + G(\omega-2) \\
\hat{g}_{2,1}(\omega) &= F(\omega+4) + \chi_{(-4,-3)}(\omega) + G(\omega+3) + F(\omega-3) + \chi_{(3,4)}(\omega) + G(\omega-4) \\
\hat{g}_{2,2}(\omega) &= F(\omega+6) + \chi_{(-6,-5)}(\omega) + G(\omega+5) + F(\omega-5) + \chi_{(5,6)}(\omega) + G(\omega-6) \\
\hat{g}_{2,3}(\omega) &= F(\omega+8) + \chi_{(-8,-7)}(\omega) + G(\omega+7) + F(\omega-7) + \chi_{(7,8)}(\omega) + G(\omega-8) \\
\hat{g}_{2,4}(\omega) &= F(\omega+5) + \chi_{(-5,-3)}(\omega) + G(\omega+3) + F(\omega-3) + \chi_{(3,5)}(\omega) + G(\omega-5) \\
\hat{g}_{2,5}(\omega) &= F(\omega+8) + \chi_{(-8,-6)}(\omega) + G(\omega+6) + F(\omega-6) + \chi_{(6,8)}(\omega) + G(\omega-8) \\
\hat{\phi}_3(\omega) &= F(\omega+9) + \chi_{(-9,9)}(\omega) + G(\omega-9)
\end{aligned}$$

Then each function is in $C_C^\infty(\hat{\mathbb{R}})$.

We numerically compute the L^1 norms of the inverse transform of the above functions using IFFT and numerical integration with stepsize 0.025: $\|\phi_1\|_1 = 1.8265$,

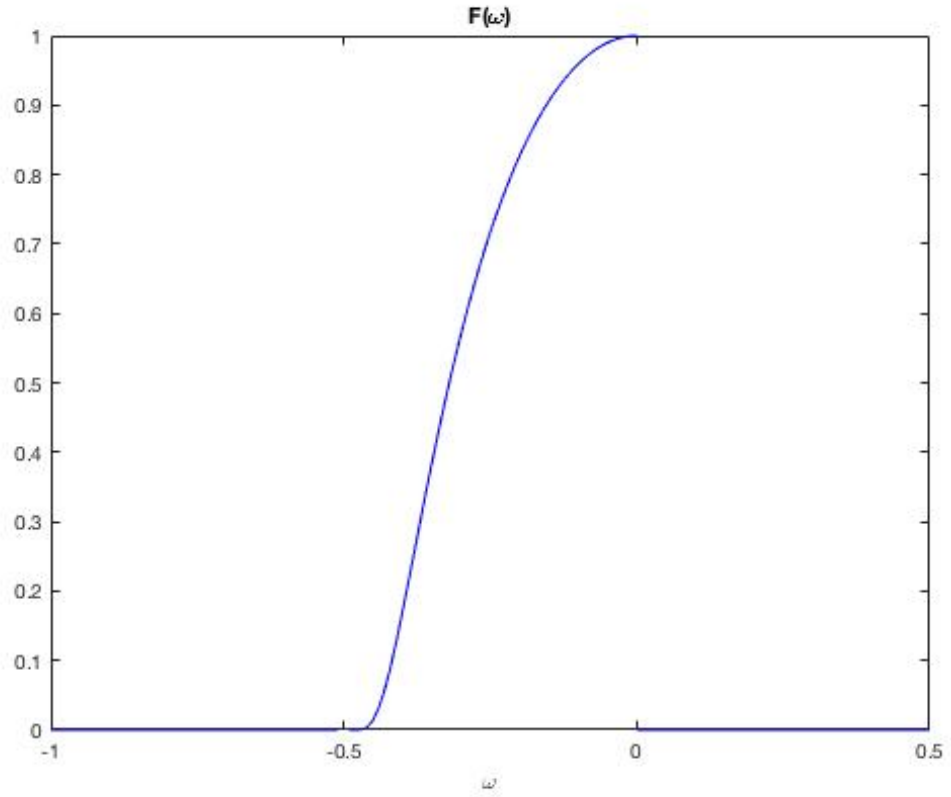

Figure 4.21: An illustration of $F(\omega)$. The functions we define in the Fourier domain are composed of the translations of F and a reflected version of F.

