OSCILLATORY DATA ANALYSIS AND FAST ALGORITHMS FOR INTEGRAL OPERATORS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF MATHEMATICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Haizhao Yang
July 2015

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Lexing Ying) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Emmanuel Candès)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Lenya Ryzhik)

Approved for the University Committee on Graduate Studies

Abstract

This dissertation consists of two independent parts: oscillatory data analysis (Part I) and fast algorithms for integral operators in computational harmonic analysis (Part II).

The first part concentrates on developing theory and efficient tools in applied and computational harmonic analysis for oscillatory data analysis. In modern data science, oscillatory data analysis aims at identifying and extracting principle wave-like components, which might be nonlinear and non-stationary, underlying a complex physical phenomenon. Estimating instantaneous properties of one-dimensional components or local properties of multi-dimensional components has been an important topic in various science and engineering problems in resent three decades. This thesis introduces several novel synchrosqueezed transforms (SSTs) with rigorous mathematical, statistical analysis, and efficient implementation to tackle challenging problems in oscillatory data analysis. Several real applications show that these transforms provide an elegant tool for oscillatory data analysis. In many applications, the SST-based algorithms are significantly faster than the existing state-of-art algorithms and obtain better results.

The second part of this thesis proposes several fast algorithms for the numerical implementation of several integral operators in harmonic analysis including Fourier integral operators (including pseudo differential operators, the generalized Radon transform, the nonuniform Fourier transform, etc.) and special function transforms (including the Fourier-Bessel transform, the spherical harmonic transform, etc.). These are useful mathematical tools in a wide range of science and engineering problems, e.g., imaging science, weather and climate modeling, electromagnetics, quantum chemistry, and phenomena modeled by wave equations. Via hierarchical domain decomposition, randomized low-rank approximations, interpolative low-rank approximations, the fast Fourier transform, and the butterfly algorithm, I propose several novel fast algorithms for applying or recovering these operators.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Prof. Lexing Ying, for his support, guidance and encouragement in my graduate study. His creative and critical thinking has been inspiring me throughout my study at Stanford. He leads me by example and teaches me how to ask questions and solve problems in the way towards a pure scientist. I also extend my gratitude to people in Prof. Ying's group for the happy time together.

I am also grateful to Prof. Ingrid Daubechies and Prof. Jianfeng Lu at Duke University for their tremendous support and fruitful discussions in the research of oscillatory data analysis and its applications. They opened my eyes to the joy and beauty of mathematics in materials science and art investigation.

I feel privileged to have many fantastic professors who taught me and inspired me a lot at Stanford. I express my special thanks to my thesis committee members, Prof. Biondo Biondi, Prof. Emmanuel Candès, Prof. Lenya Ryzhik, and Prof. Andràs Vasy for their help and insights in the mathematical science.

I would like to thank Prof. Kui Ren at the University of Texas at Austin, Prof. Weinan E at Princeton University, Prof. David Cai in Courant Institute of Mathematical Science at New York University, and Prof. Jingfang Huang at the University of North Carolina, Chapel Hill for their help, care and encouragement.

I would also like to thank my collaborators and friends for their help and discussions in my research: William P. Brown, Prof. Xiaoling Cheng, Prof. Charles K. Chui, Prof. Sergey Fomel, Dr. Kenneth L. Ho, Prof. Jingwei Hu, Yingzhou Li, Eileen R. Martin, Prof. Benedikt Wirth, Prof. Hau-Tieng Wu, Prof. Zhenli Xu.

Most of all, I would like to thank my parents Zhenqiang Yang, Guiying Yang and my wife Tian Yang for their endless source of love and support. I dedicate this thesis to them.

Contents

Part I

Oscillatory Data Analysis

Chapter 1

Introduction

The first part of this thesis is concerned with oscillatory data analysis arising in a wide range of science and engineering problems. Let $f(x) : \mathbb{R}^d \to \mathbb{R}$ be a function of d variables. We consider a class of oscillatory functions that are superpositions of several nonlinear and non-stationary wave-like components contaminated with addictive noise, i.e.,

$$f(x) = \sum_{k=1}^{K} \alpha_k(x) e^{2\pi i N_k \phi_k(x)} + T(x) + e(x),$$

where for each k $\alpha_k(x)$ is a smooth amplitude function, $2\pi N_k \phi_k(x)$ is a phase function with a smooth instantaneous frequency $N_k \phi_k'(x)$ (or a smooth local wave vector $N_k \nabla \phi_k(x)$ for $d > 1$), $T(x)$ is a smooth trend function, and $e(x)$ is a noisy perturbation term. Our goal is to identify $\alpha_k e^{2\pi i \phi_k(x)}$, $\alpha_k(x)$, $\phi_k(x)$ and $T(x)$ from the superposition above. In more complicated applications, the exponential waveform $e^{2\pi i \phi_k(x)}$ is replaced with an unknown waveform $s_k(2\pi\phi_k(x))$ to be estimated.

We propose and analyze a series of synchrosqueezed transforms (SSTs) to tackle this problem. The SST is a special time-frequency reassignment method that sharpens a linear time-frequency representation with a synchrosqueezing procedure based upon the local oscillation of the original time-frequency representation. This procedure enjoys a simple and efficient reconstruction formula, which is especially important to high dimensional applications. Synchrosqueezed transforms are local and non-parametric transforms that adapt to different data characteristics by choosing suitable linear time-frequency transforms before synchrosqueezing. Finally, synchrosqueezed transforms are visually informative with good readability - a good concentration of spectral energy and no misleading interference. They allow human interaction in spectral analysis for better understanding of the data.

We will apply this new technique to address various real problems, e.g., clinical data [171, 173],

seismic data [86, 153, 184], climate data [178], atomic materials science [126, 180] and art investigation in canvas [179]. In many cases, our algorithms are significantly faster than the existing state-of-art algorithms and obtain better results. To simplify the introduction, we will motivate the present work with dimension $d = 1$ in this chapter.

1.1 Time-Frequency Geometry

We can enjoy the beauty of music by perceiving its time-varying frequencies. But in general it is impossible to exactly "hear" the instantaneous frequency at a given time by the Heisenberg uncertainty principle [75]. The concept of instantaneous frequency is even not well defined in mathematics. In a simple cosine modulation

$$f(x) = \alpha \cos(2\pi(Nx + \theta)) := \alpha \cos(2\pi N\phi(x)),$$

it has a frequency equal to $\phi'(x) = N$. In a more general situation when

$$f(x) = \alpha(x) \cos(2\pi N\phi(x)), \tag{1.1}$$

a naive attempt is to define the time-varying instantaneous frequency of $f(x)$ to be the derivative of the phase function $N\phi(x)$. However, this definition is not unique because there are many possible choices of $\alpha(x)$ and $\phi(x)$ to satisfy (1.1). One possible solution is to consider the analytic signal for $x \in \mathbb{R}$

$$f(x) = \alpha(x)e^{2\pi i N\phi(x)} \tag{1.2}$$

corresponding to (1.1) when $\alpha(x)$ is smooth enough, i.e., $|\alpha'(x)| \ll N\phi'(x)$. The analytic signal can be computed by the Hilbert transform. For this analytic signal, the definitions of the instantaneous frequency $N\phi'(x)$ and the instantaneous amplitude $\alpha(x)$ are unique. Nevertheless, when a signal $f(x)$ contains two analytic components with two phase functions $N_1\phi_1(x)$ and $N_2\phi_2(x)$, the instantaneous frequencies $N_1\phi_1'(x)$ and $N_2\phi_2'(x)$ are well-defined only if $|N_1\phi_1'(x) - N_2\phi_2'(x)| \geq \delta(N_1\phi_1'(x) + N_2\phi_2'(x))$ for some pre-assumed constant $\delta > 0$. Otherwise, $f(x)$ can be considered as only one wave-like component with a slightly oscillatory instantaneous frequency determined by $N_1\phi_1'(x)$ and $N_2\phi_2'(x)$ if they are too close [128, 172]. This leads to the following definition.

Definition 1.1.1. *(Instantaneous frequencies in a superposition) Suppose $f(x)$ is a superposition of K wave-like components in the following form*

$$f(x) = \sum_{k=1}^{K} \alpha_k(x)e^{2\pi i N_k\phi_k(x)}$$

with $N_k\phi_k'(x) \gg |\alpha_k'(x)|$, $\alpha_k(x) > 0$, and $N_{k+1}\phi_{k+1}'(x) - N_k\phi_k'(x) \geq \delta(N_{k+1}\phi_{k+1}'(x) + N_k\phi_k'(x))$ for

some $\delta > 0$ and all k. Then the instantaneous frequencies of $f(x)$ are $N_k\phi_k'(x)$ for $1 \le k \le K$.

In signal community, analyzing time-frequency geometry of instantaneous frequencies has been a traditional research line dating back to the Gabor transform (the windowed Fourier transform). Recent challenges in various science and engineering problems have drawn people's new attention. There have been various powerful tools trying to bypass the curse of Heisenberg uncertainty principle or to reduce its effect. Most of them fall into four categories: linear time-frequency transforms, quadratic time-frequency transforms, time-frequency reassignment methods, and time-frequency optimization with pursuits.

Linear methods, e.g., taking the energy spectrogram of a Gabor transform or a wavelet transform, are typically efficient but provide poor resolution to visualize time-frequency geometry due to the Heisenberg uncertainty principle. Ridge extraction methods were proposed by Delprat, Escudiè, Guillenmain, Kronland-Martinet, Tchamitchianm, and Torrèsani [48, 81] based on the observation that ridges of the spectrogram reveal instantaneous frequencies if the window size of these transforms in the time domain is sufficiently small. More recently, the idea of ridge extraction was revisited by Aoi, Lepage, Lim, Eden and Gardner in [5] using the chirplet transform, by Chui and Mhaskar in [35] using a special windowed Fourier transform. Since ridges are not well-defined in a noisy time-frequency plane, many efforts have been made for robust ridge extraction. Statistical analysis of these methods are still under active research.

Quadratic methods mainly belong to the Cohen's class of bilinear time-frequency energy distributions [38], among which the Wigner-Ville distribution [167, 168] and its variants [39, 87] are most commonly used. For an individual wave-like component, its instantaneous frequency is exactly the "average" frequency computed relative to the Wigner-Ville distribution. However, this nice property is not true in a superposition of several components due to the interference between different components. Although this interference could be attenuated with a smoothing process, the smoothed distribution gets blurry and loses its accuracy as a trade-off. In spite of many nice properties in the theory of quadratic methods, their applications to real problems are limited by the computational efficiency and the lack of straightforward reconstruction.

Time-frequency reassignment methods [6, 7, 8, 27, 28, 44, 74] are post-processing techniques to improve the readability of the original linear or quadratic time-frequency transform by modifying the original spectral energy distribution. After reassignment, the spectral energy will concentrate around instantaneous frequencies without artificial interference. The reassignment idea was originally proposed by Kodera, Gendrin, and Villedary in [112, 113] and was revisited by Auger and Flandrin in [7] for wider applications, both conceptually and computationally. In parallel with [7], other techniques in the framework of reassigning time-frequency representations were developed independently, e.g., the differential reassignment [28] by Chassande-Mottin, Daubechies, Auger, and Flandrin; the synchrosqueezed transform [44] by Daubechies and Maes. An introductory review with recent development of reassignment methods is presented in [8] by Auger et al.

Time-frequency optimization with pursuits [20, 31, 93, 129] is a category of various optimization models based on sparsity in a redundant time-frequency dictionary, smoothness of object functions, etc. The matching pursuit introduced in [129] by Mallat and Zhang computes a sparse time-frequency representation from a redundant dictionary by iteratively selecting the most prominent atom in the dictionary. This process prevents frequency smearing and leakage in the time-frequency plane. By embedding the energy of each atom in the time-frequency plane, one can create a localized spectral energy distribution. More recently, a nonlinear matching pursuit in [93] by Hou and Shi adaptively learns a good dictionary instead of assuming a fixed redundant dictionary a priori according to the smoothness of object functions. The basis pursuit proposed in [31] by Chen, Donoho, and Saunders computes a nearly optimal sparse time-frequency representation via ℓ_1 optimization and creates a localized time-frequency representation similar to the matching pursuit. The basis pursuit method is a convex optimization that is computationally more efficient. In the scope of time-frequency detection with heavy noise, Candès, Charlton, and Helgason proposed a robust and efficient path pursuit method for detecting a single wave-like component and estimating its instantaneous frequency in [20] . A multi-component detection technique is still under development.

1.2 Mode Decomposition

For a superposition of several nonlinear and non-stationary wave-like components contaminated with addictive noise, i.e.,

$$f(x) = \sum_{k=1}^{K} \alpha_k(x) e^{2\pi i N_k \phi_k(x)} + T(x) + e(x), \tag{1.3}$$

the mode decomposition problem aims at extracting the smooth trend function $T(x)$ and each oscillatory component $\alpha_k(x)e^{2\pi i N_k \phi_k(x)}$, in addition to analyzing the time-frequency geometry $\{N_k\phi_k'(x)\}$. An ideal analysis tool should give a time-frequency representation with good readability (spectral energy concentrating around instantaneous frequencies without artificial interference) and has an efficient numerical implementation to transform, separate and reconstruct signals. An important example is the wave field or seismic event separation problem in seismic data analysis, i.e., a seismic record is decomposed into elementary wave-like components corresponding to individual wave arrivals [72, 73, 116, 143, 163, 184]. In these problems, an amplitude function $\alpha_k(x)$ may have a localized support enlarging the frequency band of the corresponding component $\alpha_k(x)e^{2\pi i\phi_k(x)}$ [116, 184]. The mode decomposition problem becomes much more complicated in such cases and many traditional time-frequency analysis tools come short of expectation.

One famous method is the empirical mode decomposition (EMD) method proposed and refined by Huang et al. in [97, 98]. The 1D EMD method decomposes a signal via a sifting process and applies the Hilbert transform to estimate the instantaneous frequency of each separated component. Starting from the most oscillatory component, the sifting process applies spline interpolation with local

extrema of the signal to identify a qualified oscillatory component and subtracts it from the signal. Although many efforts have been made [47, 54, 70, 71], the mathematical analysis of this method is still under development. In spite of many successful applications of the EMD method, its application to noisy data is limited by the lack of robustness due to the dependence of local extrema. To improve the robustness, recent variants of the EMD method were proposed, e.g., [174] sifts an ensemble of white noise-added signal for many times; [94] uses a proper smooth function to approximate the noisy signal by least-square-spline-fit. In the spirit of 1D EMD, the mode decomposition problem in higher-dimensional cases has also been extensively studied recently: either based on surface interpolation [123, 124, 135, 136] or based on the decomposition of 1D data slices [96, 175]. However, the application of 2D EMD methods is limited due to the lack of ability to distinguish two wave-like components with similar wave numbers but different wave vectors as illustrated in detail in [182].

For the purpose of designing an alternative tool for the mode decomposition problem with more rigorous analysis and mathematical understanding, Daubechies, Lu and Wu revisited the synchrosqueezed wavelet transform (SSWT) in [44] and proved that the SSWT can accurately extract wave-like components and estimate their instantaneous frequencies from their superposition in [43]. This was the beginning of a systematic study of various synchrosqueezed transforms based on 1D windowed Fourier transforms [156], 1D and 2D wave packet transforms [178, 182], 1D vanishing-moment and minimum-supported spline-wavelet transform [34], 2D generalized curvelet transform [184], 2D monogenic wavelet transform [37], both mathematically and computationally. Although the synchrosqueezing operator is not Lipschitz continuous in mathematics, its robustness against non-stationary Gaussian random noise (colored) with bounded Fourier spectrum is reasonable [32, 155, 183] and can be significantly improved by designing a highly redundant time-frequency dictionary [183]. The synchrosqueezing technique was further improved for better accuracy in [8, 115, 137] considering time-frequency group delay, phase warping, and higher order differential reassignment, respectively.

Another substantial research branch for mode decomposition problems is based on optimization. Following the methodology of sifting modes from the most oscillatory one, Hou and Shi proposed several 1D optimization models using total variations and matching pursuit in [92], using signal sparsity in a data-driven time-frequency dictionary and matching pursuit in [93]. In the spirit of recovering all components at one time, Dragomiretskiy and Zosso proposed 1D and 2D variational mode decomposition methods based on the smoothness of each component after frequency shifting in [58, 59]; Li and Demanet studied a nonlinear least-squares optimization model based on the smoothness of amplitude and frequency functions in [116].

Other than those methods in the above research lines, many other creative methods for mode decompositions have been proposed, [35, 36, 77, 78] to name a few.

In spite of considerable successes of modeling signals in the form of (1.3), a superposition of a few wave-like components is too limited to describe general oscillatory patterns. In some cases,

a decomposition in the form of (1.3) would lose important physical information as discussed in [170, 176]. To be more general, it is natural to consider a mode decomposition of the form

$$f(x) = \sum_{k=1}^{K} \alpha_k(x) s_k(2\pi N_k \phi_k(x)) + T(x) + e(x), \tag{1.4}$$

where $\{s_k(x)\}_{1\le k\le K}$ are 2π-periodic wave shape functions. Considering the Fourier series of $s_k(x)$, the form of (1.4) essentially becomes the form of (1.3) with a superposition of infinite terms, i.e.,

$$f(x) = \sum_{k=1}^{K} \sum_{n=-\infty}^{\infty} \widehat{s_k}(n) \alpha_k(x) e^{2\pi i n N_k \phi_k(x)} + T(x) + e(x). \tag{1.5}$$

One could combine terms with similar time-frequency geometry in the form of (1.3) to obtain a more efficient and more meaningful decomposition in the form of (1.4). This is referred to as the general mode decomposition in this thesis. This problem is first studied by Wu in [170] and is related to the intrawave modulation discussed in [98].

A straightforward question would be whether the existing methods for mode decompositions can extract general modes $\{\alpha_k(t)s_k(2\pi N_k\phi_k(x))\}$, identify wave shape functions $\{s_k(x)\}$ and estimate instantaneous frequencies $\{N_k\phi_k'(x)\}$. It was conjectured that the EMD methods could decompose signals into general components of the form of (1.4) instead of the form of (1.3) based on some case study. However, this advantage is frangible and worth more effort to understand the EMD methods on general mode decompositions. Articles in [34, 170] show that the synchrosqueezed wavelet transform together with a functional least-square method can be used to solve the general mode decomposition problem for a superposition of general modes with wave shape functions $s_k(x)$ sufficiently close to the exponential function e^{ix}, i.e., a few terms of the Fourier series of $s_k(x)$ are sufficient to approximate $s_k(x)$. However, this class of band-limited wave shape functions in [170] is too restrictive in some situations, e.g. ECG signals. This motivates the work in [178] that applies the synchrosqueezed wave packet transform and a novel diffeomorphism-based spectral analysis method to solve the general mode decomposition problem for a wide range of wave shape functions.

1.3 Contributions

In the first part of this thesis, we focus on designing and analyzing synchrosqueezed transforms (SST) to solve a few open problems for mode decompositions. I'm the main contributor of the theory and numerical tools in this part. These SSTs enjoy simple formulas that allow fast algorithms for forward and inverse transforms. This is especially important to many real problems in high dimensions. The smooth trend function of the oscillatory data becomes negligible if the linear time-frequency transform before synchrosqueezing has enough vanishing moment. This advantage waives the trouble

of estimating the trend function. SSTs inherit the localness of the linear time-frequency transform before synchrosqueezing. They are able to detect local events, e.g. sudden changes in data. It is also flexible to choose different linear transforms according to different data characteristics, e.g., ECG signals with spikes [170, 178], waveforms even with discontinuity [178], wave propagation with defects and sharp boundaries [126, 180, 184]. Unlike many mode decomposition methods that are algorithmic, SSTs are visually informative in the sense that they allow flexible human interaction in spectral analysis. This could inspire new thoughts for better understanding of oscillatory signals.

First, we develop new theory and algorithms for 1D general mode decompositions. This is the first constructive and effective method that is suitable for a wide range of general modes. We introduce a 1D synchrosqueezed wave packet transform in Section 2.1. This transform consists of a wave packet transform of a certain geometric scaling and a reallocation technique for sharpening time-frequency representations. It is proved that this transform is able to estimate instantaneous information from a superposition of general modes. It has a better capacity of distinguishing high frequency wave-like components than the synchrosqueezed wavelet transform. Based on diffeomorphisms through smooth phase functions, a new spectral analysis method for estimating wave shape functions is proposed in Chapter 5. These two analysis tools lead to a framework for general mode decompositions if these modes satisfy certain separation conditions.

Second, we introduce multi-dimensional synchrosqueezed wave packet transforms as the first method for "truly" multi-dimensional mode decomposition problems with rigorous mathematical analysis in Section 2.2. Existing methods cannot separate two modes if they have similar wave numbers but different wave vectors. We introduce a class of superpositions of several wave-like components satisfying certain separation conditions and prove that the multi-dimensional synchrosqueezed wave packet transform identifies each component and estimates its local wavevector accurately.

Third, 2D synchrosqueezed curvelet transform is designed in Section 2.3 as an ideal tool for 2D mode decompositions of wavefronts or banded wave-like components. The synchrosqueezed curvelet transform is a combination of a generalized curvelet transform with application-dependent geometric scaling parameters and a synchrosqueezing process for a sharpened phase space representation. In the case of a superposition of banded wave-like components with well-separated wave-vectors, we show that the synchrosqueezed curvelet transform is able to separate each component and estimate its local wave-vector.

Fourth, we study several fundamental robustness properties of synchrosqueezed transforms in Chapter 3. Although the mathematical analysis of these newly developed transforms is well developed, there is relatively little study on their robustness against noise. Assuming a generalized Gaussian random noise, we estimate the probability of a good instantaneous frequency or local wave vector estimate given by these transforms. The probability analysis shows that their robustness is determined by the geometric scaling parameters and can be improved by tuning their multiscale geometry in the frequency domain. This dependence is demonstrated by numerical experiments as

well. Finally, we provide new insights and numerical implementations for better and more robust estimates.