$\|g_{1,1}\|_1 = 2.0781$, $\|g_{1,2}\|_1 = 2.0808$, $\|g_{1,3}\|_1 = 2.0518$, $\|g_{1,4}\|_1 = 2.0720$, $\|\phi_2\|_1 = 2.0572$, $\|g_{2,1}\|_1 = 2.0784$, $\|g_{2,2}\|_1 = 2.0734$, $\|g_{2,3}\|_1 = 2.0889$, $\|g_{2,4}\|_1 = 2.2390$, $\|g_{2,5}\|_1 = 2.3175$, $\|\phi_3\|_1 = 2.6378$. Then the constant on the right-hand side of Inequality (4.54) is 966.26, and by taking the square root we conclude that the Lipschitz constant computed using the first approach is less than or equal to $\Gamma_1 = 31.1$.

It is no effort to conclude that in the second approach, $\tilde{B}_1 = \tilde{B}_2 = \tilde{B}_4 = 1$ and $\tilde{B}_3 = 2$. Therefore the Lipschitz constant computed using the second approach

is $\Gamma_2 = \sqrt{2}$.

A numerical experiment suggests that the Lipschitz bound associated with our setting of filters is about $\Gamma_3 = 1.1937$. In the experiment we numerically compute the output of the network and record the largest ratio $|||\Phi(f) - \Phi(\tilde{f})|||/||f - \tilde{f}||_2$ over one million iterations. Numerically, we consider the range $[-20, 20]$ for both the time domain and the frequency domain and take stepsize to be 0.025. For each iteration we generate two randomly signals on $[-20, 20]$ with stepsize 1 and then upsample to the same scale with stepsize 0.025.

We conclude that the naïve first approach may lead to a much larger Lipschitz bound for analysis, and the second approach gives a more reasonable estimation.

4.4.3 A comparison between Theorem 4.3.1 and Corollary 4.3.2

In the examples above, the approximation in Corollary 4.3.2 readily gives the tightest Lipschitz bound. However, it is not always the case. We shall use the same network as in the last example but a different group of filters.

Define the function on the Fourier domain supported on $(-1, 1)$ as

$$\begin{aligned} F(\omega) \;=\; & \exp\left(\frac{4\omega^2 + 4\omega + 1}{4\omega^2 + 4\omega}\right)\chi_{(-1,-1/2)}(\omega) + \\ & \chi_{(-1/2,1/2)}(\omega) + \\ & \exp\left(\frac{4\omega^2 - 4\omega + 1}{4\omega^2 - 4\omega}\right)\chi_{(1/2,1)}(\omega) \end{aligned} \tag{4.56}$$

as illustrated in Figure 4.22.

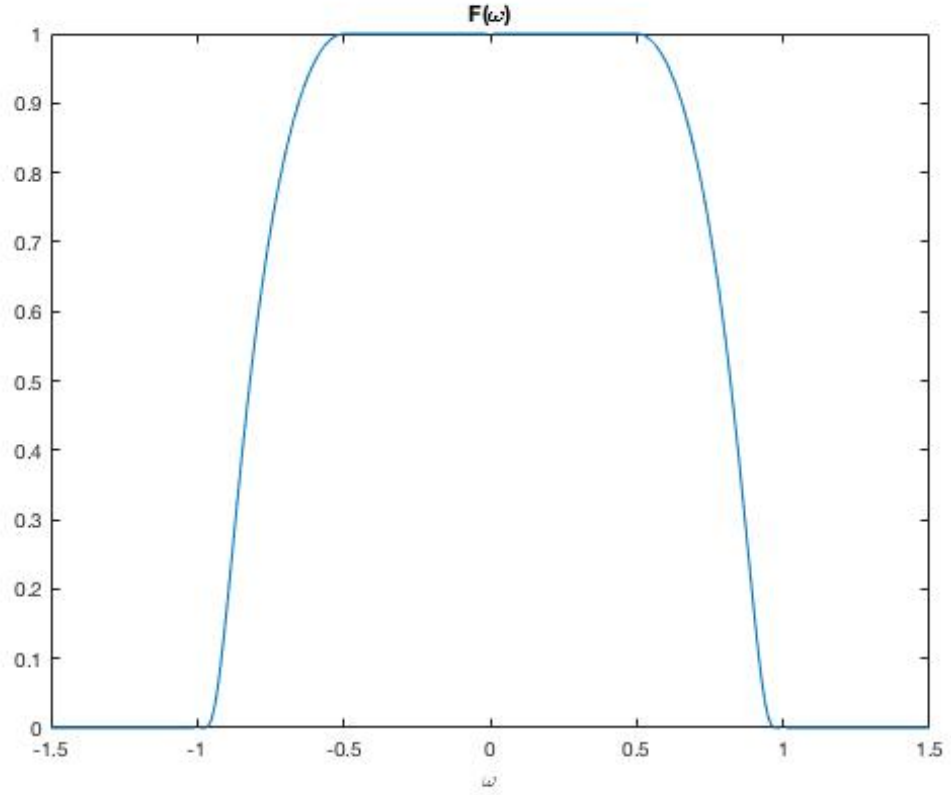

Figure 4.22: An illustration of $F(\omega)$. The functions we define in the Fourier domain all shape like F and are composed of translations of (part of) F.

With that done, we define the filters in the Fourier domain to be

$$\begin{aligned} \hat{\phi}_1(\omega) &= F(\omega) \\ \hat{g}_{1,j}(\omega) &= F(\omega + 2j - 1/2) + F(\omega - 2j + 1/2) \\ & \qquad\qquad j = 1, 2, 3, 4. \end{aligned}$$

$$\hat{\phi}_2(\omega) = \exp\left(\frac{4\omega^2+12\omega+9}{4\omega^2+12\omega+8}\right)\chi_{(-2,-3/2)}(\omega)+$$
$$\chi_{(-3/2,3/2)}(\omega)+$$
$$\exp\left(\frac{4\omega^2-12\omega+9}{4\omega^2-12\omega+8}\right)\chi_{(3/2,2)}(\omega)$$
$$\hat{g}_{2,j}(\omega) = F(\omega+2j)+F(\omega-2j)$$
$$j=1,2,3.$$
$$\hat{g}_{2,4}(\omega) = F(\omega+2)+F(\omega-2)$$
$$\hat{g}_{2,5}(\omega) = F(\omega+5)+F(\omega-5)$$
$$\hat{\phi}_3(\omega) = \exp\left(\frac{4\omega^2+20\omega+25}{4\omega^2+20\omega+24}\right)\chi_{(-3,-5/2)}(\omega)+$$
$$\chi_{(-5/2,5/2)}(\omega)+$$
$$\exp\left(\frac{4\omega^2-20\omega+25}{4\omega^2-20\omega+25}\right)\chi_{(5/2,3)}(\omega)\ .$$

Then we have $B_m^{(1)} = 2\exp(-1/3)$, $B_m^{(2)} = B_m^{(3)} = 1$ for each m. We execute a linear program using MATLAB and find out that the optimal Lipschitz bound is 2.2992, while the Lipschitz bound as derived in Corollary 4.3.2 is $8[\exp(-1/3)]^3 = 2.9430$. Therefore, in general the output of the linear program (4.21) is more optimal than the product given in Corollary 4.3.2.

4.5 Stationary processes

Signals (audio or image) are often modeled as random processes [82]. In our case, there are two ways to model the input signal of a CNN: one is to consider $X(t)$ as a random process (field) with some underlying space $(\Omega, \mathfrak{F}, \mathbb{P})$ with finite

second-order moments (see [78], Chapter 4); the other is to regard X as a random variable such that

$$X : (\Omega, \mathfrak{F}, \mathbb{P}) \to L^2(\mathbb{R}^d) \ .$$

The latter treatment makes it easy to make use of concentration inequalities (see [71, 72]). We shall present the former model for our general framework of CNN. We find out that the analysis is much parallel to the way

In this section, we shall use the notation $X(t)$ to emphasize the time (space) variable $t \in \mathbb{R}^d$ and $X_t(\omega)$ to emphasize $\omega \in \Omega$. In general, if we have dilations for random processes then after we merge the signals we lose stationarity. Thus in the section we assume that there is no dilation in our CNN.

Fix a trajectory $X(t) = X_\omega(t)$ for some $\omega \in \Omega$. Then we can define $\Phi(X)$ the same way as in (4.19). We first show that the output of a CNN is SSS provided that the input X is SSS. This is stated as the following lemma.

Lemma 4.5.1. *Suppose there is no dilation in CNN. If X is an SSS process, then so is $\Phi(X)$.*

Proof. This lemma lies on the following two facts.