Finally, discrete analogues with efficient implementations of these synchrosqueezed transforms are proposed in Chapter 4. A software package SynLab for fast synchrosqueezed transforms has been published online. It is available at https://github.com/HaizhaoYang/SynLab. We apply this package to many real problems in seismic data/image processing, atomic crystal image analysis in materials science and canvas analysis in art investigation. They obtain better results than existing state-of-the-art algorithms. These examples will be introduced in Chapter 6 after analyzing major properties of synchrosqueezed transforms.

Chapter 2

Theory of Synchrosqueezed Transforms

In this chapter, we present the theory of multi-dimensional continuous synchrosqueezed wave packet transforms and 2D synchrosqueezed curvelet transforms to analyze wave-like components from their superposition. This is joint work with Lexing Ying in [178, 182, 184]. Since the smooth trend function becomes insignificant after a time-frequency transform with sufficient vanishing moments or it can be estimated and eliminated before synchrosqueezed transform using the methods in [35], we consider a superposition of wave-like components without a smooth trend in this chapter.

Recall that a signal to be analyzed is a complex signal

$$f(x) = \sum_{k=1}^{K} \alpha_k(x) e^{2\pi i N_k \phi_k(x)}, \tag{2.1}$$

where $\alpha_k(x)$ is the instantaneous amplitude, $2\pi N_k \phi_k(x)$ is the instantaneous phase and $N_k \phi_k'(x)$ is the instantaneous frequency. One wishes to decompose the signal $f(x)$ to obtain each component $\alpha_k(x) e^{2\pi i N_k \phi_k(x)}$ and its corresponding instantaneous properties. The synchrosqueezed transform provides a time-frequency representation that concentrates non-zero energy around each instantaneous frequency $N_k \phi_k'(x)$ or local wave vector $N_k \nabla \phi_k(x)$. The time-frequency geometry of each component is a direct result of this localized representation. Each wave-like component can be recovered by an inverse transform on the information restricted in each component in the time-frequency representation. Hence, the theory of synchrosqueezed transforms focuses on the accuracy of energy concentration on $N_k \phi_k'(x)$ or $N_k \nabla \phi_k(x)$.

To motivate the synchrosqueezed transform, we will start with the 1D case.

2.1 1D Synchrosqueezed Wave Packet Transform

2.1.1 Motivation

The synchrosqueezed wavelet transform (SSWT) was introduced in [44] to process auditory signals for a sharpened time-frequency representation by reallocating wavelet coefficients. This idea is inspired by the observation that the local oscillation of the phase of the wavelet coefficients is able to reveal the instantaneous frequency of a wave-like component.

Suppose $\phi(x)$ is an appropriately chosen analytic wavelet (e.g., a Morlet wavelet), then the continuous wavelet transform of a signal $f(x)$ is

$$W_f(a,b) = \int f(x) a^{1/2} \overline{\phi((x-b)a)}\,\mathrm{d}x,$$

where $a^{1/2}\phi((x-b)a)$ for $a \in (0,+\infty)$ and $b \in \mathbb{R}$ is a wavelet that has an essential support $[b-\frac{1}{a}, b+\frac{1}{a}]$ in space and $[\frac{a}{2}, 2a]$ in frequency. We refer to [42, 128] for a detailed introduction to wavelets.

For a purely harmonic analytic signal $f(x) = \alpha e^{2\pi i N x}$, its wavelet transform is

$$\begin{aligned} W_f(a,b) &= \int \alpha e^{2\pi i N x} \sqrt{a}\overline{\phi((x-b)a)}\,\mathrm{d}x \\ &= \frac{\alpha}{\sqrt{a}} \int e^{2\pi i N(\frac{y}{a}+b)} \overline{\phi(y)}\,\mathrm{d}y \\ &= \frac{\alpha}{\sqrt{a}} e^{2\pi i N b} \overline{\widehat{\phi}(\frac{N}{a})}. \end{aligned}$$

For fixed a, notice that $W_f(a,b)$ is a purely harmonic analytic signal with an amplitude $\frac{\alpha}{\sqrt{a}}\overline{\widehat{\phi}(\frac{N}{a})}$ and frequency N. Hence, the local oscillation of $W_f(a,b)$ recovers the instantaneous frequency of $f(x)$ in the sense that

$$\frac{\partial_b W_f(a,b)}{2\pi i W_f(a,b)} = N.$$

Daubechies, Lu and Wu revisited the idea of the SSWT and proved that $\frac{\partial_b W_f(a,b)}{2\pi i W_f(a,b)}$ can act as an instantaneous frequency information function that reveals the instantaneous frequencies $\phi_k'(x)$ in a class of superpositions in (2.1) if these wave-like components satisfy a certain well-separation condition in the time-frequency plane [43]. The well-separation condition is requiring that the gap between adjacent instantaneous frequencies $N_k\phi_k'(x)$ and $N_{k+1}\phi_{k+1}'(x)$ is sufficiently large such that $W_{f_k}(a,b)$ and $\partial_b W_{f_k}(a,b)$ have essential supports well separated from others. Hence, there is only one dominant component $W_{f_k}(a,b)$ in $W_f(a,b)$ and one dominant partial derivative $\partial_b W_{f_k}(a,b)$ in $\partial_b W_f(a,b)$. This leads to the following approximation

$$v_f(a,b) := \frac{\partial_b W_f(a,b)}{2\pi i W_f(a,b)} \approx \frac{\partial_b W_{f_k}(a,b)}{2\pi i W_{f_k}(a,b)} \approx N_k \phi_k'(b),$$

when (a, b) is in the essential support of $W_{f_k}(a, b)$. Hence, by summing the spectral energy of $W_f(a, b)$ according to $v_f(a, b)$, we can obtain a sharpened spectral energy in the time-frequency plane

$$T_f(v, b) = \int_{\mathbb{R}} |W_f(a, b)|^2 \delta(\mathfrak{Re} v_f(a, b) - v)\, \mathrm{d}a$$

with essential supports concentrating around $N_k \phi_k'(b)$ independent of v for some k. Here $\mathfrak{Re} v_f(a, b)$ means the real part of a complex number $v_f(a, b)$.

By the stationary phase approximation and the smoothness of $\alpha_k(x)$ and $\phi_k(x)$, $W_{f_k}(a, b)$ and $\partial_b W_{f_k}(a, b)$ have an essential support in $[\frac{N_k \phi_k'(x)}{2}, 2N_k \phi_k'(x)]$. To satisfy the well-separation condition, it is required that

$$2N_k \phi_k'(x) \lesssim N_{k+1} \phi_{k+1}'(x),$$

i.e., instantaneous frequencies should be exponentially increasing in k. This requirement limits the application of the SSWT to analyze superpositions of wave-like components with close instantaneous frequencies. This motivates the design of synchrosqueezed wave packet transforms (SSWPT) with a weaker well-separation condition.

2.1.2 Definition of 1D SSWPT

We briefly introduce the 1D synchrosqueezed wave packet transform (SSWPT) in this section and will analyze it in the next section. Wave packets here are built on an appropriately chosen mother wave packet defined below.

Definition 2.1.1. *A mother wave packet $w(x) \in C^m(\mathbb{R})$ is of type (ϵ, m) for some $\epsilon > 0$, and some non-negative integer m, if $\widehat{w}(\xi)$ is a real-valued function with an essential support in the ball $B_1(0)$ centered at the origin with a radius 1 satisfying that:*

$$|\widehat{w}(\xi)| \leq \frac{\epsilon}{(1 + |\xi|)^m},$$

for $|\xi| > 1$.

Since $w \in C^m(\mathbb{R})$, the above decaying requirement is easy to satisfy. Actually, we can further assume $\widehat{w}(\xi)$ is essentially supported in a ball $B_d(0)$ with a support parameter $d \in (0, 1]$ for signals with close instantaneous frequencies. However, d is just a constant in later asymptotic analysis. Hence, we omit its discussion and consider it as 1 in the analysis but implement it in our numerical tool. We can use this mother wave packet $w(x)$ to define a family of wave packets through scaling, modulation, and translation, controlled by a geometric parameter s.

Definition 2.1.2. *Given the mother wave packet $w(x)$ of type (ϵ, m) and a parameter $s \in (1/2, 1)$,*

the family of wave packets $\{w_{ab}(x) : |a| \geq 1, b \in \mathbb{R}\}$ *is defined as*

$$w_{ab}(x) = |a|^{s/2} w\left(|a|^s (x-b)\right) e^{2\pi i (x-b) a},$$

or equivalently, in the Fourier domain as

$$\widehat{w_{ab}}(\xi) = |a|^{-s/2} e^{-2\pi i b \xi} \widehat{w}\left(|a|^{-s}(\xi - a)\right).$$

These definitions allow us to construct a family of compactly supported wave packets, which will be useful in practice. It is clear from the definition that the Fourier transform $\widehat{w_{ab}}(\xi)$ is essentially supported in $(a - |a|^s, a + |a|^s)$. On the other hand, $w_{ab}(x)$ is centered in space at b with an essential support of width $O(|a|^{-s})$. $\{w_{ab}(x) : |a| \geq 1, b \in \mathbb{R}\}$ are all appropriately scaled to have the same L^2 norm with the mother wave packet $w(x)$.

The instantaneous frequency of the low frequency part of a signal is not well defined as discussed in [139]. For this reason, it is enough to consider wave packets with $|a| \geq 1$. High frequency components can be identified and extracted independently of the low frequency part so that the low frequency part can be recovered by removing high frequency components.

Notice that if s were equal to 1, these functions would be qualitatively similar to the standard wavelets. On the other hand, if s were equal to $1/2$, we would obtain the wave atoms defined in [49]. But $s \in (1/2, 1)$ is essential as we shall see in the main theorems later. The lower bound $s > 1/2$ makes the support of the wave packets sufficiently small for instantaneous frequency estimation, while the upper bound $s < 1$ allows better resolution to distinguish close instantaneous frequencies than wavelets, which is the purpose for proposing the SSWPT. See Figure 2.1 for an illustration of the comparison of wavelets and wave packets in the frequency domain.

Definition 2.1.3. *The 1D wave packet transform of a function* $f \in L^\infty(\mathbb{R})$ *is a function*

$$W_f(a,b) = \langle f, w_{ab} \rangle = \int_{\mathbb{R}} f(x) \overline{w_{ab}(x)} dx$$

for $|a| \geq 1, b \in \mathbb{R}$.

Definition 2.1.4. *Instantaneous frequency information function:*

Let $f \in L^\infty(\mathbb{R})$. *The instantaneous frequency estimation function* $v_f(a,b)$ *for* $|a| \geq 1$ *and* $b \in \mathbb{R}$ *of* f *is defined by*

$$v_f(a,b) = \begin{cases} \frac{\partial_b W_f(a,b)}{2\pi i W_f(a,b)}, & \text{for } |W_f(a,b)| > 0; \\ \infty, & \text{otherwise.} \end{cases}$$

It will be proved that, for a class of wave-like functions $f(x) = \alpha(x) e^{2\pi i N \phi(x)}$, $v_f(a,b)$ precisely approximates $N\phi'(b)$ independently of a as long as $|W_f(a,b)|$ is large enough. Hence, if we squeeze the coefficients $W_f(a,b)$ together based upon the same instantaneous frequency information function

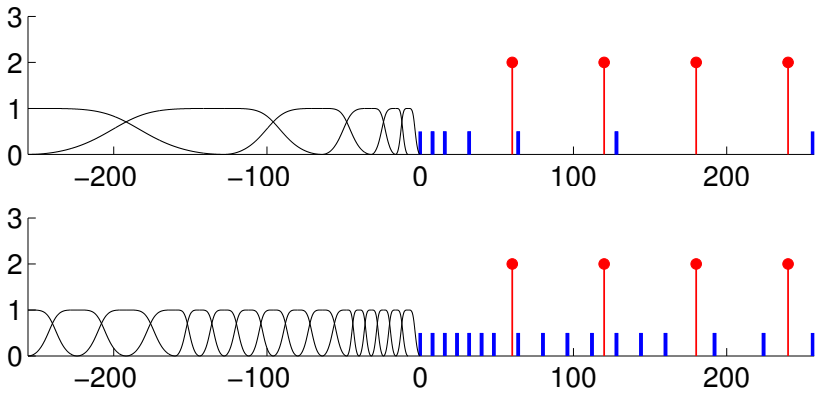


Figure 2.1: Top: Wavelet (wave packet when $s = 1$) tiling and bump functions in the high frequency Fourier domain. Each short budding in the positive part indicates the center of one bump function, while the bump functions in the negative part are plotted. Red dots denote the support of the Fourier transform of $\sum_{n=1}^{4} e^{2\pi i n N t}$, where $N = 60$. The well-separation condition for the SSWT is not satisfied, because the wavelet transform of these wave-like components are overlapping in the time-frequency domain. Bottom: Wave packet tiling and bump functions with $s = \frac{1}{2}$. The well-separation condition holds.

$v_f(a, b)$, then we would obtain a sharpened time-frequency representation of $f(x)$. This motivates the definition of the synchrosqueezed energy distribution as follows.

Definition 2.1.5. *Given $f \in L^\infty$, the synchrosqueezed energy distribution $T_f(v, b)$ is defined by*

$$T_f(v, b) = \int_{\mathbb{R}\setminus(-1,1)} |W_f(a, b)|^2 \delta\left(\mathfrak{Re} v_f(a, b) - v\right) da$$

for $v, b \in \mathbb{R}$.

For a multi-component signal $f(x)$, the synchrosqueezed energy of each component will concentrate around its corresponding instantaneous frequency. Hence, the SSWPT can provide information about their instantaneous frequencies.

2.1.3 Analysis

In this section, we provide rigorous analysis of the 1D SSWPT generated from mother wave packets of type (ϵ, m) to analyze a noiseless superposition of wave-like components.

Definition 2.1.6. *A function $f(x) = \alpha(x)e^{2\pi i N \phi(x)}$ is an intrinsic mode type function (IMT) of*

type (M, N)*, if* $\alpha(x)$ *and* $\phi(x)$ *satisfy the conditions below.*

$$\alpha(x) \in C^\infty, \quad |\alpha'(x)| \le M, \quad 1/M \le \alpha(x) \le M$$
$$\phi(x) \in C^\infty, \quad 1/M \le |\phi'(x)| \le M, \quad |\phi''(x)| \le M.$$

Definition 2.1.7. *A function* $f(x)$ *is a well-separated superposition of type* (M, N, K, s)*, if*

$$f(x) = \sum_{k=1}^{K} f_k(x),$$

where each $f_k(x) = \alpha_k(x)e^{2\pi i N_k \phi_k(x)}$ *is an IMT of type* (M, N_k) *such that* $N_k \ge N$ *and the phase functions satisfy the separation condition: for any pair* (a, b)*, there exists at most one* k *such that*

$$|a|^{-s}|a - N_k\phi_k'(b)| < 1.$$

We denote by $F(M, N, K, s)$ *the set of all such functions.*

Theorem 2.1.8 below shows that the SSWPT is able to estimate instantaneous frequencies $\{N_k\phi_k'(x)\}_{k=1}^K$ of well-separated superposition of IMTs accurately. In what follows, when we write $O(\cdot)$, $\lesssim$, or $\gtrsim$, the implicit constants may depend on M, m and K.

Theorem 2.1.8. *Suppose the mother wave packet is of type* (ϵ, m)*, for any fixed* $\epsilon \in (0, 1)$ *and any fixed integer* $m \ge 0$*. For a function* $f(x)$*, we define*

$$R_\epsilon = \{(a, b) : |W_f(a, b)| \ge |a|^{-s/2}\sqrt{\epsilon}\},$$

$$S_\epsilon = \{(a, b) : |W_f(a, b)| \ge \sqrt{\epsilon}\},$$

and

$$Z_k = \{(a, b) : |a - N_k\phi_k'(b)| \le |a|^s\}$$

for $1 \le k \le K$*. For fixed* M*,* m *and* K*, there exists a constant* $N_0(M, m, K, s, \epsilon) \simeq \max\left\{\epsilon^{\frac{-1}{2s-1}}, \epsilon^{\frac{-1}{1-s}}\right\}$ *such that for any* $N > N_0(M, m, K, s, \epsilon)$ *and* $f(x) \in F(M, N, K, s)$ *the following statements hold.*

(i) $\{Z_k : 1 \le k \le K\}$ *are disjoint and* $S_\epsilon \subset R_\epsilon \subset \bigcup_{1\le k\le K} Z_k$*;*

(ii) For any $(a, b) \in R_\epsilon \cap Z_k$*,*

$$\frac{|v_f(a, b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \lesssim \sqrt{\epsilon};$$

(iii) For any $(a, b) \in S_\epsilon \cap Z_k$*,*

$$\frac{|v_f(a, b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \lesssim \frac{\sqrt{\epsilon}}{N_k^{s/2}}.$$

The proof of Theorem 2.1.8 relies on two lemmas as follows to estimate the asymptotic behavior of $W_f(a,b)$ and $\partial_b W_f(a,b)$ as N going to infinity.

Lemma 2.1.9. *Suppose* $\Omega_a = \{k : a \in [\frac{N_k}{2M}, 2MN_k]\}$. *Under the assumption of Theorem 2.1.8, we have*

$$W_f(a,b) = |a|^{-s/2}\left(\sum_{k\in\Omega_a}\alpha_k(b)e^{2\pi i N_k\phi_k(b)}\widehat{w}\left(\left(a-N_k\phi_k'(b)\right)|a|^{-s}\right)+O(\epsilon)\right),$$

when $N > N_0(M,m,K,s,\epsilon) \simeq \max\left\{\epsilon^{\frac{-1}{2s-1}}, \epsilon^{\frac{-1}{1-s}}\right\}$.

Proof. Without loss of generality, we can simply assume $N_k = N$ for all k and only prove the case for $a > 1$. Because $w(x)$ decays rapidly, the wave packet transform $W_f(a,b)$ is well defined. By the change of variables, we have

$$\begin{aligned} W_f(a,b) &= \int_{\mathbb{R}}\sum_{k=1}^{K}\alpha_k(x)e^{2\pi i N\phi_k(x)}|a|^{s/2}w(|a|^s(x-b))e^{-2\pi i(x-b)a}dx \\ &= |a|^{-s/2}\sum_{k=1}^{K}\int_{\mathbb{R}}\alpha_k(|a|^{-s}x+b)w(x)e^{2\pi i(N\phi_k(|a|^{-s}x+b)-|a|^{1-s}x)}dx. \end{aligned}$$

Let us estimate $I_k = \int_{\mathbb{R}}\alpha_k(|a|^{-s}x+b)w(x)e^{2\pi i(N\phi_k(|a|^{-s}x+b)-|a|^{1-s}x)}dx$. Let

$$h(x) = \alpha_k(|a|^{-s}x+b)w(x)$$

and

$$g(x) = 2\pi(N\phi_k(|a|^{-s}x+b) - |a|^{1-s}x),$$

then

$$I_k = \int_{\mathbb{R}} h(x)e^{ig(x)}dx,$$

and

$$g'(x) = 2\pi|a|^{-s}(N\phi_k'(|a|^{-s}x+b) - a).$$

If $a < \frac{N}{2M}$, then $|g'(x)| \gtrsim |a|^{-s}N \gtrsim N^{1-s}$. If $a > 2MN$, then $|g'(x)| \gtrsim |a|^{1-s} \gtrsim (N)^{1-s}$. So, if $a \notin [\frac{N}{2M}, 2MN]$, then $|g'(x)| \gtrsim (N)^{1-s}$. For real smooth functions $g(x)$, we define the differential operator

$$L = \frac{1}{i}\frac{\partial_x}{g'}.$$

Because $h(x)$ decays sufficiently fast at infinity, we perform integration by parts r times to get

$$\int_{\mathbb{R}} he^{ig}dx = \int_{\mathbb{R}} h(L^r e^{ig})dx = \int_{\mathbb{R}}\left((L^*)^r h\right)e^{ig}dx,$$

where L^* is the adjoint of L. A few algebraic calculation shows that L^* contributes a factor of order

$\frac{1}{|g'|} \lesssim \frac{1}{(N)^{1-s}}$ if $a \notin [\frac{N}{2M}, 2MN]$, and we therefore have

$$|I_k| = \left| \int_{\mathbb{R}} e^{ig} \left((L^*)^r h \right) dx \right| \lesssim (N)^{-(1-s)r} \lesssim \epsilon.$$

Since $s < 1$, if $N \gtrsim \epsilon^{\frac{-1}{(1-s)r}}$, then

$$|a|^{-s/2} \sum_{k \notin \Omega_a} I_k \lesssim |a|^{-s/2} \sum_{k \notin \Omega_a} O(\epsilon) \lesssim |a|^{-s/2} O(\epsilon). \tag{2.2}$$

Now let us estimate I_k when $a \in [\frac{N}{2M}, 2MN]$. Recall that

$$I_k = \int_{\mathbb{R}} \alpha_k(|a|^{-s}x + b) w(x) e^{2\pi i (N\phi_k(|a|^{-s}x+b) - |a|^{1-s}x)} dx.$$

By Taylor expansion,

$$\alpha_k(|a|^{-s}x + b) = \alpha_k(b) + \alpha_k'(b^*)|a|^{-s}x$$

and

$$\phi_k(|a|^{-s}x + b) = \phi_k(b) + \phi_k'(b)|a|^{-s}x + \frac{1}{2}\phi_k''(b^{**})|a|^{-2s}x^2$$

for some b^* and b^{**}. Notice that, if $N \gtrsim \epsilon^{-1/s}$, then

$$\begin{aligned} & |I_k - \alpha_k(b) \int_{\mathbb{R}} w(x) e^{2\pi i (N\phi_k(|a|^{-s}x+b) - |a|^{1-s}x)} dx| \\ \lesssim \; & \alpha_k'(b^*)|a|^{-s} \int_{\mathbb{R}} |x||w(x)| dx \\ \lesssim \; & O(\epsilon). \end{aligned}$$

This implies that

$$I_k = \left(\alpha_k(b) \int_{\mathbb{R}} w(x) e^{2\pi i (N\phi_k(|a|^{-s}x+b) - |a|^{1-s}x)} dx + O(\epsilon) \right)$$

for $a \in [\frac{N}{2M}, 2MN]$ and $N \gtrsim \epsilon^{-1/s}$. Since $|e^{ix} - 1| \leq |x|$, if $N \gtrsim \epsilon^{-1/(2s-1)}$, then we have

$$\begin{aligned} & |I_k - \alpha_k(b) \int_{\mathbb{R}} w(x) e^{2\pi i (N\phi_k(b) + N\phi_k'(b)|a|^{-s}x - |a|^{1-s}x)} dx| \\ \lesssim \; & \left(O(\epsilon) + \left| \alpha_k(b) \int_{\mathbb{R}} w(x) e^{2\pi i (N\phi_k(b) + N\phi_k'(b)|a|^{-s}x - |a|^{1-s}x)} \left(e^{2\pi i N \frac{1}{2}\phi_k''(b^{**})|a|^{-2s}x^2} - 1 \right) dx \right| \right) \\ \lesssim \; & \left(O(\epsilon) + N|a|^{-2s} \int_{\mathbb{R}} x^2 |w(x)| dx \right) \\ \lesssim \; & O(\epsilon). \end{aligned}$$

Hence, it holds that

$$I_k = \left(\alpha_k(b)e^{2\pi iN\phi_k(b)}\widehat{w}\left((a-N\phi_k'(b))|a|^{-s}\right)+O(\epsilon)\right), \tag{2.3}$$

if $a\in[\frac{N}{2M},2MN]$ and $N\gtrsim\max\{\epsilon^{-1/s},\epsilon^{-1/(2s-1)}\}=\epsilon^{-1/(2s-1)}$.

In sum, by (2.2) and (2.3), we arrive at

$$\begin{aligned} W_f(a,b) &= |a|^{-s/2}\left(\sum_{k\in\Omega_a} I_k+\sum_{k\notin\Omega_a} I_k\right)\\ &= |a|^{-s/2}\left(\sum_{k\in\Omega_a}\alpha_k(b)e^{2\pi iN\phi_k(b)}\widehat{w}\left((a-N\phi_k'(b))\,|a|^{-s}\right)+O(\epsilon)\right), \end{aligned}$$

if $N\gtrsim\max\{\epsilon^{\frac{-1}{(1-s)r}},\epsilon^{\frac{-1}{2s-1}}\}$.

Similar argument can prove the above conclusion for $a<-1$ and it is simple to generalize it for different N_k to complete the proof. □

The next lemma is to estimate $\partial_b W_f(a,b)$ when $\Omega_a=\{k:a\in[\frac{N_k}{2M},2MN_k]\}$ is not empty, i.e., when $W_f(a,b)$ is relevant.

Lemma 2.1.10. *Suppose $\Omega_a=\{k:a\in[\frac{N_k}{2M},2MN_k]\}$ is not empty. Under the assumption of Theorem 2.1.8, we have*

$$\begin{aligned} &\partial_b W_f(a,b)\\ =\;& |a|^{-s/2}\left(\sum_{k\in\Omega_a}2\pi iN_k\alpha_k(b)\phi_k'(b)e^{2\pi iN_k\phi_k(b)}\widehat{w}\left((a-N_k\phi_k'(b))\,|a|^{-s}\right)+aO\left(\epsilon\right)\right), \end{aligned}$$

when $N>N_0\left(M,m,K,s,\epsilon\right)\simeq\max\left\{\epsilon^{\frac{-1}{2s-1}},\epsilon^{\frac{-1}{1-s}}\right\}$.

Proof. Similar to the proof of Lemma 2.1.9, we can assume $N_k=N$ for all k and only need to prove the case when $a>1$. By the definition of the wave packet transform, we have

$$\begin{aligned} \partial_b W_f(a,b) &= \sum_{k=1}^{K}2\pi i|a|^{1+s/2}\int_{\mathbb{R}}\alpha_k(x)e^{2\pi iN\phi_k(x)}w(|a|^s(x-b))e^{-2\pi i(x-b)a}dx\\ &\quad-\sum_{k=1}^{K}|a|^{3s/2}\int_{\mathbb{R}}\alpha_k(x)e^{2\pi iN\phi_k(x)}w'(|a|^s(x-b))e^{-2\pi i(x-b)a}dx. \end{aligned}$$

Denote the first term by T_1 and the second term by T_2. By a similar discussion in the proof of

Lemma 2.1.9, we have the following asymptotic estimates when N is sufficiently large.

$$\begin{aligned}
T_2 &= -|a|^{s/2}\sum_{k=1}^{K}\int_{\mathbb{R}}\alpha_k(|a|^{-s}x+b)w'(x)e^{2\pi i(N\phi_k(|a|^{-s}x+b)-|a|^{1-s}x)}dx \\
&= -|a|^{s/2}\sum_{k\in\Omega_a}\int_{\mathbb{R}}\alpha_k(|a|^{-s}x+b)w'(x)e^{2\pi i(N\phi_k(|a|^{-s}x+b)-|a|^{1-s}x)}dx + |a|^{s/2}O(\epsilon) \\
&= |a|^{s/2}\sum_{k\in\Omega_a}\int_{\mathbb{R}} w(x)\alpha_k(|a|^{-s}x+b)e^{2\pi i(N\phi_k(|a|^{-s}x+b)-|a|^{1-s}x)} \\
&\quad \left(2\pi iN\phi_k'(|a|^{-s}x+b)|a|^{-s}-2\pi i|a|^{1-s}\right)dx + |a|^{-s/2}\sum_{k\in\Omega_a}\int_{\mathbb{R}} w(x) \\
&\quad \alpha_k'(|a|^{-s}x+b)e^{2\pi i(N\phi_k(|a|^{-s}x+b)-|a|^{1-s})}dx + |a|^{s/2}O(\epsilon) \\
&= |a|^{-s/2}\sum_{k\in\Omega_a} 2\pi iN\int_{\mathbb{R}}\phi_k'(|a|^{-s}x+b)\alpha_k(|a|^{-s}x+b)w(x)e^{2\pi i(N\phi_k(|a|^{-s}x+b)-|a|^{1-s}x)}dx \\
&\quad -|a|^{1-s/2}\sum_{k\in\Omega_a} 2\pi i\int_{\mathbb{R}} w(x)\alpha_k(|a|^{-s}x+b)e^{2\pi i(N\phi_k(|a|^{-s}x+b)-|a|^{1-s}x)}dx \\
&\quad +|a|^{-s/2}O(1)+|a|^{s/2}O(\epsilon) \\
&= |a|^{-s/2}\sum_{k\in\Omega_a} 2\pi iN\left(\phi_k'(b)\alpha_k(b)e^{2\pi iN\phi_k(b)}\widehat{w}(|a|^{-s}(a-N\phi_k'(b)))+O(\epsilon)\right) \\
&\quad -|a|^{1+s/2}\sum_{k\in\Omega_a} 2\pi i\int_{\mathbb{R}}\alpha_k(x)w(|a|^s(x-b))e^{2\pi i(N\phi_k(x)-(x-b)a)}dx \\
&\quad +|a|^{-s/2}O(1)+|a|^{s/2}O(\epsilon),
\end{aligned}$$

if $N \gtrsim \max\{\epsilon^{\frac{-1}{(1-s)r}}, \epsilon^{\frac{-1}{2s-1}}\}$. The third equality holds by integration by parts and the last equality holds by changing variables. Notice that

$$\begin{aligned}
T_1 &= |a|^{1+s/2}\sum_{k\in\Omega_a} 2\pi i\int_{\mathbb{R}}\alpha_k(x)w(|a|^s(x-b))e^{2\pi i(N\phi_k(x)-(x-b)a)}dx \\
&\quad +\sum_{k\notin\Omega_a} 2\pi i|a|^{1-s/2}\int_{\mathbb{R}}\alpha_k(|a|^{-s}x+b)w(x)e^{2\pi i(N\phi_k(|a|^{-s}x+b)-|a|^{1-s}x)}dx \\
&= |a|^{1+s/2}\sum_{k\in\Omega_a} 2\pi i\int_{\mathbb{R}}\alpha_k(x)w(|a|^s(x-b))e^{2\pi i(N\phi_k(x)-(x-b)a)}dx + |a|^{1-s/2}O(\epsilon),
\end{aligned}$$

if $N \gtrsim \epsilon^{\frac{-1}{(1-s)r}}$ for any $r >= 1$. Hence $T_1 + T_2$ results in

$$\begin{aligned} & \partial_b W_f(a,b) \\ = \; & |a|^{-s/2} \sum_{k\in\Omega_a} 2\pi i N \left(\phi_k'(b)\alpha_k(b) e^{2\pi i N \phi_k(b)} \widehat{w}(|a|^{-s}(a - N\phi_k'(b))) + O(\epsilon) \right) \\ & + |a|^{-s/2} O(1) + |a|^{s/2} O(\epsilon) + |a|^{1-s/2} O(\epsilon) \\ = \; & |a|^{-s/2} \left(\sum_{k\in\Omega_a} 2\pi i N \alpha_k(b)\phi_k'(b) e^{2\pi i N\phi_k(b)} \widehat{w}\left((a - N\phi_k'(b))\,|a|^{-s} \right) + |a| O(\epsilon) \right), \end{aligned}$$

if N is sufficiently large. So, the Lemma 2.1.10 is proved. □

We are now ready to prove Theorem 2.1.8 with Lemma 2.1.9 and Lemma 2.1.10.

Proof. Let us first consider (i). The well-separation condition implies that $\{Z_k : 1 \le k \le K\}$ are disjoint. Let (a,b) be a point in R_ϵ, then $|W_f(a,b)| \ge |a|^{-s/2}\sqrt{\epsilon}$, which means that Ω_a is not empty and $\exists k \in \Omega_a$ such that $\widehat{w}((a - N_k\phi_k'(b))|a|^{-s}) \neq 0$. Because the support of $\widehat{w}(\xi)$ is $(-1,1)$, we know $|a - N_k\phi_k'(b)| \le |a|^s$, i.e., $(a,b) \in Z_k$. Hence, $R_\epsilon \subset \bigcup_{1\le k\le K} \bigcup_{\neq 0} Z_k$.

To show (ii), let us recall that $v_f(a,b)$ is defined as

$$v_f(a,b) = \frac{\partial_b W_f(a,b)}{2\pi i W_f(a,b)},$$

for $W_f(a,b) \neq 0$. If $(a,b) \in R_\epsilon \bigcap Z_k$, then by Lemma 2.1.9

$$\begin{aligned} W_f(a,b) &= |a|^{-s/2} \left(\sum_{k\in\Omega_a} \alpha_k(b) e^{2\pi i N_k \phi_k(b)} \widehat{w}\left((a - N_k\phi_k'(b))\,|a|^{-s} \right) + O(\epsilon) \right) \\ &= |a|^{-s/2} \left(\alpha_k(b) e^{2\pi i N_k\phi_k(b)} \widehat{w}\left((a - N_k\phi_k'(b))\,|a|^{-s} \right) + O(\epsilon) \right), \end{aligned}$$

as the other terms drop out, since $\{Z_k\}$ are disjoint. Similarly, by Lemma 2.1.10

$$\begin{aligned} & \partial_b W_f(a,b) \\ = \; & |a|^{-s/2} \left(2\pi i N_k \alpha_k(b)\phi_k'(b) e^{2\pi i N_k\phi_k(b)} \widehat{w}\left((a - N_k\phi_k'(b))\,|a|^{-s} \right) + |a| O(\epsilon) \right). \end{aligned}$$

Let g denote the term $\alpha_k(b) e^{2\pi i N_k\phi_k(b)} \widehat{w}\left((a - N_k\phi_k'(b))\,|a|^{-s}\right)$, then

$$\begin{aligned} v_f(a,b) &= \frac{N_k\phi_k'(b) g + |a| O(\epsilon)}{g + O(\epsilon)} \\ &= \frac{N_k\phi_k'(b)(g + O(\epsilon))}{g + O(\epsilon)}, \end{aligned}$$

since $a \in [\frac{N_k}{2M}, 2MN_k]$. Because $|W_f(a,b)| \geq |a|^{-s/2}\sqrt{\epsilon}$ for $(a,b) \in R_\epsilon$, then $|g| \gtrsim \sqrt{\epsilon}$. Therefore

$$\frac{|v_f(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \lesssim \left|\frac{O(\epsilon)}{g + O(\epsilon)}\right| \lesssim \sqrt{\epsilon}.$$

Similarly, if $(a,b) \in S_\epsilon \cap Z_k$, then

$$\frac{|v_f(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \lesssim \left|\frac{O(\epsilon)}{g + O(\epsilon)}\right| \lesssim \frac{\sqrt{\epsilon}}{N_k^{s/2}},$$

since $|g| \gtrsim N_k^{s/2}\sqrt{\epsilon}$ for $(a,b) \in S_\epsilon \cap Z_k$. □

Theorem 2.1.8 shows that the instantaneous frequency information function $v_f(a,b)$ can estimate $N_k\phi_k'(x)$ accurately for a class of superpositions of IMTs if their phases are sufficiently steep. This guarantees the well concentration of the synchrosqueezed energy distribution $T_f(v,b)$ around $N_k\phi_k'(x)$. The assumption $s \in (1/2, 1)$ is essential to the proof. The upper bound $s < 1$ enables the wave packets to detect oscillations in different directions. The lower bound $s > 1/2$ ensures that the support of the wave packets is sufficiently small in space so that the second order properties of the phase function (such as the curvature of the wave front) do not affect the synchrosqueezing estimate of the local wavevectors.

In [43], the authors show that, for synchrosqueezed wavelet transform, each intrinsic mode function or component can be reconstructed from the synchrosqueezed coefficients by making use of a reconstruction formula that integrates the continuous wavelet coefficient over the scale parameter with an appropriate weight. They also prove an error bound on the reconstructed intrinsic mode functions. In the current setting, however, we are not aware of a similar reconstruction formula for the wave packet. Therefore, our reconstruction step is based on a Calderon-type reconstruction formula for the wave packets as discussed in the next chapter. A similar approach based on the Calderon reconstruction formula for the wavelets is in fact used in the numerical examples of [43] as it is more robust in the noisy case. However, we have not been able to derive a rigorous error bound for this Calderon-type reconstruction formula for the wave packets at this point.

Since we require N to be sufficiently large in Theorem 2.1.8, a function defined in Definition 2.1.7 is a superposition of highly oscillatory components. In practical applications, a function might also contain a low-frequency component. For such a low-frequency component, the local wavevector is not well-defined as it is impossible to perform a phase-amplitude decomposition as given in Definition 2.1.6 for a low-frequency signal. Thus Theorem 2.1.8 does not apply to such a superposition. However, in practice, we observe that the synchrosqueezing step can still separate the support of different components quite well: typically the support of high frequency components are squeezed into regions Z_k while the support of the low frequency component remains at the low-frequency part of the Fourier domain. Therefore, by applying the reconstruction formula to the coefficients of the low-frequency component, one is still able to identify the low-frequency component quite accurately

even though one cannot estimate its local wavevector.

2.2 Multi-Dimensional Synchrosqueezed Wave Packet Transform

2.2.1 Motivation

An obvious question, which is motivated by applications in geophysics [143, 163], is whether the synchrosqueezing idea can be extended to multi-dimensional images. For example, in seismic imaging analysis, different local wavevectors correspond to different seismic events, which typically link to different geological features. A straightforward attempt would simply combine the multi-dimensional wavelet transform with the synchrosqueezing approach. The resulting synchrosqueezed multi-dimensional wavelet transform would be capable of separating components that have different wavevectors at each location, just as the 1D transform does for 1D signals. However, in many situations this is not enough since a typical multi-dimensional image can have components whose wavevectors have the same magnitude but point in different directions, as shown in Figure 2.2(left). Another simple idea is to synchrosqueeze an appropriately designed directional wavelet transforms (e.g., multi-dimensional Gabor wavelets). However, the dyadic scaling property of these transforms would still give poor resolution to distinguish wave-like components with close local wave vectors. This phenomenon has been shown in the 1D case in last section. In fact, images from many applications related to high-frequency wave propagation have wave-like components with close local wave vectors.

In order to design synchrosqueezed transforms that can separate multi-dimensional wave-like components once they have different local wave vectors, we propose the multi-dimensional synchrosqueezed wave packet transform (SSWPT). Similar to the 1D SSWPT, it combines the synchrosqueezing idea with multi-dimensional wave packets of an appropriate geometric scaling s. The key feature is that these wave packets have finer and, more importantly directional, support in the multi-dimensional Fourier domain, which allows the anisotropic angular separation in the Fourier domain, i.e., distinguishing components oscillating in different directions, as shown in Figure 2.2(right). As we know of, the synchrosqueezed wave packet transform is the first method equipped with this ability so far.

2.2.2 Definition

We will briefly introduce the multi-dimensional synchrosqueezed wave packet transform proposed in this section and analyze it in the next section. Similar to the 1D case, we can also introduce an n-dimensional mother wave packet $w(x) \in C^m(\mathbb{R}^n)$ of type (ϵ, m) such that $\widehat{w}(\xi)$ has an essential

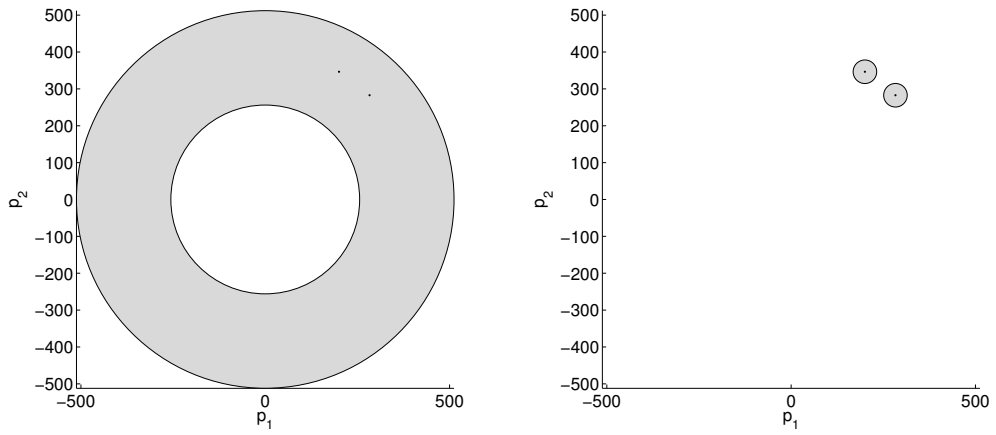


Figure 2.2: Comparison of the resolutions of 2D continuous wavelets (left) and 2D continuous wave packets (right) in the Fourier domain. Consider the superposition of two plane waves $e^{2\pi ip\cdot x}$ and $e^{2\pi iq\cdot x}$ with the same frequency ($|p| = |q|$) but different wavevectors ($p \neq q$). Left: The two dots in each plot show the support of the Fourier transforms of these two plane waves and the gray region stands for the support of a continuous wavelet. Since the isotropic support of each wavelet either covers or misses both points p and q, the wavelet transform is not able to distinguish these two plane waves. Right: Each gray region represents the support of a wave packet. As long as p and q are well separated, they are in the support of two different wave packets. Hence these two plane waves can be distinguished from each other by the wave packet transform.

support in the ball $B_1(0)$ centered at the frequency origin with a radius 1, i.e.,

$$|\widehat{w}(\xi)| \leq \frac{\epsilon}{(1+|\xi|)^m},$$

for $|\xi| > 1$ and some non-negative integer m. A family of n-dimensional wave packets is obtained by isotropic dilation, rotations and translations of the mother wave packet as follows, controlled by a geometric parameter s.

Definition 2.2.1. *Given the mother wave packet $w(x)$ of type (ϵ, m) and the parameter $s \in (1/2, 1)$, the family of wave packets $\{w_{ab}(x) : a, b \in \mathbb{R}^n, |a| \geq 1\}$ are defined as*

$$w_{ab}(x) = |a|^{ns/2} w\left(|a|^s (x-b)\right) e^{2\pi i(x-b)\cdot a},$$

or equivalently in the Fourier domain

$$\widehat{w_{ab}}(\xi) = |a|^{-ns/2} e^{-2\pi ib\cdot\xi} \widehat{w}\left(|a|^{-s}(\xi - a)\right).$$

In this definition, we require $|a| \geq 1$. The reason is that, when $|a| < 1$, the above consideration

regarding the shape of the wave packets is no longer valid. However, since we are mostly concerned with the high frequencies as the signals of interest here are oscillatory, the case $|a| < 1$ is essentially irrelevant.

Some properties can be seen immediately from the definition: the Fourier transform $\widehat{w_{ab}}(\xi)$ is essentially supported in $B_{|a|^s}(a)$, a ball centered at a with a radius $|a|^s$; $w_{ab}(x)$ is centered in space at b with an essential support of width $O(|a|^{-s})$; $\{w_{ab}(x) : a, b \in \mathbb{R}^n, |a| \geq 1\}$ are all appropriately scaled to have the same L^2 norm with the mother wave packet $w(x)$. Notice that if s were equal to $1/2$, we would obtain the wave atoms defined in [49]. If s were equal to 1, these functions would be qualitatively similar to the standard multi-dimensional wavelets. In general, an n-dimensional SSWPT with a smaller s value is better distinguishing two IMTs with close propagating directions. This is the motivation to propose n-dimensional SSWPT rather than directly generalizing the 1D SSWT in [37, 43, 44, 172].