1. If X is SSS, then $\sigma(X(t))$, where σ is a pointwise function, is also SSS;

2. If X is SSS, then $X * g(t)$ defined as

$$(X * g)_\omega(t) = \int X_\omega(t-s)g(s)ds \ , \tag{4.57}$$

is also SSS.

To see 1, we need to show

$$\begin{aligned} &\mathbb{P}\Big\{\sigma(X_{t_1+\tau}) \in A_1, \cdots, \sigma(X_{t_n+\tau}) \in A_n\Big\} \\ =\ &\mathbb{P}\Big\{\sigma(X_{t_1}) \in A_1, \cdots, \sigma(X_{t_n}) \in A_n\Big\} \end{aligned} \tag{4.58}$$

for any $t_1, \cdots, t_n, \tau \in \mathbb{R}^d$ and any $A_1, \cdots, A_n \in \mathfrak{F}$. Let $B_j = \sigma^{-1}(A_j) = \{c \in \mathbb{C} : \sigma(c) \in A_j\}$ for $j = 1, \cdots, n$. The above equality reads

$$\begin{aligned} &\mathbb{P}\Big\{X_{t_1+\tau} \in B_1, \cdots, X_{t_n+\tau} \in B_n\Big\} \\ =\ &\mathbb{P}\Big\{X_{t_1} \in B_1, \cdots, X_{t_n} \in B_n\Big\}\ , \end{aligned} \tag{4.59}$$

which holds true due to the assumption that X is SSS.

To see 2, note that since X is SSS there exists a semigroup of measure-preserving transformation

$$\left\{T^t : \Omega \to \Omega\right\}_{t \in \mathbb{R}^d}$$

associated with X such that

$$T^s T^t = T^{s+t}$$

for each $s, t \in \mathbb{R}^d$; and a function f such that

$$f(T^t \omega) = X_t(\omega)\ , \tag{4.60}$$

for each $\omega \in \Omega$, $t \in \mathbb{R}^d$. Thus

$$X * g(t) = \int f\left(T^{t-s}\omega\right) g(s) ds\ . \tag{4.61}$$

For any $t_1, \cdots, t_n \in \mathbb{R}^d$, $A_1, \cdots, A_n \in \mathfrak{F}$, let

$$\tilde{\Omega}_\tau = \{\omega \in \Omega : (X * g)_{t_1+\tau}(\omega) \in A_1, \cdots, (X * g)_{t_n+\tau}(\omega) \in A_n\}\ . \tag{4.62}$$

For $\omega \in \tilde{\Omega}_\tau$, note that $T^\tau \omega$ satisfies

$$(X * g)_{t_1}(\omega) \in A_1, \cdots, (X * g)_{t_n}(\omega) \in A_n \ .$$

Since T^τ is measure-preserving, we have $\mathbb{P}(\tilde{\Omega}_\tau) = \mathbb{P}(\tilde{\Omega}_0)$. Thus $X * g$ is SSS.

Given the two facts, the lemma is proved by tracking from the input to each output of the CNN. □

Theorem 4.5.2. *Assume there is no dilation in CNN. Let X and Y be SSS processes with finite second-order moments. Then*

$$\mathbb{E}\left(\left|\left|\left|\Phi(X) - \Phi(Y)\right|\right|\right|^2\right) \leq L \cdot \mathbb{E}\left(|X - Y|^2\right) \ . \tag{4.63}$$

In particular, $|||\Phi(X)|||^2 \leq L \cdot \mathbb{E}\left(|X|^2\right)$.

Proof. Since the input X and Y are SSS, so are the signals at all input and output nodes of the CNN. Therefore we can use the Wiener-Khinchin Theorem (Theorem 2.4.4) to relate the auto-correlation with the power spectrum.