With this family of wave packets, we define the wave packet transform as follows.

Definition 2.2.2. *The wave packet transform of a function $f(x)$ is a function*

$$W_f(a,b) = \langle f, w_{ab} \rangle = \int_{\mathbb{R}^n} f(x)\overline{w_{ab}(x)}dx$$

for $a, b \in \mathbb{R}^n, |a| \geq 1$.

If the Fourier transform $\widehat{f}(\xi)$ vanishes for $|\xi| < 1$, it is easy to check that the L^2 norms of $W_f(a,b)$ and $f(x)$ are equivalent, up to a uniform constant factor, i.e.,

$$\int_{\mathbb{R}^{2n}} |W_f(a,b)|^2 dadb \approx \int_{\mathbb{R}^n} |f(x)|^2 dx. \tag{2.4}$$

Definition 2.2.3. *The local wave vector estimation of a function $f(x)$ at $(a,b) \in \mathbb{R}^{2n}$ is*

$$v_f(a,b) = \begin{cases} \frac{\nabla_b W_f(a,b)}{2\pi i W_f(a,b)}, & \text{for } W_f(a,b) \neq 0; \\ (\infty, \infty), & \text{otherwise.} \end{cases}$$

Given the wave vector estimation $v_f(a,b)$, the synchrosqueezing step reallocates the information in the phase space and provides a sharpened phase space representation of $f(x)$ in the following way.

Definition 2.2.4. *Given $f(x)$, the synchrosqueezed energy distribution $T_f(v,b)$ is defined by*

$$T_f(v,b) = \int_{\mathbb{R}^n \setminus B_1(0)} |W_f(a,b)|^2 \delta\left(\mathfrak{Re} v_f(a,b) - v\right) da$$

for $v, b \in \mathbb{R}^n$.

As we shall see, for $f(x) = \alpha(x)e^{2\pi i N\phi(x)}$ with sufficiently smooth amplitude $\alpha(x)$ and sufficiently steep phase $N\phi(x)$, we can show that for each b, the estimation $v_f(a,b)$ indeed approximates $N\nabla\phi(b)$

independently of a as long as $W_f(a,b)$ is non-negligible. As a direct consequence, for each b, the essential support of $T_f(v,b)$ in the v variable concentrates near $N\nabla\phi(b)$ (see Figure 2.3 for an example). In addition, we have the following property

$$\int T_f(v,b)dvdb = \int |W_f(a,b)|^2 \delta(\mathfrak{Re} v_f(a,b) - v)dvdadb = \int |W_f(a,b)|^2 dadb \approx \|f\|_2^2$$

from Fubini's theorem and the norm equivalence (2.4), for any $f(x)$ with its Fourier transform vanishing for $|\xi| < 1$.

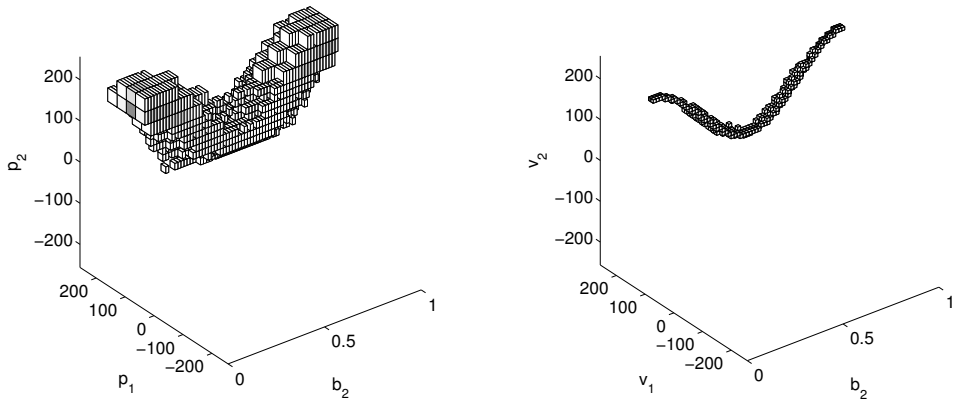


Figure 2.3: 2D example: Synchrosqueezed wave packet transform applied to a deformed plane wave $f(x) = \alpha(x)e^{2\pi i N\phi(x)}$. Left: The essential support of the wave packet transform $W_f(a,b)$ at $b_1 = 1$. Right: The essential support of the synchrosqueezed energy distribution $T_f(v,b)$ at the same b_1 value. $W_f(a,b)$ has been reallocated to form a sharp phase space representation $T_f(v,b)$.

2.2.3 Analysis

In this section, we show that the synchrosqueezed wave packet transform can distinguish well-separated local wavevectors from a superposition of multiple components.

Definition 2.2.5. *A function $f(x) = \alpha(x)e^{2\pi i N\phi(x)}$ is an intrinsic mode type function (IMT) of type (M,N) if $\alpha(x)$ and $\phi(x)$ satisfy*

$$\alpha(x) \in C^\infty, \quad |\nabla\alpha(x)| \le M, \quad 1/M \le \alpha(x) \le M$$
$$\phi(x) \in C^\infty, \quad 1/M \le |\nabla\phi(x)| \le M, \quad |\nabla^2\phi(x)| \le M.$$

Definition 2.2.6. *A function $f(x)$ is a well-separated superposition of type (M, N, K, s) if*

$$f(x) = \sum_{k=1}^{K} f_k(x)$$

where each $f_k(x) = \alpha_k(x)e^{2\pi i N_k \phi_k(x)}$ is an IMT of type (M, N_k) with $N_k \geq N$ and the phase functions satisfy the separation condition: for any $(a, b) \in \mathbb{R}^{2n}$, there exists at most one f_k satisfying that

$$|a|^{-s} |a - N_k \nabla \phi_k(b)| \leq 1.$$

We denote by $F(M, N, K, s)$ the set of all such functions.

The following theorem illustrates the main results of n-dimensional SSWPT for a superposition of IMTs without noise or perturbation. In what follows, when we write $O(\cdot)$, $\lesssim$, or $\gtrsim$, the implicit constants may depend on M, m and K.

Theorem 2.2.7. *Suppose the n-dimensional mother wave packet is of type (ϵ, m), for any fixed $\epsilon \in (0, 1)$ and any fixed integer $m \geq 0$. For a function $f(x)$, we define*

$$R_\epsilon = \{(a, b) : |W_f(a, b)| \geq |a|^{-ns/2}\sqrt{\epsilon}\},$$

$$S_\epsilon = \{(a, b) : |W_f(a, b)| \geq \sqrt{\epsilon}\},$$

and

$$Z_k = \{(a, b) : |a - N_k \nabla \phi_k(b)| \leq |a|^s\}$$

for $1 \leq k \leq K$. For fixed M, m, and K there exists a constant $N_0(M, m, K, s, \epsilon) \simeq \max\left\{\epsilon^{\frac{-2}{2s-1}}, \epsilon^{\frac{-1}{1-s}}\right\}$ such that for any $N > N_0$ and $f(x) \in F(M, N, K, s)$ the following statements hold.

(i) $\{Z_k : 1 \leq k \leq K\}$ are disjoint and $S_\epsilon \subset R_\epsilon \subset \bigcup_{1 \leq k \leq K} Z_k$;

(ii) For any $(a, b) \in R_\epsilon \cap Z_k$,

$$\frac{|v_f(a, b) - N_k \nabla \phi_k(b)|}{|N_k \nabla \phi_k(b)|} \lesssim \sqrt{\epsilon};$$

(iii) For any $(a, b) \in S_\epsilon \cap Z_k$,

$$\frac{|v_f(a, b) - N_k \nabla \phi_k(b)|}{|N_k \nabla \phi_k(b)|} \lesssim N_k^{-ns/2}\sqrt{\epsilon}.$$

Lemma 2.2.8. *Suppose $\Omega_a = \{k : |a| \in [\frac{N_k}{2M}, 2MN_k]\}$. Under the assumption of Theorem 2.2.7, we have*

$$W_f(a, b) = |a|^{-ns/2} \left(\sum_{k \in \Omega_a} \alpha_k(b) e^{2\pi i N_k \phi_k(b)} \widehat{w}\left(|a|^{-s}(a - N_k \nabla \phi_k(b))\right) + O(\epsilon) \right),$$

when $N > N_0\left(M, m, K, s, \epsilon\right) \simeq \max\left\{\epsilon^{\frac{-2}{2s-1}}, \epsilon^{\frac{-1}{1-s}}\right\}$.

Proof. Let us first estimate $W_f(a,b)$ assuming that $f(x)$ contains a single intrinsic mode function of type (M,N)

$$f(x) = \alpha(x)e^{2\pi i N\phi(x)}.$$

Using the definition of the wave packet transform, we have the following expression for $W_f(a,b)$.

$$\begin{aligned} W_f(a,b) &= \int \alpha(x)e^{2\pi i N\phi(x)}|a|^{ns/2}w(|a|^s(x-b))e^{-2\pi i(x-b)\cdot a}dx \\ &= \int \alpha(b+|a|^{-s}y)e^{2\pi i N\phi(b+|a|^{-s}y)}|a|^s w(y)e^{-2\pi i|a|^{-s}y\cdot a}d(|a|^{-s}y) \\ &= |a|^{-ns/2}\int \alpha(b+|a|^{-s}y)w(y)e^{2\pi i(N\phi(b+|a|^{-s}y)-|a|^{-s}y\cdot a)}dy. \end{aligned}$$

We claim that when N is sufficiently large

$$W_f(a,b) = \begin{cases} |a|^{-ns/2}O(\epsilon), & |a| \notin [\frac{N}{2M}, 2MN] \\ |a|^{-ns/2}\left(\alpha(b)e^{2\pi i N\phi(b)}\widehat{w}\left(|a|^{-s}(a-N\nabla\phi(b))\right)+O(\epsilon)\right), & |a| \in [\frac{N}{2M}, 2MN]. \end{cases} \tag{2.5}$$

First, let us consider the case $|a| \notin [\frac{N}{2M}, 2MN]$. Consider the integral

$$\int h(y)e^{ig(y)}dy$$

for smooth real functions $h(y)$ and $g(y)$, along with the differential operator

$$L = \frac{1}{i}\frac{\langle \nabla g, \nabla\rangle}{|\nabla g|^2}.$$

If $|\nabla g|$ does not vanish, we have

$$Le^{ig} = \frac{\langle \nabla g, i\nabla g e^{ig}\rangle}{i|\nabla g|^2} = e^{ig}.$$

Assuming that $h(y)$ decays sufficiently fast at infinity, we perform integration by parts r times to get

$$\int he^{ig}dy = \int h(L^r e^{ig})dy = \int ((L^*)^r h)e^{ig}dy,$$

where L^* is the adjoint of L. In the current setting, $W_f(a,b) = |a|^{-ns/2}\int h(y)e^{ig(y)}dy$ with

$$h(y) = \alpha(b+|a|^{-s}y)w(y), \quad g(y) = 2\pi(N\phi(b+|a|^{-s}y) - |a|^{-s}y\cdot a),$$

where $h(y)$ clearly decays rapidly at infinity since $w(y)$ is in the Schwartz class. In order to understand the impact of L and L^*, we need to bound the norm of

$$\nabla g(y) = 2\pi \left(N\nabla\phi(b + |a|^{-s}y) - a\right)|a|^{-s}$$

from below when $|a| \notin [\frac{N}{2M}, 2MN]$. If $|a| < \frac{N}{2M}$, then

$$|\nabla g| \gtrsim (|N\nabla\phi| - |a|)|a|^{-s} \gtrsim |N\nabla\phi||a|^{-s}/2 \gtrsim N^{1-s}.$$

If $|a| > 2MN$, then

$$|\nabla g| \gtrsim (|a| - |N\nabla\phi|)|a|^{-s} \gtrsim |a| \cdot |a|^{-s}/2 \gtrsim (|a|)^{1-s} \gtrsim N^{1-s}.$$

Hence $|\nabla g| \gtrsim N^{1-s}$ if $|a| \notin [\frac{N}{2M}, 2MN]$. Since $|\nabla g| \neq 0$ and each L^* contributes a factor of order $1/|\nabla g|$

$$\left|\int e^{ig(y)}((L^*)^r h)(y)dy\right| \lesssim N^{-(1-s)r}.$$

When

$$N \gtrsim \epsilon^{-1/((1-s)r)}, \tag{2.6}$$

we obtain

$$\left|\int e^{ig(y)}((L^*)^r h)(y)dy\right| \lesssim \epsilon.$$

Using the fact $W_f(a,b) = |a|^{-ns/2}\int h(y)e^{ig(y)}dy$, we have $|W_f(a,b)| \lesssim |a|^{-ns/2}\epsilon$.

Second, let us address the case $|a| \in [\frac{N}{2M}, 2MN]$. We want to approximate $W_f(a,b)$ with

$$|a|^{-ns/2}\alpha(b)e^{2\pi iN\phi(x)}\widehat{w}\left(|a|^{-s}(a - N\nabla\phi(b))\right).$$

Since $w(y)$ is in the Schwartz class, we can assume that $|w(y)| \leq \frac{C_u}{|y|^u}$ for some sufficient large u with C_u for $|y| \geq 1$. Therefore, the integration over $|y| \gtrsim \epsilon^{-1/u}$ yields a contribution of at most order $O(\epsilon)$. We can then estimate

$$|W_f(a,b)| = |a|^{-ns/2}\left(\int_{|y|\lesssim\epsilon^{-1/u}} \alpha(b + |a|^{-s}y)w(y)e^{2\pi i(N\phi(b+|a|^{-s}y) - |a|^{-s}y\cdot a)}dy + O(\epsilon)\right).$$

A Taylor expansion of $\alpha(x)$ and $\phi(x)$ shows that

$$\alpha(b + |a|^{-s}y) = \alpha(b) + \nabla\alpha(b^*)\cdot|a|^{-s}y$$

and

$$\phi(b + |a|^{-s}y) = \phi(b) + \nabla\phi(b)\cdot(|a|^{-s}y) + \frac{1}{2}(|a|^{-s}y)^t\nabla^2\phi(b^*)(|a|^{-s}y),$$

where in each case b^* is a point between b and $b + |a|^{-s}y$. We want to drop the last term from the above formulas without introducing a relative error larger than $O(\epsilon)$. We begin with the estimate

$$\int_{|y|\lesssim\epsilon^{-1/u}} |\nabla\alpha \cdot |a|^{-s} y w(y)| dy \lesssim \epsilon,$$

which holds if $\epsilon^{-n/u}|\nabla\alpha \cdot |a|^{-s}y| \lesssim \epsilon$, which is true when $|a|^{-s} \lesssim \epsilon^{1+(n+1)/u}$. Since $|a| \in [\frac{N}{2M}, 2MN]$, the above holds if

$$N \gtrsim \epsilon^{-(1+(n+1)/u)/s}. \tag{2.7}$$

We also need

$$\int_{|y|\lesssim\epsilon^{-1/u}} |\alpha(b)w(y)e^{2\pi i(N\phi(b)+N\nabla\phi(b)\cdot|a|^{-s}y-|a|^{-s}y\cdot a)}| \cdot |e^{2\pi iN/2(|a|^{-s}y)^t\nabla^2\phi(|a|^{-s}y)} - 1| dy \lesssim \epsilon.$$

Since $|e^{ix} - 1| \leq |x|$, the above inequality is equivalent to

$$\int_{|y|\lesssim\epsilon^{-1/u}} \alpha(b)w(y)e^{2\pi i(N\phi(b)+N\nabla\phi(b)\cdot|a|^{-s}y-|a|^{-s}y\cdot a)}|2\pi N/2(|a|^{-s}y)^t\nabla^2\phi(|a|^{-s}y)|dy \lesssim \epsilon,$$

which is true if $\epsilon^{-n/u}N(|a|^{-s}y)^t\nabla^2\phi(|a|^{-s}y) \lesssim \epsilon$, which in turn holds if $N|a|^{-2s}|y|^2 \lesssim \epsilon^{1+n/u}$. Because $|y| \lesssim \epsilon^{-\frac{1}{u}}$ and $|a| \in [\frac{N}{2M}, 2MN]$, the above inequality is valid when

$$N \gtrsim \epsilon^{-(1+(n+2)/u)/(2s-1)}. \tag{2.8}$$

In summary, for N larger than the maximum of the right hand sides of (2.6), (2.7) and (2.8), if $|a| \in [\frac{N}{2M}, 2MN]$ then we have

$$\begin{aligned}
W_f(a,b) &= |a|^{-ns/2}\left(\int_{|y|\lesssim\epsilon^{-1/u}} \alpha(b)w(y)e^{2\pi i(N\phi(b)+N\nabla\phi(b)\cdot|a|^{-s}y-|a|^{-s}y\cdot a)}dy + O(\epsilon)\right)\\
&= |a|^{-ns/2}\left(\int_{|y|\lesssim\epsilon^{-1/u}} \left(\alpha(b)e^{2\pi iN\phi(b)}\right) w(y)e^{2\pi i(N\nabla\phi(b)-a)\cdot|a|^{-s}y}dy + O(\epsilon)\right)\\
&= |a|^{-ns/2}\left(\int_{\mathbb{R}^n} \left(\alpha(b)e^{2\pi iN\phi(b)}\right) w(y)e^{2\pi i(N\nabla\phi(b)-a)\cdot|a|^{-s}y}dy + O(\epsilon)\right)\\
&= |a|^{-ns/2}\left(\alpha(b)e^{2\pi iN\phi(b)}\widehat{w}\left(|a|^{-s}(a - N\nabla\phi(b))\right) + O(\epsilon)\right),
\end{aligned}$$

where the third line uses the fact that the integration of $w(y)$ outside the set $\{y : |y| \lesssim \epsilon^{-1/u}\}$ is again of order $O(\epsilon)$.

Now let us return to the general case, where $f(x)$ is a superposition of K well-separated intrinsic

mode components:

$$f(x) = \sum_{k=1}^{K} f_k(x) = \sum_{k=1}^{K} \alpha_k(x) e^{2\pi i N \phi_k(x)}.$$

By linearity of the wave packet transform and (2.5), we find:

$$W_f(a,b) = |a|^{-ns/2} O(\epsilon),$$

if $|a| \notin [\frac{N}{2M}, 2MN]$;

$$W_f(a,b) = |a|^{-ns/2} \left(\sum_{k=1}^{K} \alpha_k(b) e^{2\pi i N \phi_k(b)} \widehat{w}\left(|a|^{-s}(a - N\nabla\phi_k(b))\right) + O(\epsilon) \right),$$

if $|a| \in [\frac{N}{2M}, 2MN]$. □

The next lemma estimates $\nabla_b W_f(a,b)$ when Ω_a is not empty, i.e., the case where $W_f(a,b)$ is non-negligible.

Lemma 2.2.9. *Suppose $\Omega_a = \{k : |a| \in [\frac{N_k}{2M}, 2MN_k]\}$ is not empty. Under the assumption of Theorem 2.2.7, we have*

$$\nabla_b W_f(a,b) = 2\pi i |a|^{-ns/2} \left(\sum_{k \in \Omega_a} N_k \nabla\phi_k(b) \alpha_k(b) e^{2\pi i N_k \phi_k(b)} \widehat{w}\left(|a|^{-s}\left(a - N_k \nabla\phi_k(b)\right)\right) + |a| O\left(\epsilon\right) \right),$$

when $N > N_0\left(M, m, K, s, \epsilon\right) \simeq \max\left\{\epsilon^{\frac{-2}{2s-1}}, \epsilon^{\frac{-1}{1-s}}\right\}$.

Proof. The proof is similar to the one of Lemma 2.2.8. Assume that $f(x)$ contains a single intrinsic mode function, i.e.,

$$f(x) = \alpha(x) e^{2\pi i N \phi(x)},$$

then

$$\begin{aligned}
\nabla_b W_f(a,b) &= \int_{\mathbb{R}^n} \alpha(x) e^{2\pi i N \phi(x)} |a|^{ns/2} \left(\nabla w(|a|^s(x-b))(-|a|^s) + 2\pi i p w(|a|^s(x-b))\right) e^{-2\pi i (x-b)\cdot a} dx \\
&= \int_{\mathbb{R}^n} \alpha(b + |a|^{-s} y) e^{2\pi i N \phi(b + |a|^{-s} y)} |a|^{-ns/2} \left(\nabla w(y)(-|a|^s) + 2\pi i p w(y)\right) e^{-2\pi i |a|^{-s} y \cdot a} dy \\
&= \int_{\mathbb{R}^n} \alpha(b + |a|^{-s} y) e^{2\pi i N \phi(b + |a|^{-s} y)} |a|^{-ns/2} \nabla w(y)(-|a|^s) e^{-2\pi i |a|^{-s} y \cdot a} dy \\
&\quad + \int_{\mathbb{R}^n} \alpha(b + |a|^{-s} y) e^{2\pi i N \phi(b + |a|^{-s} y)} |a|^{-ns/2} 2\pi i a w(y) e^{-2\pi i |a|^{-s} y \cdot a} dy.
\end{aligned}$$

Forming a Taylor expansion and following the same argument as in the proof of Lemma 2.2.8 gives

the following approximation for $|a| \in [\frac{N}{2M}, 2MN]$

$$\begin{aligned}\nabla_b W_f(a,b) =& \left(-2\pi i|a|^{-ns/2}(a - N\nabla\phi(b))\alpha(b)e^{2\pi i N\phi(b)}\widehat{w}(|a|^{-s}(a - N\nabla\phi(b))) + O(\epsilon)\right) \\ &+ 2\pi i|a|^{-ns/2}a\left(\alpha(b)e^{2\pi i N\phi(b)}\widehat{w}\left(|a|^{-s}(a - N\nabla\phi(b))\right) + O(\epsilon)\right) \\ =&2\pi i|a|^{-ns/2}\left(N\nabla\phi(b)\alpha(b)e^{2\pi i N\phi(b)}\widehat{w}(|a|^{-s}(a - N\nabla\phi(b))) + |a|O(\epsilon)\right).\end{aligned}$$

For $f(x) = \sum_{k=1}^{K} f_k(x) = \sum_{k=1}^{K} \alpha_k(x)e^{2\pi i N\phi_k(x)}$, taking sum over K terms gives

$$\nabla_b W_f(a,b) = 2\pi i|a|^{-ns/2}\left(\sum_{k\in\Omega_a}\left(N\nabla\phi_k(b)\alpha_k(x)e^{2\pi i N\phi_k(b)}\widehat{w}(|a|^{-s}(a - N\nabla\phi_k(b)))\right) + |a|O(\epsilon)\right)$$

for $|a| \in [\frac{N}{2M}, 2MN]$. □

We are now ready to prove the theorem.