Consider an SSS process Z that are filtered by some fixed $g \in \mathcal{B}$. Denote $W = Z * g$. Then by Theorem 2.4.4 we have $R_W(0) = \int \hat{S}_W(\omega) d\omega$. Note that we have the transfer relation

$$\hat{S}_W(\omega) = \hat{S}_Z(\omega) \cdot |\hat{g}(\omega)|^2 \ . \tag{4.64}$$

That is to say,

$$\mathbb{E}\left(|W|^2\right) = \int \hat{R}_W(\omega) \, |\hat{g}(\omega)|^2 \, d\omega \ . \tag{4.65}$$

More generally, due to linearity of $\mathbb{E}$, if we have two inputs Z and $\tilde{Z}$ and a family

of filters $\{g_j\}_{j\in J}$, we have

$$\begin{aligned}
\mathbb{E}\left(\sum_j \left|Z * g_j - \tilde{Z} * g_j\right|^2\right) &= \sum_j \int \hat{S}_{Z-\tilde{Z}}(\omega)\left|\hat{g}_j(\omega)\right|^2 d\omega \\
&= \int \hat{S}_{Z-\tilde{Z}}(\omega) \sum_j \left|\hat{g}_j\right|^2(\omega) d\omega \\
&\leq \int \hat{S}_{Z-\tilde{Z}}(\omega) d\omega \cdot \left\|\sum_j \left|\hat{g}_j\right|^2\right\|_\infty \\
&= \mathbb{E}\left(\left|Z-\tilde{Z}\right|^2\right) \cdot \left\|\sum_j \left|\hat{g}_j\right|^2\right\|_\infty .
\end{aligned} \tag{4.66}$$

With this, we can compare the correlation on the first input nodes with the outputs of the CNN similar to what we did in the proof of Theorem 4.3.1. Note that for merging, the inequalities (4.24), (4.27), (4.28), (4.30) still hold when $\|\cdot\|_2^2$ are replaced with $\mathbb{E}\,|\cdot|^2$. □

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*, arXiv preprint arXiv:1603.04467, (2016).

[2] R. Alaifari, I. Daubechies, P. Grohs, and G. Thakur, *Reconstructing real-valued functions from unsigned coefficients with respect to wavelet and other frames*, Journal of Fourier Analysis and Applications, (2016), pp. 1–15.

[3] R. Alaifari, I. Daubechies, P. Grohs, and R. Yin, *Stable phase retrieval in infinite dimensions*, arXiv preprint arXiv:1609.00034, (2016).

[4] R. Alaifari and P. Grohs, *Phase retrieval in the general setting of continuous frames for banach spaces*, arXiv preprint arXiv:1604.03163, (2016).

[5] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, *Phase retrieval with polarization*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 35–66.

[6] S. T. Ali, J.-P. Antoine, and J.-P. Gazeau, *Continuous frames in hilbert space*, Annals of physics, 222 (1993), pp. 1–37.

[7] J. Andén and S. Mallat, *Multiscale scattering for audio classification.*, in ISMIR, 2011, pp. 657–662.

[8] ——, *Deep scattering spectrum*, IEEE Transactions on Signal Processing, 62 (2014), pp. 4114–4128.

[9] R. Balan, *The noncommutative wiener lemma, linear independence, and spectral properties of the algebra of time-frequency shift operators*, Transactions of the American Mathematical Society, 360 (2008), pp. 3921–3941.

[10] ——, *On signal reconstruction from its spectrogram*, in Information Sciences and Systems (CISS), 2010 44th Annual Conference on, IEEE, 2010, pp. 1–4.

[11] ——, *Reconstruction of signals from magnitudes of redundant representations*, arXiv preprint arXiv:1207.1134, (2012).

[12] ——, *The fisher information matrix and the crlb in a non-awgn model for the phase retrieval problem*, in Sampling Theory and Applications (SampTA), 2015 International Conference on, IEEE, 2015, pp. 178–182.

[13] ——, *Reconstruction of signals from magnitudes of redundant representations: The complex case*, Foundations of Computational Mathematics, 16 (2016), pp. 677–721.

[14] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, *Fast algorithms for signal reconstruction without phase*, in Optical Engineering+ Applications, International Society for Optics and Photonics, 2007, pp. 67011L–67011L.

[15] ——, *Painless reconstruction from magnitudes of frame coefficients*, Journal of Fourier Analysis and Applications, 15 (2009), pp. 488–501.

[16] R. Balan, P. Casazza, and D. Edidin, *On signal reconstruction without phase*, Applied and Computational Harmonic Analysis, 20 (2006), pp. 345–356.

[17] R. Balan, M. Singh, and D. Zou, *Lipschitz properties for deep convolutional networks*, arXiv preprint arXiv:1701.05217, (2017).