Proof. For (i), the well-separation condition implies that $\{Z_k : 1 \le k \le K\}$ are disjoint.

Let (a,b) be a point in $R_\epsilon = \{(a,b) : |W_f(a,b)| \ge |a|^{-ns/2}\sqrt{\epsilon}\}$. From the above lemma, we have

$$W_f(a,b) = |a|^{-ns/2}\left(\sum_{k\in\Omega_a}\alpha_k(b)e^{2\pi i N_k\phi_k(b)}\widehat{w}\left(|a|^{-s}(a - N_k\nabla\phi_k(b))\right) + O(\epsilon)\right).$$

Therefore, there exists k between 1 and K such that $\widehat{w}\left(|a|^{-s}(a - N_k\nabla\phi_k(b))\right)$ is non-zero. From the definition of $\widehat{w}(\xi)$, we see that this implies $(a,b) \in Z_k$. Hence $R_\epsilon \subset \bigcup_{k=1}^{K} Z_k$. It's obvious that $S_\epsilon \subset R_\epsilon$.

To show (ii), let us recall that $v_f(a,b)$ is defined as

$$v_f(a,b) = \frac{\nabla_b W_f(a,b)}{2\pi i W_f(a,b)}$$

for $W_f(a,b) \neq 0$. If $(a,b) \in R_\epsilon \cap Z_k$, then

$$W_f(a,b) = |a|^{-ns/2}\left(\alpha_k(b)e^{2\pi i N_k\phi_k(b)}\widehat{w}\left(|a|^{-s}(a - N_k\nabla\phi(b))\right) + O(\epsilon)\right)$$

and

$$\nabla_b W_f(a,b) = 2\pi i|a|^{-ns/2}\left(N_k\nabla\phi_k(b)\alpha_k(b)e^{2\pi i N_k\phi_k(b)}\widehat{w}(|a|^{-s}(a - N_k\nabla\phi_k(b))) + |a|O(\epsilon)\right)$$

as the other terms drop out since $\{Z_k\}$ are disjoint. Hence

$$v_f(a,b) = \frac{N_k\nabla\phi_k(b)\left(\alpha_k(b)e^{2\pi i N_k\phi_k(b)}\widehat{w}\left(|a|^{-s}(a - N_k\nabla\phi_k(b))\right) + O(\epsilon)\right)}{\left(\alpha_k(b)e^{2\pi i N_k\phi_k(b)}\widehat{w}\left(|a|^{-s}(a - N_k\nabla\phi_k(b))\right) + O(\epsilon)\right)}.$$

Let us denote the term $\alpha_k(b)e^{2\pi i N_k\phi_k(b)}\widehat{w}\left(|a|^{-s}(a-N_k\nabla\phi_k(b))\right)$ by g. Then

$$v_f(a,b)=\frac{N_k\nabla\phi_k(b)\left(g+O(\epsilon)\right)}{g+O(\epsilon)}.$$

Since $|W_f(a,b)|\geq|a|^{-ns/2}\sqrt{\epsilon}$ for $(a,b)\in R_\epsilon$, $|g|\gtrsim\sqrt{\epsilon}$, and therefore

$$\frac{|v_f(a,b)-N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|}\lesssim\left|\frac{O(\epsilon)}{g+O(\epsilon)}\right|\lesssim\sqrt{\epsilon}.$$

Similarly, if $(a,b)\in S_\epsilon\cap Z_k$, then

$$\frac{|v_f(a,b)-N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|}\lesssim\left|\frac{O(\epsilon)}{g+O(\epsilon)}\right|\lesssim\frac{\sqrt{\epsilon}}{N_k^{ns/2}},$$

since $|g|\gtrsim N_k^{ns/2}\sqrt{\epsilon}$ for $(a,b)\in S_\epsilon\cap Z_k$. □

2.3 2D Synchrosqueezed Curvelet Transform

2.3.1 Motivation

In some applications such as wave field separation problems [143, 163] and ground roll removal problems [17, 72, 185] in geophysics, it is required to separates overlapping wavefronts or banded wave-like components. In this case, the boundary of these components gives rise to many nonzero coefficients of wave packet transform, which results in unexpected interferential synchrosqueezed energy distribution (see Figure 2.4 top-right). This would dramatically reduce the accuracy of local wave-vector estimation, because the locations of nonzero energy provide estimation of local wave-vectors. As shown in Figure 2.4 (top-right), there exist misleading local wave-vector estimates at the location where the signal is negligible. Even if at the location where the signal is relevant, the relative error is still unacceptable.

To solve this problem, an empirical idea is that, good basis elements in the synchrosqueezed transform should look like the components, i.e., they should appear in a needle-like shape. An optimal solution is curvelets. The curvelet transform is anisotropic (as shown in Figure 2.5 right), and is designed for optimally representing curved edges [23, 148] and banded wavefronts [19]. This motivates the design of the synchrosqueezed curvelet transform (SSCT) as a better tool to estimate local wave-vectors of wavefronts or banded wave-like components in this paper. The estimate of local wave-vectors provided by SSCT is much better than that by SSWPT as shown in Figure 2.4 (bottom).

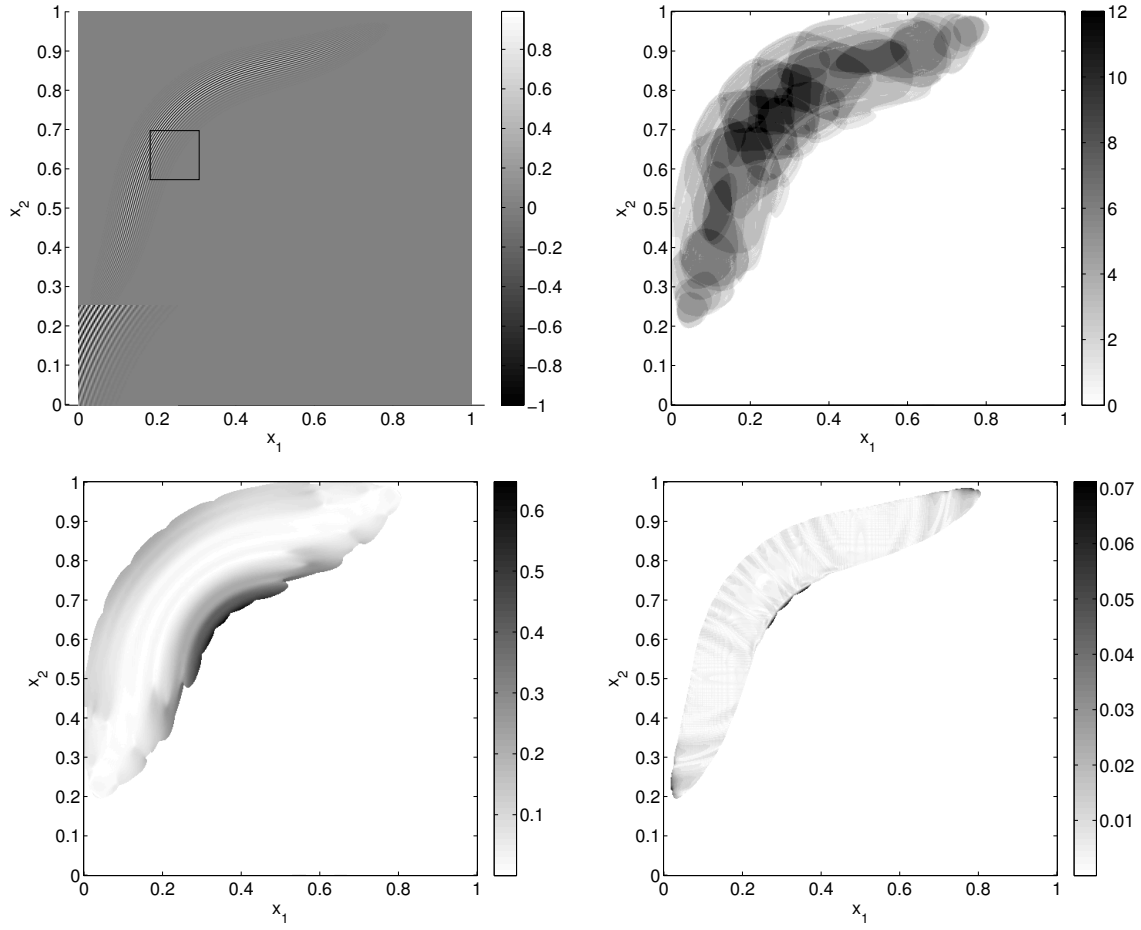


Figure 2.4: Top-left: A banded deformed plane wave, $f(x) = e^{-\frac{(\phi(x)-0.7)^2}{(4/N)^2}} e^{2\pi i N \phi(x)}$, where $N = 135$ and $\phi(x) = x_1 + (1-x_2) + 0.1\sin(2\pi x_1) + 0.1\sin(2\pi(1-x_2))$. Top-right: Number of nonzero discrete synchrosqueezed energy of SSWPT at each grid point of space domain. Bottom-left: Relative error between the mean local wave-vector estimate (defined in [182]) and the exact local wave-vector using SSWPT. Bottom-right: Relative error between the mean local wave-vector estimate and the exact local wave-vector using SSCT.

2.3.2 Definition

Below is a brief introduction to the generalized curvelet transform with a radial scaling parameter $t < 1$ and an angular scaling parameter $s \in (\frac{1}{2}, t)$. Similar to the discussion in [182], it is crucial to assume $\frac{1}{2} < s < t < 1$, so as to obtain accurate estimates of local wave-vectors for reasonable large wavenumbers. It is proved in the next section, $s < t$ guarantees precise estimates in the case of banded wave-like components. Here are some notations for the generalized curvelet transform.

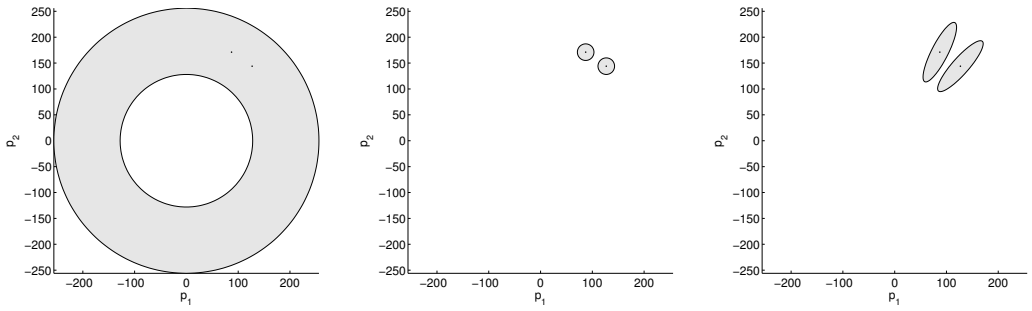


Figure 2.5: Comparison of localized supports of continuous wavelets (left), wave packets (middle) and curvelets (right) in the Fourier domain. Two dots in each plot show the support of the Fourier transforms of the superposition of two plane waves $e^{2\pi ip\cdot x}$ and $e^{2\pi iq\cdot x}$ with the same wave-number ($|p| = |q|$) but different wave-vectors ($p \neq q$).

1. The scaling matrix
$$A_a = \begin{pmatrix} a^t & 0 \\ 0 & a^s \end{pmatrix},$$
where a is the distance from the center of one curvelet to the origin of Fourier domain.

2. The rotation angle θ and rotation matrix
$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

3. The unit vector $u_\theta = (\cos\theta, \sin\theta)^T$ of rotation angle θ and T denotes a transpose.

4. θ_α represents the argument of given vector α.

5. $w(x)$ of $x \in \mathbb{R}^2$ denotes the mother curvelet, which belongs to the class of mother wave packets of some type (ϵ, m) in Section 2.2 and obeys the admissibility condition: $\exists 0 < c_1 < c_2 < \infty$ such that
$$c_1 \leq \int_0^{2\pi} \int_1^\infty a^{-(t+s)} |\widehat{w}\left(A_a^{-1} R_\theta^{-1} (\xi - a \cdot u_\theta)\right)|^2 a da d\theta \leq c_2$$
for any $|\xi| \geq 1$.

With the notations above, it is ready to define a family of curvelets through scaling, modulation, and translation as follows, controlled by the geometric parameter s and t.

Definition 2.3.1. *Given geometric scaling parameters $\frac{1}{2} < s < t < 1$ and a mother curvelet of type (ϵ, m), the family of curvelets $\{w_{a\theta b}(x), a \in [1, \infty), \theta \in [0, 2\pi), b \in \mathbb{R}^2\}$ is constructed as*

$$w_{a\theta b}(x) = a^{\frac{t+s}{2}} e^{2\pi i a(x-b)\cdot u_\theta} w\left(A_a R_\theta^{-1}(x-b)\right),$$

or equivalently, in the frequency domain

$$\widehat{w_{a\theta b}}(\xi) = \widehat{w}\left(A_a^{-1}R_\theta^{-1}\left(\xi - a\cdot u_\theta\right)\right)e^{-2\pi i b\cdot\xi}a^{-\frac{t+s}{2}}.$$

It is clear from the definition that the Fourier transform $\widehat{w_{a\theta b}}(\xi)$ has an ellipse-like essential support $\{\xi : |A_a^{-1}R_\theta^{-1}\left(\xi - a\cdot u_\theta\right)| \leq 1\}$ centered at $a\cdot u_\theta$ with a major radius a^t and a minor radius a^s. Meanwhile, $w_{a\theta b}(x)$ is centered in space at b with an essential support of length $O\left(a^{-s}\right)$ and width $O\left(a^{-t}\right)$. By this appropriate construction, each curvelet is scaled to have the same L^2 norm with the mother curvelet $w(x)$. The generalized curvelet transform can also be considered as a generalization of the wave packet transform in Section 2.2 with two different scaling parameters s and t. This family of functions is quantitatively similar to wavelets when $s = t = 1$, wave atoms [49] when $s = t = \frac{1}{2}$, and curvelets [19, 24, 25] when $s = \frac{1}{2}$ and $t = 1$. In real applications, it is beneficial to adaptively tune s and t for better estimates of local wave vectors in complex data structures.

Similar to the curvelet transform, the generalized curvelet transform is defined as follows.

Definition 2.3.2. *The generalized curvelet transform of a function $f(x)$ is a function*

$$W_f(a,\theta,b) \quad = \quad \langle f, w_{a\theta b}\rangle = \int_{\mathbb{R}^2} f(x)\overline{w_{a\theta b}(x)}dx$$

for $a \in [1,\infty)$, $\theta \in [0,2\pi)$, $b \in \mathbb{R}^2$.

If the Fourier transform $\widehat{f}(\xi)$ vanishes for $|\xi| < 1$, one can check the following L^2 norms equivalence up to a uniform constant factor following the proof of Theorem 1 in [25], i.e.,

$$c_1\int |f(x)|^2dx \leq \int |W_f(a,\theta,b)|^2 a da d\theta db \leq c_2\int |f(x)|^2dx.$$

Definition 2.3.3. *The local wave vector information function of a function $f(x)$ at (a,θ,b) for $a \in [1,\infty)$, $\theta \in [0,2\pi)$, $b \in \mathbb{R}^2$ is*

$$v_f(a,\theta,b) = \begin{cases} \frac{\nabla_b W_f(a,\theta,b)}{2\pi i W_f(a,\theta,b)}, & \text{for } W_f(a,\theta,b) \neq 0; \\ (\infty,\infty), & \text{otherwise.} \end{cases}$$

Since $v_f(a,\theta,b)$ estimates the local wave vectors accurately, as we shall see, reallocating the coefficients with the same v_f together would generate a sharpened phase space representation of $f(x)$. This motivates the design of the synchrosqueezed energy distribution as follows.

Definition 2.3.4. *Given $f(x)$, the synchrosqueezed energy distribution $T_f(v,b)$ is*

$$T_f(v,b) = \int |W_f(a,\theta,b)|^2\delta\left(\mathfrak{Re}v_f(a,\theta,b) - v\right)a da d\theta$$

for $v \in \mathbb{R}^2$, $b \in \mathbb{R}^2$.

For $f(x)$ with Fourier transform vanishing for $|\xi| < 1$, the following norm equivalence holds

$$\int T_f(v,b)dvdb = \int |W_f(a,\theta,b)|^2 a da d\theta db \approx \|f\|_2^2$$

as a consequence of the L^2 norm equivalence between $W_f(a,\theta,b)$ and $f(x)$.

2.3.3 Analysis

To model a wave-like component with a band-shape support, we are going to analyze components of the form

$$f(x) = e^{-(\phi(x)-c)^2/\sigma^2}\alpha(x)e^{2\pi i N\phi(x)},$$

where $\alpha(x)$ is a smooth amplitude function, $\phi(x)$ a smooth phase function, and σ is a band parameter that controls the width of the signal.

To understand how large the bandwidth should be so as to obtain accurate local wave vector estimates by the SSCT, we assume $\sigma = \Theta(N^{-\eta})$ and show that the SSCT gives good estimates when $\eta < t$ and N is sufficiently large. In the space domain, a generalized curvelet at the scale $a = O(N)$ has a width $O(N^{-t})$. $\sigma \geq N^{-\eta}$ with $\eta < t$ indicates that the bandwidth σ of $e^{-(\phi(x)-c)^2/\sigma^2}\alpha(x)e^{2\pi i N\phi(x)}$ can be almost as narrow as the width of a generalized curvelet that sharing the same wave number $O(N)$, when N is sufficiently large.

Definition 2.3.5. *For any $c \in \mathbb{R}$, $N > 0$ and $M > 0$, $f(x) = e^{-(\phi(x)-c)^2/\sigma^2}\alpha(x)e^{2\pi i N\phi(x)}$ is a banded intrinsic mode function of type (M,N,η), if $\alpha(x)$ and $\phi(x)$ satisfy*

$$\begin{aligned} &\alpha(x) \in C^\infty, \qquad |\nabla\alpha(x)| \leq M, \qquad 1/M \leq \alpha(x) \leq M, \\ &\phi(x) \in C^\infty, \quad 1/M \leq |\nabla\phi(x)| \leq M, \quad |\nabla^2\phi(x)| \leq M, \\ &\qquad\qquad \textit{and} \quad \sigma \geq N^{-\eta}. \end{aligned}$$

Definition 2.3.6. *A function $f(x)$ is a well-separated superposition of type (M,N,η,s,t,K) if*

$$f(x) = \sum_{k=1}^{K} f_k(x),$$

where each $f_k(x) = e^{-(\phi_k(x)-c_k)^2/\sigma_k^2}\alpha_k(x)e^{2\pi i N\phi_k(x)}$ is a banded intrinsic mode function (IMT) of type (M,N_k,η) with $N_k \geq N$ and they satisfy the separation condition: $\forall a \in [1,\infty)$ and $\forall\theta \in [0,2\pi)$, there is at most one banded intrinsic mode function f_k satisfying that

$$|A_a^{-1}R_\theta^{-1}(a\cdot u_\theta - N_k\nabla\phi_k(b))| \leq 1.$$

We denote by $F(M, N, \eta, s, t, K)$ *the set of all such functions.*

The first theorem in this section demonstrates the accuracy of the local wave vector estimation of the banded IMTs when the given data does not contain noise.

Theorem 2.3.7. *Suppose the* $2D$ *mother curvelet is of type* (ϵ, m)*, for any fixed* $\epsilon \in (0,1)$ *and any fixed integer* $m \geq 0$*. For a function* $f(x)$*, we define*

$$R_\epsilon = \left\{(a,\theta,b) : |W_f(a,\theta,b)| \geq a^{-\frac{s+t}{2}}\sqrt{\epsilon}\right\},$$

$$S_\epsilon = \left\{(a,\theta,b) : |W_f(a,\theta,b)| \geq \sqrt{\epsilon}\right\},$$

and

$$Z_k = \left\{(a,\theta,b) : |A_a^{-1}R_\theta^{-1}\left(a\cdot u_\theta - N_k\nabla\phi_k(b)\right)| \leq 1\right\}$$

for $1 \leq k \leq K$*. For fixed* M*,* m*,* s*,* t*,* η*, and* ϵ*, there exists*

$$N_0(M, m, s, t, \eta, \epsilon) \simeq \max\left\{\epsilon^{\frac{-1}{1-t}}, \epsilon^{\frac{-2}{t-\eta}}, \epsilon^{\frac{-2}{2s-1}}\right\}$$

such that for any $N > N_0(M, m, s, t, \eta, \epsilon)$ *and* $f(x) \in F(M, N, \eta, s, t, K)$ *the following statements hold.*

(i) $\{Z_k : 1 \leq k \leq K\}$ *are disjoint and* $S_\epsilon \subset R_\epsilon \subset \bigcup_{1\leq k\leq K} Z_k$*.*

(ii) For any $(a,\theta,b) \in R_\epsilon \cap Z_k$*,*

$$\frac{|v_f(a,\theta,b) - N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|} \lesssim \sqrt{\epsilon}.$$

(iii) For any $(a,\theta,b) \in S_\epsilon \cap Z_k$*,*

$$\frac{|v_f(a,\theta,b) - N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|} \lesssim N_k^{-\frac{s+t}{2}}\sqrt{\epsilon}.$$

For simplicity, the notations $O(\cdot)$, $\lesssim$ and $\gtrsim$ are used when the implicit constants may only depend on M, m and K. The proof of Theorem 2.3.7 relies on two lemmas below to estimate $W_f(a,\theta,b)$ and $\nabla_b W_f(a,\theta,b)$.

Lemma 2.3.8. *Suppose*

$$\Omega_{a\theta b} = \left\{k : a \in \left(\frac{N_k}{2M}, 2MN_k\right), \left|\theta_{\nabla\phi_k(b)} - \theta\right| < \arcsin\left(\left(\frac{M}{N_k}\right)^{t-s}\right)\right\}.$$

Under the assumption of Theorem 2.3.7, the following estimation of $W_f(a,\theta,b)$ *holds when* $N >$

$N_0\left(M,m,s,t,\eta,\epsilon\right) \simeq \max\left\{\epsilon^{\frac{-1}{1-t}}, \epsilon^{\frac{-2}{t-\eta}}, \epsilon^{\frac{-2}{2s-1}}\right\}$:

$$W_f(a,\theta,b) = a^{-\frac{s+t}{2}}\left(\sum_{k\in\Omega_{a\theta b}} f_k(b)\widehat{w}\left(A_a^{-1}R_\theta^{-1}\left(a\cdot u_\theta - N_k\nabla\phi_k(b)\right)\right) + O\left(\epsilon\right)\right).$$

Proof. We only need to discuss the case when $K=1$. The result for general K is an easy extension by the linearity of generalized curvelet transform. Suppose $f(x)$ contains a single banded intrinsic mode function of type (M,N,η)

$$f(x) = e^{-(\phi(x)-c)^2/\sigma^2}\alpha(x)e^{2\pi iN\phi(x)}.$$

We claim that when N is large enough, the approximation of $W_f(a,\theta,b)$ holds. By the definition of generalized curvelet transform, it holds that

$$\begin{aligned} W_f(a,\theta,b) &= \int_{R^2} f(x)a^{\frac{s+t}{2}}w(A_aR_\theta^{-1}(x-b))e^{-2\pi ia(x-b)\cdot u_\theta}dx \\ &= a^{-\frac{s+t}{2}}\int_{R^2} f(b+R_\theta A_a^{-1}y)w(y)e^{-2\pi ia^{1-t}y_1}dy. \end{aligned}$$

Step 1: We start with the proof of (2) first.

Let $h(y) = w(y)e^{-(\phi(b+R_\theta A_a^{-1}y)-c)^2/\sigma^2}\alpha(b+R_\theta A_a^{-1}y)$ and $g(y) = 2\pi(N\phi(b+R_\theta A_a^{-1}y) - a^{1-t}y_1)$, then we have

$$W_f(a,\theta,b) = a^{-\frac{s+t}{2}}\int_{\mathbb{R}^2} h(y)e^{ig(y)}dy,$$

with real smooth functions $h(y)$ and $g(y)$. Consider the differential operator

$$L = \frac{1}{i}\frac{\langle\nabla g,\nabla\rangle}{|\nabla g|^2}.$$

If $|\nabla g|$ does not vanish, we have

$$Le^{ig} = \frac{\langle\nabla g, i\nabla g e^{ig}\rangle}{i|\nabla g|^2} = e^{ig}.$$

By the definition of $w(y)$, we know $h(y)$ is decaying rapidly at infinity. Then we can apply integration by parts to get

$$\int_{\mathbb{R}^2} he^{ig}dy = \int_{\mathbb{R}^2} h(Le^{ig})dy = -\int_{\mathbb{R}^2}\nabla\cdot\left(\frac{h\nabla g}{i|\nabla g|^2}\right)e^{ig}dy.$$

Hence, we need to estimate $\left|\nabla\cdot\left(\frac{h\nabla g}{i|\nabla g|^2}\right)\right|$. Because

$$\nabla\cdot\left(\frac{h\nabla g}{i|\nabla g|^2}\right) = \frac{1}{i}\left(\frac{\nabla h\cdot\nabla g}{|\nabla g|^2} + h\nabla\cdot\left(\frac{\nabla g}{|\nabla g|^2}\right)\right)$$

and $|h(y)| \lesssim 1$, we only need to estimate $\left|\frac{\nabla h \cdot \nabla g}{|\nabla g|^2}\right|$ and $\left|\frac{\partial^2 g}{\partial y_i \partial y_j}\frac{1}{|\nabla g|^2}\right|$ for $i, j = 1, 2$.

Let $z = (z_1, z_2)^T = R_\theta^{-1}\nabla\phi(b + R_\theta A_a^{-1} y)$, $v_1 = NA_a^{-1}R_\theta^{-1}\nabla\phi(b + R_\theta A_a^{-1} y)$ and $v_2 = (a^{1-t}, 0)^T$, then $\nabla g(y) = 2\pi(v_1 - v_2) = 2\pi((Nz_1 - a)a^{-t}, Na^{-s}z_2)$.

Case 1: $a \notin (\frac{N}{2M}, 2MN)$.

When $a \geq 2MN$, then

$$|\nabla g(y)| \geq a^{1-t} - MNa^{-t} = \frac{a^{1-t}}{2} + (\frac{a}{2} - MN)a^{-t} \geq \frac{a^{1-t}}{2} \gtrsim N^{1-t}.$$

When $a \leq \frac{N}{2M}$, then

$$|\nabla g(y)| \gtrsim \frac{Na^{-t}}{M} - a^{1-t} \geq \frac{Na^{-t}}{2M} \gtrsim N^{1-t}.$$

So

$$|\nabla g(y)| \gtrsim N^{1-t} \tag{2.9}$$

for $a \notin (\frac{N}{2M}, 2MN)$.

If $a \geq 2MN$, then $\left|\frac{\partial^2 g}{\partial y_i \partial y_j}\right| \lesssim Na^{-2s} \lesssim N^{1-2s}$, implying that

$$\left|\frac{\partial^2 g}{\partial y_i \partial y_j}\frac{1}{|\nabla g|^2}\right| \lesssim N^{1-2s}/N^{2-2t} = \frac{1}{N^{1-2(t-s)}}.$$

Since $|z| \geq \frac{1}{M}$, then either $|z_1| \geq \frac{1}{\sqrt{2}M}$ or $|z_2| \geq \frac{1}{\sqrt{2}M}$ holds. If $a \leq \frac{N}{2M}$, then

$$\begin{aligned}
\left|\frac{\partial^2 g}{\partial y_i \partial y_j}\frac{1}{|\nabla g|^2}\right| &\lesssim \frac{Na^{-2s}}{(Nz_1 - a)^2 a^{-2t} + N^2 a^{-2s} z_2^2} \\
&= \frac{1}{(z_1 - \frac{a}{N})^2 Na^{-2(t-s)} + Nz_2^2} \\
&\lesssim \max\{\frac{1}{N^{1-2(t-s)}}, \frac{1}{N}\}. \\
&= \frac{1}{N^{1-2(t-s)}}.
\end{aligned}$$

In sum,

$$\left|\frac{\partial^2 g}{\partial y_i \partial y_j}\frac{1}{|\nabla g|^2}\right| \lesssim \frac{1}{N^{1-2(t-s)}} \tag{2.10}$$

for $a \notin (\frac{N}{2M}, 2MN)$.

Notice that the dominant term of ∇h is

$$w(y)\alpha(b + R_\theta A_a^{-1} y)e^{-(\phi(b+R_\theta A_a^{-1} y)-c)^2/\sigma^2} \cdot \frac{-2(\phi(b + R_\theta A_a^{-1} y) - c)}{\sigma^2} A_a^{-1} z$$

and the other terms are of order 1. Because $e^{-\frac{x^2}{\sigma^2}} \cdot \frac{|x|}{\sigma^2} \leq e^{-\frac{1}{2}} \cdot \frac{1}{\sigma\sqrt{2}}$, then

$$\left|\frac{\nabla h \cdot \nabla g}{|\nabla g|^2}\right| \lesssim \frac{1}{\sigma}\left|\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2}\right| + \left|\frac{1}{|\nabla g|}\right| \lesssim N^{\eta}\left|\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2}\right| + \frac{1}{N^{1-t}}.$$

Recall that $\nabla g = 2\pi(NA_a^{-1}z - (a^{1-t},0)^T)$, then

$$\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2} \approx \frac{(Nz_1-a)a^{-2t}z_1 + Na^{-2s}z_2^2}{(Nz_1-a)^2a^{-2t} + N^2a^{-2s}z_2^2}.$$

If $z_1z_2 \neq 0$, then $\left|\frac{Na^{-2s}z_2^2}{N^2a^{-2s}z_2^2}\right| = \frac{1}{N}$ and $\left|\frac{(Nz_1-a)a^{-2t}z_1}{(Nz_1-a)^2a^{-2t}}\right| \approx \frac{1}{|Nz_1-a|} \approx \frac{1}{N}$, which implies that $\left|\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2}\right| \lesssim \frac{1}{N}$. If $z_1z_2 = 0$, then it is easy to check that $\left|\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2}\right| \approx \frac{1}{N}$. Hence,

$$\left|\frac{\nabla h \cdot \nabla g}{|\nabla g|^2}\right| \lesssim N^{\eta}\left|\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2}\right| + \frac{1}{N^{1-t}} \lesssim \frac{1}{N^{1-\eta}} + \frac{1}{N^{1-t}} \lesssim \frac{1}{N^{1-t}} \tag{2.11}$$

for $a \notin (\frac{N}{2M}, 2MN)$.

By (2.10) and (2.11), we have

$$\left|\int_{\mathbb{R}^2} he^{ig}dy\right| = \left|\int_{\mathbb{R}^2} \nabla\cdot\left(\frac{h\nabla g}{i|\nabla g|^2}\right)e^{ig}dy\right| \lesssim \left|\nabla\cdot\left(\frac{h\nabla g}{i|\nabla g|^2}\right)\right|\left(||w||_{L^1} + ||\nabla w||_{L^1}\right) \lesssim \frac{1}{N^{1-t}}$$

for $a \notin (\frac{N}{2M}, 2MN)$. So,

$$W_f(a,\theta,b) = a^{-\frac{s+t}{2}}O(\epsilon),$$

when $N \gtrsim \epsilon^{\frac{-1}{1-t}}$ and $a \notin (\frac{N}{2M}, 2MN)$.

Case 2: $a \in (\frac{N}{2M}, 2MN)$ and $|\theta_{\nabla\phi(b)} - \theta| \geq \theta_0$.

Observing that $\nabla g(y) = 2\pi A_a^{-1}R_\theta^{-1}(N\nabla\phi(b + R_\theta A_a^{-1}y) - a\cdot u_\theta)$, we can expect $|\nabla g|$ is large when $\theta_{\nabla\phi(b)}$ is far away from θ. Notice that $w(y)$ is in the Schwartz class, then $\exists C_u > 0$ such that $|w(y)| \leq \frac{C_u}{y^u}$ for $|y| \geq 1$ and any u large enough. So

$$W_f(a,\theta,b) \quad = \quad a^{-\frac{s+t}{2}}\left(\int_{|y|\lesssim\epsilon^{-1/u}} f(b + R_\theta A_a^{-1}y)w(y)e^{-2\pi i a^{1-t}y_1}dy + O(\epsilon)\right).$$

Define $D = \{y : |y| \lesssim \epsilon^{-1/u}\}$ and $D_+ = \{y : |y| \lesssim \epsilon^{-1/u} + 1\}$. Suppose $X_D(y)$ is a positive and smooth function compactly supported in D_+ such that $X_D(y) = 1$ if $y \in D$, $||X_D||_{L^\infty} \leq 1$, then

$$W_f(a,\theta,b) = a^{-\frac{s+t}{2}}\left(O(\epsilon) + \int_{D_+} X_D(y)h(y)e^{ig(y)}dy\right).$$

If $|\nabla g(y)|$ is not vanishing in D_+, then apply the integral by parts to get

$$\int_{D_+} X_D h e^{ig} dy = \int_{D_+} X_D h (L e^{ig}) dy = -\int_{D_+} \nabla \cdot \left(\frac{X_D h \nabla g}{i|\nabla g|^2}\right) e^{ig} dy.$$

We are going to estimate $|\nabla g(y)|$ when $a \in (\frac{N}{2M}, 2MN)$ and $|\theta_{\nabla\phi(b)} - \theta| \geq \theta_0$. By Taylor expansion,

$$\nabla\phi(b + R_\theta A_a^{-1} y) = \nabla\phi(b) + \nabla^2\phi(b^*) R_\theta A_a^{-1} y,$$

where b^* is between b and $b + R_\theta A_a^{-1} y$. Notice that

$$|\nabla^2\phi(b^*) R_\theta A_a^{-1} y| \leq a^{-s} |\nabla^2\phi(b^*)||y| \lesssim M a^{-s} (\epsilon^{-1/u} + 1) \leq \frac{\sin(\theta_0)}{2M},$$

when $|y| \lesssim \epsilon^{-1/u} + 1$ and $(\frac{2M^2}{\sin(\theta_0)})^{1/s} (\epsilon^{-1/u} + 1)^{1/s} \leq a$. The latter one holds when $N \gtrsim (\epsilon^{-1/u} + 1)^{1/(2s-t)}$ for $a \in (\frac{N}{2M}, 2MN)$. So, when these conditions are satisfied, we have

$$\nabla\phi(b + R_\theta A_a^{-1} y) = \nabla\phi(b) + v,$$

with $|v| \leq \frac{\sin(\theta_0)}{2M}$. Recall the fact $|\theta_{\nabla\phi(b)} - \theta| \geq \theta_0$, then it holds that

$$\begin{aligned}
& |A_a^{-1} R_\theta^{-1} (N \nabla\phi(b + R_\theta A_a^{-1} y) - a \cdot u_\theta)| \\
\geq\ & |N A_a^{-1} R_\theta^{-1} \nabla\phi(b) - (a^{1-t}, 0)^T| - N |A_a^{-1} R_\theta^{-1} v| \\
\geq\ & \sqrt{(r\cos\alpha - a)^2 a^{-2t} + r^2 a^{-2s} \sin^2\alpha} - \frac{N}{2M} \sin\theta_0 a^{-s} \\
\geq\ & r a^{-s} \sin\theta_0 - \frac{N}{2M} \sin\theta_0 a^{-s} \\
\geq\ & \frac{N}{2M} \sin\theta_0 a^{-s} \\
\gtrsim\ & N^{1-t},
\end{aligned}$$

where $\alpha = \theta_{\nabla\phi(b)} - \theta$ and $r = |N\nabla\phi(b)| \geq \frac{N}{M}$. Hence, we have

$$|\nabla g(y)| \gtrsim N^{1-t} \tag{2.12}$$

when $a \in (\frac{N}{2M}, 2MN)$, $|\theta_{\nabla\phi(b)} - \theta| \geq \theta_0$, $N \gtrsim (\epsilon^{-1/u} + 1)^{1/(2s-t)}$ and $y \in D_+$.

Next, we move on to estimate $\left|\frac{\nabla(X_D h)\cdot\nabla g}{|\nabla g|^2}\right|$ and $\left|\frac{\partial^2 g}{\partial y_i \partial y_j} \frac{1}{|\nabla g|^2}\right|$ for $i, j = 1, 2$, under the conditions that $a \in (\frac{N}{2M}, 2MN)$, $|\theta_{\nabla\phi(b)} - \theta| \geq \theta_0$, $N \gtrsim (\epsilon^{-1/u} + 1)^{1/(2s-t)}$ and $y \in D_+$. First,

$$\left|\frac{\partial^2 g}{\partial y_i \partial y_j} \frac{1}{|\nabla g|^2}\right| \leq \frac{N a^{-2s}}{|\nabla g|^2} \leq \frac{N^{1-2s}}{N^{2-2t}} = \frac{1}{N^{1-2(t-s)}}. \tag{2.13}$$

Second, as for $\left|\frac{\nabla(X_D h)\cdot\nabla g}{|\nabla g|^2}\right|$, we only need to estimate $\left|\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2}\right|$ for the similar reason in the last case. As we have shown,

$$\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2} \approx \frac{(Nz_1-a)a^{-2t}z_1 + Na^{-2s}z_2^2}{(Nz_1-a)^2a^{-2t} + N^2a^{-2s}z_2^2}.$$

If $z_1 = 0$, then $\left|\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2}\right| \approx \frac{1}{N}$. If $z_1 \neq 0$ and $\left|\frac{z_2}{z_1}\right| \gtrsim \frac{a^s}{a^t}$, then $|z_2| \gtrsim \frac{a^s}{Ma^t}$, since $|z| \geq \frac{1}{M}$. Hence,

$$\begin{aligned}
\left|\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2}\right| &\lesssim \frac{|(Nz_1-a)a^{-2t}z_1| + |Na^{-2s}z_2^2|}{N^2a^{-2s}z_2^2} \\
&\lesssim \frac{|Nz_1-a|\cdot|z_1|}{N^2a^{2(t-s)}z_2^2} + \frac{1}{N} \\
&\lesssim \frac{1}{Na^{t-s}|z_2|} + \frac{1}{N} \\
&\lesssim \frac{1}{N}.
\end{aligned}$$

If $z_1 \neq 0$ and $\left|\frac{z_2}{z_1}\right| \lesssim \frac{a^s}{a^t}$, then

$$\begin{aligned}
\left|\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2}\right| &\leq \frac{|(Nz_1-a)a^{-2t}z_1| + |Na^{-2s}z_2^2|}{|\nabla g|^2} \\
&\lesssim \frac{(|Nz_1|+a)a^{-2t}|z_1| + Na^{-2t}z_1^2}{N^{2-2t}} \\
&\lesssim \frac{1}{N}.
\end{aligned}$$

In sum,

$$\left|\frac{(A_a^{-1}z)\cdot\nabla g}{|\nabla g|^2}\right| \lesssim \frac{1}{N},$$

which implies that

$$\left|\frac{\nabla(X_D h)\cdot\nabla g}{|\nabla g|^2}\right| \lesssim \frac{1}{N^{1-t}}. \tag{2.14}$$

By (2.13) and (2.14), we have

$$\left|\int_{D_+} \nabla\cdot\left(\frac{X_D h\nabla g}{i|\nabla g|^2}\right)e^{ig}dy\right| \lesssim \left|\nabla\cdot\left(\frac{X_D h\nabla g}{i|\nabla g|^2}\right)\right|\left(||X_D w||_{L^1} + ||\nabla(X_D w)||_{L^1}\right) \lesssim \frac{1}{N^{1-t}}$$

for $a \in (\frac{N}{2M}, 2MN)$, $|\theta_{\nabla\phi(b)} - \theta| \geq \theta_0$ and $N \gtrsim (\epsilon^{-1/u}+1)^{1/(2s-t)}$. So,

$$W_f(a,\theta,b) = a^{-\frac{s+t}{2}}O(\epsilon),$$

when $a \in (\frac{N}{2M}, 2MN)$, $|\theta_{\nabla\phi(b)} - \theta| \geq \theta_0$ and

$$N \gtrsim \max\{(\epsilon^{\frac{-1}{u}} + 1)^{\frac{1}{2s-t}}, \epsilon^{\frac{-1}{1-t}}\}.$$

.

From the discussion in the two cases above, we see that

$$W_f(a, \theta, b) = a^{-\frac{s+t}{2}} O(\epsilon),$$

if $a \notin (\frac{N}{2M}, 2MN)$ or $|\theta_{\nabla\phi(b)} - \theta| \geq \theta_0$, when

$$N \gtrsim \max\{(\epsilon^{\frac{-1}{u}} + 1)^{\frac{1}{2s-t}}, \epsilon^{\frac{-1}{1-t}}\}, \tag{2.15}$$

where u is any fixed positive integer. Hence, the proof of (2) when $K = 1$ is done.

Step2: Henceforth, we move on to prove (1), i.e., to discuss the approximation of $W_f(a, \theta, b)$, when $a \in (\frac{N}{2M}, 2MN)$ and $|\theta_{\nabla\phi(b)} - \theta| < \theta_0$. Recall that

$$W_f(a, \theta, b) \quad = \quad a^{-\frac{s+t}{2}} \bigg(\int_{y \in D} f(b + R_\theta A_a^{-1} y) w(y) e^{-2\pi i a^{1-t} y_1} dy + O(\epsilon) \bigg).$$

Our goal is to get the following estimate

$$W_f(a, \theta, b) = a^{-\frac{s+t}{2}} \bigg(\int_{y \in D} f(b) w(y) e^{2\pi i (N \nabla\phi(b) \cdot (R_\theta A_a^{-1} y) - a^{1-t} y_1)} dy + O(\epsilon) \bigg), \tag{2.16}$$

for N large enough.