[18] ——, *On stability of general deep convolutional networks*, (In preparation).

[19] R. Balan and Y. Wang, *Invertibility and robustness of phaseless reconstruction*, Applied and Computational Harmonic Analysis, 38 (2015), pp. 469–488.

[20] R. Balan and D. Zou, *On lipschitz inversion of nonlinear redundant representations*, Contemporary Mathematics, 650 (2015), pp. 15–22.

[21] ——, *On lipschitz analysis and lipschitz synthesis for the phase retrieval problem*, Linear Algebra and its Applications, 496 (2016), pp. 152–181.

[22] A. S. Bandeira, J. Cahill, D. G. Mixon, and A. A. Nelson, *Saving phase: Injectivity and stability for phase retrieval*, Applied and Computational Harmonic Analysis, 37 (2014), pp. 106–125.

[23] D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2012.

[24] J. J. Benedetto, *Wavelets: mathematics and applications*, vol. 13, CRC press, 1993.

[25] ——, *Harmonic analysis and applications*, vol. 23, CRC Press, 1996.

[26] Y. Benyamini and J. Lindenstrauss, *Geometric nonlinear functional analysis*, vol. 48, American Mathematical Soc., 1998.

[27] R. Bhatia, *Matrix analysis*, vol. 169, Springer Science & Business Media, 2013.

[28] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[29] B. G. Bodmann and N. Hammen, *Stable phase retrieval with low-redundancy frames*, Advances in computational mathematics, 41 (2015), pp. 317–331.

[30] O. Bousquet, U. Von Luxburg, and G. Rätsch, *Advanced lectures on machine learning*, in ML Summer Schools 2003, 2004.

[31] J. Bruna, S. Chintala, Y. LeCun, S. Piantino, A. Szlam, and M. Tygert, *A theoretical argument for complex-valued convolutional networks*, CoRR, abs/1503.03438 (2015).

[32] J. Bruna and S. Mallat, *Invariant scattering convolution networks*, IEEE transactions on pattern analysis and machine intelligence, 35 (2013), pp. 1872–1886.

[33] J. Cahill, P. Casazza, and I. Daubechies, *Phase retrieval in infinite-dimensional hilbert spaces*, Transactions of the American Mathematical Society, Series B, 3 (2016), pp. 63–76.

[34] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, *Phase retrieval via matrix completion*, SIAM review, 57 (2015), pp. 225–251.

[35] E. J. Candes, X. Li, and M. Soltanolkotabi, *Phase retrieval via wirtinger flow: Theory and algorithms*, IEEE Transactions on Information Theory, 61 (2015), pp. 1985–2007.

[36] E. J. Candes, T. Strohmer, and V. Voroninski, *Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming*, Communications on Pure and Applied Mathematics, 66 (2013), pp. 1241–1274.

[37] P. G. Casazza, *The art of frame theory*, Taiwanese Journal of Mathematics, (2000), pp. 129–201.

[38] P. G. Casazza, G. Kutyniok, and F. Philipp, *Finite Frames*, Springer, 2013.

[39] J. Chen, H. Dawkins, Z. Ji, N. Johnston, D. Kribs, F. Shultz, and B. Zeng, *Uniqueness of quantum states compatible with given measurement results*, Physical Review A, 88 (2013), p. 012109.

[40] Y. Chen, C. Cheng, Q. Sun, and H. Wang, *Phase retrieval of real-valued signals in a shift-invariant space*, arXiv preprint arXiv:1603.01592, (2016).

[41] O. Christensen, *Frames and bases: An introductory course*, Springer Science & Business Media, 2008.

[42] A. Conca, D. Edidin, M. Hering, and C. Vinzant, *An algebraic characterization of injectivity in phase retrieval*, Applied and Computational Harmonic Analysis, 38 (2015), pp. 346–356.

[43] D. Cox, J. Little, and D. O'shea, *Ideals, varieties, and algorithms*, vol. 3, Springer, 1992.

[44] I. Daubechies, *Ten lectures on wavelets*, SIAM, 1992.

[45] N. Dunford, J. T. Schwartz, W. G. Bade, and R. G. Bartle, *Linear operators*, Wiley-interscience New York, 1971.

[46] Y. C. Eldar and S. Mendelson, *Phase retrieval: Stability and recovery guarantees*, Applied and Computational Harmonic Analysis, 36 (2014), pp. 473–494.