First, we are going to show

$$W_f(a, \theta, b) = a^{-\frac{s+t}{2}} \bigg(\int_{y \in D} e^{-\frac{(\phi(b)-c)^2}{\sigma^2}} \alpha(b + R_\theta A_a^{-1} y) w(y) e^{2\pi i (N\phi(b + R_\theta A_a^{-1} y) - a^{1-t} y_1)} dy + O(\epsilon) \bigg) \tag{2.17}$$

for sufficiently large N. Taylor expansion is applied again to obtain the following three expansions.

$$\phi(b + R_\theta A_a^{-1} y) = \phi(b) + \nabla\phi(b) \cdot (R_\theta A_a^{-1} y) + \frac{1}{2} (R_\theta A_a^{-1} y)^T \nabla^2 \phi(b^*) (R_\theta A_a^{-1} y),$$

where b^* is between b and $b + R_\theta A_a^{-1} y$.

$$\begin{aligned} & e^{-(\phi(b + R_\theta A_a^{-1} y) - c)^2 / \sigma^2} \\ = \quad & e^{-(\phi(b) + \nabla\phi(b) \cdot (R_\theta A_a^{-1} y) + \frac{1}{2} (R_\theta A_a^{-1} y)^T \nabla^2 \phi(b^*) (R_\theta A_a^{-1} y) - c)^2 / \sigma^2} \\ = \quad & e^{-\frac{(\phi(b)-c)^2}{\sigma^2}} + e^{-\frac{(\lambda - c)^2}{\sigma^2}} \cdot \frac{-2(\lambda - c)}{\sigma^2} \big(\nabla\phi(b) \cdot (R_\theta A_a^{-1} y) + \frac{1}{2} (R_\theta A_a^{-1} y)^T \nabla^2 \phi(b^*) (R_\theta A_a^{-1} y) \big), \end{aligned}$$

where $\lambda \in [\phi(b), \phi(b) + \nabla\phi(b) \cdot (R_\theta A_a^{-1} y) + \frac{1}{2}(R_\theta A_a^{-1} y)^T \nabla^2 \phi(b^*)(R_\theta A_a^{-1} y)]$.

$$\alpha(b + R_\theta A_a^{-1} y) = \alpha(b) + \nabla\alpha(b^{**}) \cdot (R_\theta A_a^{-1} y),$$

where b^{**} is between b and $b + R_\theta A_a^{-1} y$.

The above Taylor expansions help us to estimate the effect of phase function $\phi(x)$ in the Gaussian term. We claim two estimates as follows.

$$I_1 = \int_{y \in D} \left| e^{-\frac{(\lambda - c)^2}{\sigma^2}} \cdot \frac{-2(\lambda - c)}{\sigma^2} \nabla\phi(b) \cdot (R_\theta A_a^{-1} y) \alpha(b + R_\theta A_a^{-1} y) w(y) \right| dy \leq O(\epsilon)$$

and

$$I_2 = \int_{y \in D} \left| e^{-\frac{(\lambda - c)^2}{\sigma^2}} \cdot \frac{-2(\lambda - c)}{\sigma^2} \frac{1}{2} (R_\theta A_a^{-1} y)^T \nabla^2 \phi(b^*)(R_\theta A_a^{-1} y) \alpha(b + R_\theta A_a^{-1} y) w(y) \right| dy \leq O(\epsilon).$$

Because $e^{-\frac{x^2}{\sigma^2}} \cdot \frac{|x|}{\sigma^2} \leq e^{-\frac{1}{2}} \cdot \frac{1}{\sigma\sqrt{2}}$, we know

$$I_2 \lesssim \frac{1}{\sigma} \int_{y \in D} |y|^2 a^{-2s} dy \lesssim \frac{1}{\sigma} a^{-2s} \epsilon^{-\frac{4}{u}} < \epsilon,$$

if $a \gtrsim \sigma^{-\frac{1}{2s}} \epsilon^{-\frac{1 + \frac{4}{u}}{2s}}$, which is true when

$$N \gtrsim \epsilon^{-\frac{1 + \frac{4}{u}}{2s - \eta}}. \tag{2.18}$$

As for I_1, notice that $|\theta_{\nabla\phi(b)} - \theta| < \theta_0$, then $|\theta_{R_\theta^{-1} \nabla\phi(b)}| < \theta_0$. Let $\tilde{\theta} = \theta_{R_\theta^{-1} \nabla\phi(b)}$ and $y = (y_1, y_2)^T$, then for $a \in (\frac{N}{2M}, 2MN)$

$$\begin{aligned}
I_1 &\lesssim \frac{1}{\sigma} \int_{y \in D} \left| \nabla\phi(b) \cdot (R_\theta A_a^{-1} y) \right| dy \\
&\lesssim \frac{M}{\sigma} \int_{y \in D} \left| \frac{y_1}{a^t} \cos\tilde{\theta} + \frac{y_2}{a^s} \sin\tilde{\theta} \right| dy \\
&\lesssim \frac{Md}{\sigma} \int_{y \in D} \max_{\gamma \in [0, 2\pi)} \left| \frac{\cos\gamma \cos\tilde{\theta}}{a^t} + \frac{\sin\gamma \sin\tilde{\theta}}{a^s} \right| dy \\
&\lesssim \frac{Md^3 L}{\sigma},
\end{aligned}$$

where $d \approx \epsilon^{-\frac{1}{u}}$ is the radius of D and

$$L = \sqrt{\frac{\cos^2\tilde{\theta}}{a^{2t}} + \frac{\sin^2\tilde{\theta}}{a^{2s}}} \leq \sqrt{\frac{1}{a^{2t}} + \frac{\sin^2\theta_0}{a^{2s}}} \lesssim \max\{\frac{1}{a^t}, \frac{|\sin\theta_0|}{a^s}\} \lesssim N^{-t}.$$

So

$$I_1 \lesssim \frac{Md^3L}{\sigma} \lesssim \frac{Md^3N^{-t}}{\sigma} \lesssim O(\epsilon),$$

if

$$N \gtrsim \epsilon^{-\frac{1+\frac{3}{u}}{t-\eta}}. \tag{2.19}$$

A direct result of the estimate of I_1 and I_2 is (2.17) for

$$N \gtrsim \max\{\epsilon^{-\frac{1+\frac{3}{u}}{t-\eta}}, \epsilon^{-\frac{1+\frac{4}{u}}{2s-\eta}}\}. \tag{2.20}$$

Second, we need to show

$$W_f(a,\theta,b) = a^{-\frac{s+t}{2}}\left(\int_{y\in D} e^{-\frac{(\phi(b)-c)^2}{\sigma^2}}\alpha(b)w(y)e^{2\pi i(N\phi(b+R_\theta A_a^{-1}y)-a^{1-t}y_1)}dy + O(\epsilon)\right), \tag{2.21}$$

which relies on the analysis of the effect of $\phi(x)$ on $\alpha(x)$ as follows. Since $a \in (\frac{N}{2M}, 2MN)$, then

$$\begin{aligned} I_3 &= \int_{y\in D} e^{-\frac{(\phi(b)-c)^2}{\sigma^2}} \left|\nabla\alpha\cdot(R_\theta A_a^{-1}y)w(y)\right| dy \\ &\lesssim \int_{y\in D} \left|R_\theta A_a^{-1}y\right| dy \\ &\lesssim a^{-s}\epsilon^{-\frac{3}{u}} \\ &\lesssim O(\epsilon) \end{aligned}$$

holds when

$$N \gtrsim \epsilon^{-\frac{1+\frac{3}{u}}{s}}.$$

Then we derive (2.21) by the estimate of I_3 and (2.17) for $N \gtrsim \epsilon^{-\frac{1+\frac{3}{u}}{s}}$.

Finally, we should estimate the nonlinear effect of $\phi(x)$ on the oscillatory pattern and show (2.16) for sufficiently large N. If

$$N \gtrsim \epsilon^{-\frac{1+\frac{4}{u}}{2s-1}},$$

then

$$\begin{aligned} I_4 &= \int_{y\in D} \left|e^{2\pi i(N\phi(b)+N\nabla\phi(b)\cdot(R_\theta A_a^{-1}y)-a^{1-t}y_1)}\right| \cdot \left|e^{2\pi i\frac{N}{2}(R_\theta A_a^{-1}y)^T\nabla^2\phi(R_\theta A_a^{-1}y)} - 1\right| dy \\ &\lesssim \int_{y\in D} \left|N(R_\theta A_a^{-1}y)^T\nabla^2\phi(R_\theta A_a^{-1}y)\right| dy \\ &\lesssim \int_{y\in D} Na^{-2s}|y|^2 dy \\ &\lesssim Na^{-2s}\epsilon^{-\frac{4}{u}} \\ &\lesssim O(\epsilon) \end{aligned}$$

holds by the fact that $|e^{ix}-1| \le |x|$ and $a \in (\frac{N}{2M}, 2MN)$. Then by (2.21) and I_4, we have

$$\begin{aligned} W_f(a,\theta,b) &= a^{-\frac{s+t}{2}}\left(f(b)\int_{y\in D} w(y)e^{2\pi i(N\nabla\phi(b)\cdot(R_\theta A_a^{-1}y)-a^{1-t}y_1)}dy + O(\epsilon)\right) \\ &= a^{-\frac{s+t}{2}}\left(f(b)\int_{R^2} w(y)e^{2\pi i(NA_a^{-1}R_\theta^{-1}\nabla\phi(b)-(a^{1-t},0)^T)\cdot y}dy + O(\epsilon)\right) \\ &= a^{-\frac{s+t}{2}}\left(f(b)\widehat{w}\big(A_a^{-1}R_\theta^{-1}(a\cdot u_\theta - N\nabla\phi(b))\big) + O(\epsilon)\right), \end{aligned}$$

for $a \in (\frac{N}{2M}, 2MN)$ and $|\theta_{\nabla\phi(b)} - \theta| < \theta_0$, if

$$N \gtrsim \max\{\epsilon^{-\frac{1+\frac{3}{u}}{t-\eta}}, \epsilon^{-\frac{1+\frac{4}{u}}{2s-\eta}}, \epsilon^{-\frac{1+\frac{3}{u}}{s}}, \epsilon^{-\frac{1+\frac{4}{u}}{2s-1}}\}, \tag{2.22}$$

where u is any fixed positive integer.

By (2.15) and (2.22), the requirement for N is

$$N \gtrsim N_0 = \max\{(\epsilon^{\frac{-1}{u}}+1)^{\frac{1}{2s-t}}, \epsilon^{\frac{-1}{1-t}}, \epsilon^{-\frac{1+\frac{3}{u}}{t-\eta}}, \epsilon^{-\frac{1+\frac{4}{u}}{2s-\eta}}, \epsilon^{-\frac{1+\frac{3}{u}}{s}}, \epsilon^{-\frac{1+\frac{4}{u}}{2s-1}}\},$$

where u is any fixed positive integer. Hence, this completes the proof of (1) when $K = 1$.

In sum, we have proved this lemma when $K = 1$. The conclusion is also true for general K by the linearity of generalized curvelet transform. □

Lemma 2.3.9. *Suppose*

$$\Omega_{a\theta b} = \left\{k : a \in \left(\frac{N_k}{2M}, 2MN_k\right), \left|\theta_{\nabla\phi_k(b)} - \theta\right| < \arcsin\left(\left(\frac{M}{N_k}\right)^{t-s}\right)\right\}$$

is not empty. Under the assumption of Theorem 2.3.7, there exists a constant $N_0(M,m,s,t,\eta,\epsilon) \simeq \max\left\{\epsilon^{\frac{-1}{1-t}}, \epsilon^{\frac{-2}{t-\eta}}, \epsilon^{\frac{-2}{2s-1}}\right\}$ *such that if* $N > N_0(M,m,s,t,\eta,\epsilon)$*, then we have*

$$\nabla_b W_f(a,\theta,b) = a^{-\frac{s+t}{2}}\left(2\pi i\sum_{k\in\Omega_{a\theta b}} N_k\nabla\phi_k(b)f_k(b)\widehat{w}\left(A_a^{-1}R_\theta^{-1}\left(a\cdot u_\theta - N_k\nabla\phi(b)\right)\right) + aO(\epsilon)\right).$$

Proof. The proof is similar to the one of Lemma 2.3.8. We only need to discuss the case $K = 1$ and the case $K > 1$ holds by the linearity of generalized curvelet transform. Suppose

$$f(x) = e^{-\frac{(\phi(x)-c)^2}{\sigma^2}}\alpha(x)e^{2\pi iN\phi(x)},$$

we have

$$\begin{aligned}
& \nabla_b W_f(a,\theta,b) \\
= & \int_{\mathbb{R}^2} f(x) a^{\frac{s+t}{2}} \Big((-R_\theta A_a) \nabla w(A_a R_\theta^{-1}(x-b)) + 2\pi i a u_\theta w(A_a R_\theta^{-1}(x-b)) \Big) e^{-2\pi i a (x-b)\cdot u_\theta} dx \\
= & \int_{\mathbb{R}^2} f(b + R_\theta A_a^{-1} y) a^{-\frac{s+t}{2}} \Big((-R_\theta A_a) \nabla w(y) + 2\pi i a e_\theta w(y) \Big) e^{-2\pi i a^{1-t} y_1} dy \\
= & a^{-\frac{s+t}{2}} \Big(f(b) \int_{\mathbb{R}^2} \big((-R_\theta A_a) \nabla w(y) + 2\pi i a e_\theta w(y) \big) e^{-2\pi i ((a^{1-t},0)^T - N A_a^{-1} R_\theta^{-1} \nabla\phi(b))\cdot y} dy + a O(\epsilon) \Big) \\
= & a^{-\frac{s+t}{2}} \Big(2\pi i N \nabla\phi(b) f(b) \widehat{w}\left(A_a^{-1} R_\theta^{-1} (a\cdot u_\theta - N\nabla\phi(b)) \right) + a O(\epsilon) \Big)
\end{aligned}$$

for $a \in (\frac{N}{2M}, 2MN)$ and $|\theta_{\nabla\phi(b)} - \theta| < \theta_0$, if N satisfies the condition in Lemma 2.3.8. Therefore, if f has K components, we know

$$\nabla_b W_f(a,\theta,b) = a^{-\frac{s+t}{2}} \Big(\sum_{k\in\Omega_{a\theta b}} 2\pi i N_k \nabla\phi_k(b) f_k(b) \widehat{w}\left(A_a^{-1} R_\theta^{-1} (a\cdot u_\theta - N_k \nabla\phi_k(b)) \right) + a O(\epsilon) \Big),$$

for N larger than the same constant N_0 in Lemma 2.3.8. □

With the above two lemmas proved, it is enough to prove Theorem 2.3.7.

Proof. We shall start from (i). $\{Z_k : 1 \le k \le K\}$ are disjoint as soon as $f(x)$ is a superposition of well-separated components. Let $(a,\theta,b) \in R_\epsilon$. By Lemma 2.3.8, we have

$$W_f(a,\theta,b) = a^{-\frac{s+t}{2}} \left(\sum_{k\in\Omega_{a\theta b}} f_k(b) \widehat{w}\left(A_a^{-1} R_\theta^{-1} (a\cdot u_\theta - N_k \nabla\phi_k(b)) \right) + O(\epsilon) \right).$$

Therefore, $\exists k$ such that $\widehat{w}\left(A_a^{-1} R_\theta^{-1} (a\cdot u_\theta - N_k \nabla\phi_k(b)) \right) \neq 0$. By the definition of Z_k, we see that $(a,\theta,b) \in Z_k$. Hence, $R_\epsilon \subset \cup_{k=1}^K Z_k$. It's obvious that $S_\epsilon \subset R_\epsilon$.

To show (ii), notice that $(a,\theta,b) \in R_\epsilon \cup Z_k$, then

$$W_f(a,\theta,b) = a^{-\frac{s+t}{2}} \Big(f_k(b) \widehat{w}\left(A_a^{-1} R_\theta^{-1} (a\cdot u_\theta - N_k \nabla\phi_k(b)) \right) + O(\epsilon) \Big),$$

and

$$\nabla_b W_f(a,\theta,b) = a^{-\frac{s+t}{2}} \Big(2\pi i N_k \nabla\phi_k(b) f_k(b) \widehat{w}\left(A_a^{-1} R_\theta^{-1} (a\cdot u_\theta - N_k \nabla\phi(b)) \right) + a O(\epsilon) \Big).$$

Let $g = f_k(b) \widehat{w}\left(A_a^{-1} R_\theta^{-1} (a\cdot u_\theta - N_k \nabla\phi(b)) \right)$, then

$$v_f(a,\theta,b) = \frac{N_k \nabla\phi_k(b)(g + O(\epsilon))}{g + O(\epsilon)}.$$

Since $|W_f(a,\theta,b)| \geq a^{-\frac{s+t}{2}}\sqrt{\epsilon}$ for $(a,\theta,b) \in R_\epsilon$, then $|g| \gtrsim \sqrt{\epsilon}$. So

$$\frac{|v_f(a,\theta,b) - N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|} \lesssim \left|\frac{O(\epsilon)}{g+O(\epsilon)}\right| \lesssim \sqrt{\epsilon}.$$

Similarly, if $(a,\theta,b) \in S_\epsilon \cap Z_k$, then

$$\frac{|v_f(a,\theta,b) - N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|} \lesssim \left|\frac{O(\epsilon)}{g+O(\epsilon)}\right| \lesssim \frac{\sqrt{\epsilon}}{N_k^{(s+t)/2}},$$

since $|g| \gtrsim N_k^{(s+t)/2}\sqrt{\epsilon}$ for $(a,\theta,b) \in S_\epsilon \cap Z_k$. □

The assumption $\frac{1}{2} < s < t < 1$ and $\eta < t$ are essential to the proof. However, we have not arrived to a clear opinion on the optimal values of these parameters. The difference $t - s$ allows us to construct directional needle-like curvelets in order to approximate banded wave-like components or wavefronts and capture the oscillatory behavior better. When t and η approach to 1, and s gets close to $\frac{1}{2}$, we can expect that the synchrosqueezed curvelet transform can separate banded components of width approximately $O(N^{-1})$, if N is large enough. On the other hand, the lower bound $s > 1/2$ ensures that the support of each curvelet is sufficiently small in space so that the second order properties of the phase function (such as the curvature of wavefronts) do not affect the estimate of local wave-vectors. The upper bound $t < 1$ guarantees sufficient resolution to detect different components with large wavenumbers.

Chapter 3

Robustness of Synchrosqueezed Transforms

3.1 Introduction

3.1.1 Motivation

In this chapter, we will focus on the robustness analysis of synchrosqueezed transform (joint work with Lexing Ying in [183]) on signals with a noisy perturbation term $e(x)$:

$$f(x) = \sum_{k=1}^{K} \alpha_k(x) e^{2\pi i N_k \phi_k(x)} + e(x). \tag{3.1}$$

It follows from the definition that $\alpha_k(x)e^{2\pi i N_k \phi_k(x)}$ is a highly oscillatory component with a frequency content also rapidly changing with x. An immediate challenge from this rapid change is that an instantaneous frequency or the magnitude of a local wave vector may quickly increase to the sampling rate, e.g., power-law chirps in gravitational waves [29]. Another challenge comes from a large number of different N_k in various scales caused by the phenomenon of wave shape functions [170, 178, 180] or equivalently intrawaves [91, 176]. When noise meets these multiscale oscillatory components, an efficient and accurate tool with multiscale robustness to identify and analyze these wave-like components is of great value.

A variety of synchrosqueezed transforms have been proposed to study signal (3.1), e.g., the synchrosqueezed wavelet transform (SSWT) in [37, 43], the synchrosqueezed short time Fourier transform (SSSTFT) in [156], the synchrosqueezed wave packet transform (SSWPT) in [178, 182] and the synchrosqueezed curvelet transform (SSCT) in [184]. Rigorous analysis has proved that these transforms can accurately decompose a class of superpositions of wave-like components and

estimate their instantaneous (local) properties if the given signal is noiseless.

Although the literature on synchrosqueezed transforms for noiseless data is well developed and they have been successfully applied to various real problems with noisy data, rigorous robustness analysis of these transforms is still limited. In a recent paper [155] that addresses the robustness analysis, it is assumed that (3.1) contains only Gaussian white noise with a variance much smaller than ϵ^2, where $\epsilon \ll 1$ is the error tolerance of the estimation accuracy in [43]. This requirement is too restricted in real applications. To deal with heavier noise, a recent paper [32] proves the robustness against a generalized stationary Gaussian noise and analyzes statistical properties of (3.1) when it has a trend with heteroscedasticity. However, this proof is valid only for the 1D SSWT in [43] in analyzing wave-like components with instantaneous frequencies of constant order.

In the analysis of 1D SSWT [32, 43, 155, 156], the authors are assuming a class of well-separated superpositions of intrinsic mode type functions. If we rephrase the definition of the 1D SSWT in [32, 43, 155] using a statement convenient for multiscale analysis, then this class of well-separated superpositions can be defined through the following definitions.

Definition 3.1.1. *(An intrinsic mode type function for the 1D SSWT). A continuous function $f : \mathbb{R} \to \mathbb{C}$, $f \in L^\infty(\mathbb{R})$ is said to be intrinsic-mode-type (IMT) with accuracy $\epsilon > 0$ if $f(x) = \alpha(x)e^{2\pi i N\phi(x)}$ with $\alpha(x)$ and $\phi(x)$ having the following properties:*

$$\alpha \in C^1(\mathbb{R}) \cap L^\infty(\mathbb{R}), \quad \phi \in C^2(\mathbb{R})$$
$$\inf_{x\in\mathbb{R}} \phi'(x) > 0, \quad \sup_{x\in\mathbb{R}} \phi'(x) < \infty, \quad \sup_{x\in\mathbb{R}} |\phi''(x)| < \infty,$$
$$|\alpha'(x)| \le \epsilon |N\phi'(x)|, |\phi''(x)| \le \epsilon |\phi'(x)|, \quad \forall x \in \mathbb{R}.$$

To guarantee accurate estimates of nonlinear wave-like components provided by the SSWT, the approach in [43] needs to assume N to be sufficiently small. To make things concrete, consider the requirement of Equation (3.5) in [43], which reads $\epsilon < N^{-3/2}$ in the language of this paper. For example, if $\epsilon = 0.01$, then N has to be less than 21.5. Since larger ϵ allows stronger nonlinearity in Definition 3.1.1, Equation (3.5) says that high frequency wave-like components have to be nearly linear, which is impractical for a superposition of multiscale nonlinear wave-like components. Indeed, multiscale components are common in nature, which motivates the work in [178, 182, 184] and this thesis.