[47] A. Fawzi, O. Fawzi, and P. Frossard, *Analysis of classifiers' robustness to adversarial perturbations*, arXiv preprint arXiv:1502.02590, (2015).

[48] H. G. Feichtinger and T. Strohmer, *Advances in Gabor analysis*, Springer Science & Business Media, 2012.

[49] J. R. Fienup, *Phase retrieval algorithms: a comparison*, Applied optics, 21 (1982), pp. 2758–2769.

[50] G. B. Folland, *A course in abstract harmonic analysis*, vol. 29, CRC press, 2015.

[51] ——, *Harmonic Analysis in Phase Space.(AM-122)*, vol. 122, Princeton university press, 2016.

[52] B. Gao, Q. Sun, Y. Wang, and Z. Xu, *Phase retrieval from the magnitudes of affine linear measurements*, arXiv preprint arXiv:1608.06117, (2016).

[53] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk, *Fast alternating direction optimization methods*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 1588–1623.

[54] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT Press, 2016.

[55] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Advances in neural information processing systems, 2014, pp. 2672–2680.

[56] D. R. Grayson and M. E. Stillman, *Macaulay 2, a software system for research in algebraic geometry*, 2002.

[57] U. Grenander, *Tutorial in pattern theory*, Brown University, Division of Applied Mathematics, 1983.

[58] K. Gröchenig, *Foundations of time-frequency analysis*, Springer Science & Business Media, 2013.

[59] J. Harris, *Algebraic geometry: a first course*, vol. 133, Springer Science & Business Media, 2013.

[60] M. H. Hayes, *Statistical digital signal processing and modeling*, John Wiley & Sons, 2009.

[61] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neural networks and learning machines*, vol. 3, Pearson Upper Saddle River, NJ, USA:, 2009.

[62] T. Heinosaari, L. Mazzarella, and M. M. Wolf, *Quantum tomography under prior information*, Communications in Mathematical Physics, 318 (2013), pp. 355–374.

[63] M. Hirn, S. Mallat, and N. Poilvert, *Wavelet scattering regression of quantum chemical energies*, arXiv preprint arXiv:1605.04654, (2016).

[64] M. J. Hirn and E. Y. Le Gruyer, *A general theorem of existence of quasi absolutely minimal lipschitz extensions*, Mathematische Annalen, 359 (2014), pp. 595–628.

[65] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neural Comput., 9 (1997), pp. 1735–1780.

[66] M. Kech, *Explicit frames for deterministic phase retrieval via phaselift*, Applied and Computational Harmonic Analysis, (2016).

[67] L. Koralov and Y. G. Sinai, *Theory of probability and random processes*, Springer Science & Business Media, 2007.

[68] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in Advances in neural information processing systems, 2012, pp. 1097–1105.

[69] R. Kueng, H. Rauhut, and U. Terstiege, *Low rank matrix recovery from rank one measurements*, Applied and Computational Harmonic Analysis, 42 (2017), pp. 88–116.

[70] Y. Lecun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, 521 (2015), pp. 436–444.

[71] M. Ledoux, *The concentration of measure phenomenon*, no. 89, American Mathematical Soc., 2005.

[72] M. Ledoux and M. Talagrand, *Probability in Banach Spaces*, Springer–Verlag, 1991.

[73] Y. Li, K. He, J. Sun, et al., *R-fcn: Object detection via region-based fully convolutional networks*, in Advances in Neural Information Processing Systems, 2016, pp. 379–387.

[74] E. H. Lieb and M. Loss, *Analysis, volume 14 of graduate studies in mathematics*, American Mathematical Society, Providence, RI,, 4 (2001).

[75] R. Livni, S. Shalev-shwartz, and O. Shamir, *On the computational efficiency of training neural networks*, in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds., Curran Associates, Inc., 2014, pp. 855–863.

[76] D. J. MacKay, *Information theory, inference and learning algorithms*, Cambridge university press, 2003.

[77] S. Mallat, *A wavelet tour of signal processing*, Academic press, 1999.

[78] ——, *Group invariant scattering*, Communications on Pure and Applied Mathematics, 65 (2012), pp. 1331–1398.

[79] ——, *Understanding deep convolutional networks*, Phil. Trans. R. Soc. A, 374 (2016), p. 20150203.

[80] M. Moscoso, A. Novikov, and G. Papanicolaou, *Coherent imaging without phases*, SIAM Journal on Imaging Sciences, 9 (2016), pp. 1689–1707.

[81] A. Newell, K. Yang, and J. Deng, *Stacked hourglass networks for human pose estimation*, in European Conference on Computer Vision, Springer, 2016, pp. 483–499.

[82] A. V. Oppenheim, *Discrete-time signal processing*, Pearson Education India, 1999.

[83] K. R. Parthasarathy, *An introduction to quantum stochastic calculus*, vol. 85, Birkhäuser, 2012.

[84] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, *How to construct deep recurrent neural networks*, arXiv preprint arXiv:1312.6026, (2013).

[85] F. Riesz and S. Nagy, *B.(1990). Functional analysis*, vol. 3, Dover Publications, Inc., New York. First published in, 1955.

[86] W. Rudin, *Real and complex analysis*, Tata McGraw-Hill Education, 1987.

[87] ——, *Functional analysis. International series in pure and applied mathematics*, McGraw-Hill, Inc., New York, 1991.

[88] T. N. Sainath, O. Vinyals, A. W. Senior, and H. Sak, *Convolutional, long short-term memory, fully connected deep neural networks*, in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015, 2015, pp. 4580–4584.

[89] O. Shamir and T. Zhang, *Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.*, in ICML (1), 2013, pp. 71–79.

[90] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, *Phase retrieval with application to optical imaging: a contemporary overview*, IEEE Signal Processing Magazine, 32 (2015), pp. 87–109.

[91] S. Sheffield and C. K. Smart, *Vector-valued optimal lipschitz extensions*, Communications on Pure and Applied Mathematics, 65 (2012), pp. 128–154.

[92] E. M. Stein, *Harmonic Analysis (PMS-43): Real-Variable Methods, Orthogonality, and Oscillatory Integrals.(PMS-43)*, vol. 43, Princeton University Press, 2016.

[93] E. M. Stein and R. Shakarchi, *Functional Analysis: Introduction to Further Topics in Analysis*, vol. 4, Princeton University Press, 2011.

[94] J. Sun, Q. Qu, and J. Wright, *A geometric analysis of phase retrieval*, in Information Theory (ISIT), 2016 IEEE International Symposium on, IEEE, 2016, pp. 2379–2383.

[95] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions*, in CVPR 2015, 2015.

[96] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, *Intriguing properties of neural networks*, CoRR, abs/1312.6199 (2013).

[97] T. Tao, *Topics in random matrix theory*, vol. 132, American Mathematical Society Providence, RI, 2012.

[98] J. A. Tropp et al., *An introduction to matrix concentration inequalities*, Foundations and Trends in Machine Learning, 8 (2015), pp. 1–230.

[99] M. Tygert, J. Bruna, S. Chintala, Y. LeCun, S. Piantino, and A. Szlam, *A mathematical motivation for complex-valued convolutional networks*, Neural computation, (2016).

[100] P. P. Vaidyanathan, *Multirate systems and filter banks*, Pearson Education India, 1993.

[101] A. Vedaldi and K. Lenc, *Matconvnet: Convolutional neural networks for matlab*, in Proceedings of the 23rd ACM international conference on Multimedia, ACM, 2015, pp. 689–692.

[102] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, arXiv preprint arXiv:1011.3027, (2010).

[103] Y. Wang and Z. Xu, *Generalized phase retrieval: measurement number, matrix recovery and beyond*, arXiv preprint arXiv:1605.08034, (2016).

[104] J. H. Wells and L. R. Williams, *Embeddings and extensions in analysis*, vol. 84, Springer Science & Business Media, 2012.

[105] T. Wiatowski and H. Bölcskei, *Deep convolutional neural networks based on semi-discrete frames*, in Proc. of IEEE International Symposium on Information Theory (ISIT), June 2015, pp. 1212–1216.

[106] ——, *A mathematical theory of deep convolutional neural networks for feature extraction*, IEEE Transactions on Information Theory, (2015).

[107] M. M. Wilde, *Quantum information theory*, Cambridge University Press, 2013.

[108] L. Zwald and G. Blanchard, *On the convergence of eigenspaces in kernel principal component analysis*, in NIPS, 2005, pp. 1649–1656.