Using our new language convenient for multiscale analysis for synchrosqueezed transforms, we shall provide rigorous probability analysis for their multiscale robustness with different geometric scaling. It will be shown that a trade-off between the multiscale robustness and the estimation accuracy has to be balanced.

3.1.2 Significance

Analyzing signals in (3.1) is also called a mode decomposition problem. A famous empirical mode decomposition (EMD) method has initialized a very active research line in advanced and adaptive data analysis. This method was first proposed by Huang et al. in [97] and refined in [98]. It has good numerical performance in decomposing a class of superpositions of oscillatory components and has been widely used in various applications, even though the mathematics behind this method is still unknown. However, the good properties of the EMD method are fragile. It is well known that EMD methods are not robust against noise. Therefore, synchrosqueezed transforms with well-developed mathematical background and reasonable robustness are important alternatives. This is illustrated in a recent review [153] by comparing several advanced tools for spectral estimations, e.g., the EMD method, the short-time Fourier transform, the SSWT, some basics pursuit method and some matching pursuit method. We expect synchrosqueezed transforms can provide new insights for oscillatory component analysis to help us understand the nature, since in some cases the EMD method would give misleading results [178, 182].

Statistical literature on oscillatory estimation is well developed, but a multiscale oscillatory estimation with a possible trend is perhaps more recent. Some existing models, e.g., the seasonal auto-regressive integrated moving average [16] and the trend and seasonal components algorithm [46], focus on forecasting. They might not be suitable for time-varying historical components as discussed in [32]. Some methods are based on a global assumption with precise known properties of the signal and perform a generalized likelihood ratio test. Global assumptions could be too restrictive in analyzing local information hidden in a general time-varying component. The resulting statistics might be sensitive to the length of the given signal. Some models are fully non-parametric and local in nature. They can even detect an oscillatory component from totally unknown and fully noisy data via the chirplet transform and path pursuit [20]. However, they are focusing on detecting and analyzing only one oscillatory component and cannot be applied to more complex data. Hence, the non-parametric robust analysis tool for multiscale components discussed in this thesis is new and adaptive to a general problem.

The rest of this chapter is organized as follows. Sections 3.2 to 3.4 present the main theorems for the 1D synchrosqueezed wave packet transform, the 2D synchrosqueezed wave packet transform and the 2D synchrosqueezed curvelet transform, respectively. In each of these sections, the robustness of synchrosqueezed transforms to bounded perturbation and generalized stationary Gaussian noise is analyzed. These theorems can be generalized to show the robustness of higher dimensional synchrosqueezed transforms.

3.2 1D Synchrosqueezed Wave Packet Transform (SSWPT)

Theorem 2.1.8 shows that the instantaneous frequency information function $v_f(a,b)$ can estimate $N_k\phi_k'(x)$ accurately for a class of noiseless superpositions of IMTs if their phases are sufficiently steep. This guarantees the well concentration of the synchrosqueezed energy distribution $T_f(v,b)$ around $N_k\phi_k'(x)$. If the superposition is perturbed slightly by a contaminant, Theorem 3.2.1 below shows that these conclusions are still valid with a reasonable error determined by the magnitude of the perturbation.

In what follows, when we write $O(\cdot)$, $\lesssim$, or $\gtrsim$, the implicit constants may depend on M, m and K.

Theorem 3.2.1. *Suppose the mother wave packet is of type (ϵ, m), for any fixed $\epsilon \in (0,1)$ and any fixed integer $m \geq 0$. Suppose $g(x) = f(x) + e(x)$, where $e(x) \in L^\infty$ is a small error term that satisfies $\|e\|_{L^\infty} \leq \sqrt{\epsilon_1}$ for some $\epsilon_1 > 0$. For any $p \in \left(0, \frac{1}{2}\right]$, let $\delta = \sqrt{\epsilon} + \epsilon_1^{\frac{1}{2}-p}$. Define*

$$R_\delta = \{(a,b) : |W_g(a,b)| \geq |a|^{-s/2}\delta\},$$

$$S_\delta = \{(a,b) : |W_g(a,b)| \geq \delta\},$$

and

$$Z_k = \{(a,b) : |a - N_k\phi_k'(b)| \leq |a|^s\}$$

for $1 \leq k \leq K$. For fixed M, m and K, there exists a constant $N_0(M,m,K,s,\epsilon) \simeq \max\left\{\epsilon^{\frac{-1}{2s-1}}, \epsilon^{\frac{-1}{1-s}}\right\}$ such that for any $N > N_0(M,m,K,s,\epsilon)$ and $f(x) \in F(M,N,K,s)$ the following statements hold.

(i) $\{Z_k : 1 \leq k \leq K\}$ are disjoint and $S_\delta \subset R_\delta \subset \bigcup_{1\leq k\leq K} Z_k$;

(ii) For any $(a,b) \in R_\delta \cap Z_k$,

$$\frac{|v_g(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \lesssim \sqrt{\epsilon} + \epsilon_1^p;$$

(iii) For any $(a,b) \in S_\delta \cap Z_k$,

$$\frac{|v_g(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \lesssim \frac{\sqrt{\epsilon} + \epsilon_1^p}{N_k^{s/2}}.$$

We introduce the parameter p to clarify the relation among the perturbation level, the threshold and the accuracy for better understanding the influence of perturbation or noise. For the same purpose, a parameter q will be introduced in the coming theorems. Theorem 3.2.1 shows that the instantaneous frequency estimates provided by the SSWPT are still reasonable when the given signal is contaminated by a bounded perturbation. Actually, if the threshold δ is larger, e.g., $\delta \geq \sqrt{\frac{\epsilon_1}{\epsilon}}$, the relative estimate errors in (ii) and (iii) are bounded by $\sqrt{\epsilon}$ and $\frac{\sqrt{\epsilon}}{N_k^{s/2}}$, respectively. This also implies that the instantaneous frequency can be better estimated by selecting the wave packet coefficient with the largest magnitude. However, when the perturbation is overwhelming, e.g., the wave packet

coefficients of a component are below the threshold in (ii), it is difficult to estimate instantaneous frequencies.

Proof. We only need to discuss the case when $a > 0$. We estimate several inequalities first. By the definition of the wave packet transform of $e(x)$, we have

$$W_e(a,b) \quad = \quad a^{-s/2}\int_{\mathbb{R}} e\left(a^{-s}y+b\right)w(y)e^{-2\pi i a^{1-s}y}dy.$$

Hence,

$$|W_e(a,b)| \lesssim \|e\|_{L^\infty} a^{-s/2} \leq \sqrt{\epsilon_1}a^{-s/2}. \tag{3.2}$$

Applying the same strategy, we have

$$|\partial_b W_e(a,b)| \lesssim \sqrt{\epsilon_1}\left(a^{1-s/2}+a^{s/2}\right). \tag{3.3}$$

If $(a,b) \in R_\delta$, then $|W_g(a,b)| \geq a^{-s/2}\delta$. Together with Equation (3.2), it holds that

$$|W_g(a,b)| \geq |W_f(a,b)| - |W_e(a,b)| \geq a^{-s/2}\left(\delta - \sqrt{\epsilon_1}\right) \geq a^{-s/2}\sqrt{\epsilon}. \tag{3.4}$$

Hence, $S_\delta \subset R_\delta \subset R_\epsilon$, where R_ϵ is defined in Theorem 2.1.8 and is a subset of $\bigcup_{1\leq k\leq K} Z_k$. So, (i) is true by Theorem 2.1.8.

Now, let us prove (ii). Since $R_\delta \subset R_\epsilon$, $(a,b) \in R_\delta \cap Z_k$ implies $(a,b) \in R_\epsilon \cap Z_k$. Hence, by Theorem 2.1.8, it holds that

$$\frac{|v_f(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \lesssim \sqrt{\epsilon}, \tag{3.5}$$

when N is larger than a constant $N_0\left(M,m,K,s,\epsilon\right) \simeq \max\left\{\epsilon^{\frac{-1}{2s-1}}, \epsilon^{\frac{-1}{1-s}}\right\}$. Notice that $(a,b) \in Z_k$ implies $a \simeq N_k$. Hence, by Equation (3.2) to (3.5),

$$\begin{aligned}
& \frac{|v_g(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \\
\leq \quad & \frac{|v_f(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} + \frac{|\frac{\partial_b W_f(a,b)}{2\pi i W_f(a,b)} - \frac{\partial_b W_g(a,b)}{2\pi i W_g(a,b)}|}{|N_k\phi_k'(b)|} \\
\lesssim \quad & \sqrt{\epsilon} + \left|\frac{\partial_b W_f(a,b)W_e(a,b) - \partial_b W_e(a,b)W_f(a,b)}{W_f(a,b)W_g(a,b)}\right| \frac{1}{|N_k\phi_k'(b)|} \\
\lesssim \quad & \sqrt{\epsilon} + \left|\frac{W_e(a,b)}{W_g(a,b)}\right| + \left|\frac{\partial_b W_e(a,b)}{N_k W_g(a,b)}\right| \\
\lesssim \quad & \sqrt{\epsilon} + \frac{\sqrt{\epsilon_1}}{\delta} + \frac{\sqrt{\epsilon_1}\left(a^{1-s/2}+a^{s/2}\right)}{N_k\delta a^{-s/2}} \\
\lesssim \quad & \sqrt{\epsilon} + \frac{\sqrt{\epsilon_1}}{\delta} \\
= \quad & \sqrt{\epsilon} + \epsilon_1^p,
\end{aligned}$$

when $N > N_0$. Hence, (ii) is proved. The proof of (iii) is similar. If $(a,b) \in S_\delta \cap Z_k$, then

$$\begin{aligned}
& \frac{|v_g(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \\
\leq\ & \frac{|v_f(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} + \frac{|\frac{\partial_b W_f(a,b)}{2\pi i W_f(a,b)} - \frac{\partial_b W_g(a,b)}{2\pi i W_g(a,b)}|}{|N_k\phi_k'(b)|} \\
\lesssim\ & \frac{\sqrt{\epsilon}}{N_k^{s/2}} + \frac{\sqrt{\epsilon_1}}{a^{s/2}\delta} + \frac{\sqrt{\epsilon_1}\left(a^{1-s/2} + a^{s/2}\right)}{N_k\delta} \\
\lesssim\ & \frac{\sqrt{\epsilon}}{N_k^{s/2}} + \frac{\sqrt{\epsilon_1}}{a^{s/2}\delta} \\
\lesssim\ & \frac{\sqrt{\epsilon} + \epsilon_1^p}{N_k^{s/2}},
\end{aligned}$$

when $N > N_0$. □

Next, we will analyze the robustness properties of the SSWPT in the presence of random perturbation. [76, 114, 134, 154, 161] are referred to for basic facts about generalized random fields and complex Gaussian processes that are used throughout this chapter. To warm up, we start with additive Gaussian white noise in Theorem 3.2.2 and extend it to a general zero mean stationary Gaussian noise in Theorem 3.2.3. Let n be the dimension of given data. $n = 1$ in this section and $n = 2$ in later sections. If we fix a probability space $(\mathbb{R}^n, \mu)$ and assume that $L^2(\mathbb{R}^n, \mu)$ is separable, a stationary Gaussian process e on $\mathbb{R}^n$ is an $L^2(\mathbb{R}^n, \mu)$-valued distribution [154], i.e., a continuous linear functional in $\mathcal{D}'(\mathbb{R}^n, L^2(\mathbb{R}^n, \mu))$ such that

$$e : C_0^\infty(\mathbb{R}^n) \to L^2(\mathbb{R}^n, \mu),$$

which can be continuously extended to

$$e : L^1 \cap C^r(\mathbb{R}^n) \to L^2(\mathbb{R}^n, \mu),$$

for some $r \in \mathbb{N}$ or $r = \infty$ depending on e. We assume that r is small enough such that the family of wave packets we constructed and their derivatives belong to $C^r(\mathbb{R}^n)$. Suppose e has a mean functional $\mathcal{T} : L^1 \cap C^r(\mathbb{R}^n) \to L^1 \cap C^r(\mathbb{R}^n)$ and a covariance functional $\mathcal{R} : L^1 \cap C^r(\mathbb{R}^n) \to L^1 \cap C^r(\mathbb{R}^n)$, then we have

1. For any finite collection $\{f_k\} \subset L^1 \cap C^r(\mathbb{R}^n)$, $\{e(f_k)\}$ are jointly Gaussian variables and their joint distribution is translation invariant for all translates of f_k;
2. $\mathbb{E}[e(f)] = \mathcal{T}f$ and $\mathbb{E}\left[e(f_1)\overline{e(f_2)}\right] = \langle f_1, \mathcal{R}f_2\rangle$, where $\langle\cdot,\cdot\rangle$ is the L^2 inner product.

Gaussian white noise is a special case of stationary Gaussian processes with $\mathcal{T} = 0$ and $\mathcal{R}$ being the identical functional. For the convenience of notations, for any wave packet $w_{ab}(x)$, $e(w_{ab})$ and

$e(\partial_b w_{ab})$ are denoted as $W_e(a,b)$ and $\partial_b W_e(a,b)$, respectively. We assume that e has an explicit power spectral function denoted by $\widehat{e}(\xi)$. $\|\cdot\|$ will represent the L^2 norm.

Theorem 3.2.2. *Suppose the mother wave packet is of type (ϵ, m), for any fixed $\epsilon \in (0,1)$ and any fixed integer $m \geq \frac{2}{1-s}+4$. Suppose $g(x) = f(x)+e$, where e is zero mean Gaussian white noise with a variance ϵ_1^{1+q} for some $q>0$ and some $\epsilon_1 > 0$. For any $p \in \left(0, \frac{1}{2}\right]$, let $\delta = \sqrt{\epsilon} + \epsilon_1^{\frac{1}{2}-p}$. Define*

$$R_\delta = \{(a,b) : |W_g(a,b)| \geq a^{-s/2}\delta\},$$

$$S_\delta = \{(a,b) : |W_g(a,b)| \geq \delta\},$$

and

$$Z_k = \{(a,b) : |a - N_k \phi_k'(b)| \leq a^s\}$$

for $1 \leq k \leq K$. For fixed M and K, there exists a constant $N_0(M,m,K,s,\epsilon) \simeq \max\left\{\epsilon^{\frac{-1}{2s-1}}, \epsilon^{\frac{-1}{1-s}}\right\}$ such that for any $N > N_0(M,m,K,s,\epsilon)$ and $f(x) \in F(M,N,K,s)$ the following statements hold.

(i) $\{Z_k : 1 \leq k \leq K\}$ are disjoint.

(ii) If $(a,b) \in R_\delta$, then $(a,b) \in \bigcup_{1\leq k\leq K} Z_k$ with a probability at least

$$1 - e^{-O\left(N_k^{-s}\epsilon_1^{-q}\right)} + O\left(\frac{\epsilon}{N_k^{m(1-s)}}\right).$$

(iii) If $(a,b) \in S_\delta$, then $(a,b) \in \bigcup_{1\leq k\leq K} Z_k$ with a probability at least

$$1 - e^{-\epsilon_1^{-q}\|w\|^{-2}} + O\left(\frac{\epsilon}{N_k^{m(1-s)}}\right).$$

(iv) If $(a,b) \in R_\delta \cap Z_k$ for some k, then

$$\frac{|v_g(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \lesssim \sqrt{\epsilon} + \epsilon_1^p$$

is true with a probability at least

$$\left(1 - e^{-O\left(N_k^{2-3s}\epsilon_1^{-q}\right)}\right)\left(1 - e^{-O\left(N_k^{-s-2}\epsilon_1^{-q}\right)}\right) + O\left(\frac{\epsilon}{N_k^{(m-4)(1-s)-2}}\right).$$

(v) If $(a,b) \in S_\delta \cap Z_k$, then

$$\frac{|v_g(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \lesssim \frac{\sqrt{\epsilon} + \epsilon_1^p}{N_k^{s/2}}$$

is true with a probability at least

$$\left(1-e^{-O\left(N_k^{2-2s}\epsilon_1^{-q}\right)}\right)\left(1-e^{-O\left(N_k^{-2}\epsilon_1^{-q}\right)}\right)+O\left(\frac{\epsilon}{N_k^{(m-4)(1-s)-2}}\right).$$

Proof. Step 1: we prove this theorem when the mother wave packet is of type $(0,m)$ first, i.e., compactly supported in the frequency domain.

Since w_{ab} and $\partial_b w_{ab}$ are in $L^1\cap C^{m-1}$, $W_e(a,b)$ and $\partial_b W_e(a,b)$ are Gaussian variables. Hence, $W_g(a,b)=W_f(a,b)+W_e(a,b)$ and $\partial_b W_g(a,b)=\partial_b W_f(a,b)+\partial_b W_e(a,b)$ can be understood as Gaussian variables. Furthermore, $W_e(a,b)$ and $(W_e(a,b),\partial_b W_e(a,b))$ are circularly symmetric Gaussian variables by checking that their pseudo-covariance matrices are zero. By the properties of Gaussian white noise, we have $\mathbb{E}\left[W_e(a,b)\right]=0$,

$$\mathbb{E}\left[W_e(a,b)\overline{W_e(a,b)}\right]=\epsilon_1^{1+q}\int_{\mathbb{R}}a^s w\left(a^s(x-b)\right)\overline{w\left(a^s(x-b)\right)}dx=\epsilon_1^{1+q}\|w\|^2,$$

and

$$\mathbb{E}\left[W_e(a,b)W_e(a,b)\right]=\epsilon_1^{1+q}\langle w_{ab},\overline{w_{ab}}\rangle=\epsilon_1^{1+q}\langle\widehat{w_{ab}},\widehat{\overline{w_{ab}}}\rangle=\epsilon_1^{1+q}\int_{\mathbb{R}}\widehat{w_{ab}}(\xi)\widehat{w_{ab}}(-\xi)d\xi=0.$$

The last equality holds because $\operatorname{supp}\left(\widehat{w_{ab}(\xi)}\right)\cap\operatorname{supp}\left(\widehat{w_{ab}(-\xi)}\right)=\emptyset$. Similarly, we know

$$\mathbb{E}\left[\partial_b W_e(a,b)\right]=\mathbb{E}\left[\left(\partial_b W_e(a,b)\right)^2\right]=\mathbb{E}\left[\partial_b W_e(a,b)W_e(a,b)\right]=0,$$

$$\mathbb{E}\left[\partial_b W_e(a,b)\overline{\partial_b W_e(a,b)}\right]=\epsilon_1^{1+q}\langle\partial_b w_{ab},\partial_b w_{ab}\rangle=\epsilon_1^{1+q}\langle 2\pi i\xi\widehat{w_{ab}},2\pi i\xi\widehat{w_{ab}}\rangle$$

and

$$\mathbb{E}\left[W_e(a,b)\overline{\partial_b W_e(a,b)}\right]=\epsilon_1^{1+p}\langle w_{ab},\partial_b w_{ab}\rangle=\epsilon_1^{1+q}\langle\widehat{w_{ab}},2\pi i\xi\widehat{w_{ab}}\rangle.$$

Hence, $W_e(a,b)$ and $(W_e(a,b),\partial_b W_e(a,b))$ have zero pseudo-covariance matrices and they are circularly symmetric. Therefore, the distribution of $W_e(a,b)$ is determined by its variance as follows

$$\frac{e^{-\epsilon_1^{-(1+q)}|z_1|^2\|w\|^{-2}}}{\pi\epsilon_1^{1+q}\|w\|^2}.$$

If we define

$$V=\begin{pmatrix}\|\widehat{w}\|^2 & \langle\widehat{w_{ab}},2\pi i\xi\widehat{w_{ab}}\rangle\\ \langle 2\pi i\xi\widehat{w_{ab}},\widehat{w_{ab}}\rangle & \langle 2\pi i\xi\widehat{w_{ab}},2\pi i\xi\widehat{w_{ab}}\rangle\end{pmatrix},$$

then $\epsilon_1^{1+q}V$ is the covariance matrix of $(W_e(a,b),\partial_b W_e(a,b))$ and its distribution is described by the

joint probability density

$$\frac{e^{-\epsilon_1^{-(1+q)} z^* V^{-1} z}}{\pi^2 \epsilon_1^{2(1+q)} \det V},$$

where $z = (z_1, z_2)^T$, T and $*$ denote the transpose operator and conjugate transpose operator. V is an invertible and self-adjoint matrix, since $W_e(a,b)$ and $\partial_b W_e(a,b)$ are linearly independent. Hence, there exist a diagonal matrix D and a unitary matrix U such that $V^{-1} = U^* D U$.

Part (i) is true by previous theorems. Define the following events

$$G_1 = \left\{ |W_e(a,b)| < a^{-s/2} \sqrt{\epsilon_1} \right\},$$

$$G_2 = \left\{ |W_e(a,b)| < \sqrt{\epsilon_1} \right\},$$

$$G_3 = \left\{ |\partial_b W_e(a,b)| < \sqrt{\epsilon_1} \left(a^{1-s/2} + a^{s/2} \right) \right\},$$

$$H_k = \left\{ \frac{|v_g(a,b) - N_k \phi_k'(b)|}{|N_k \phi_k'(b)|} \lesssim \sqrt{\epsilon} + \epsilon_1^p \right\},$$

and

$$J_k = \left\{ \frac{|v_g(a,b) - N_k \phi_k'(b)|}{|N_k \phi_k'(b)|} \lesssim \frac{\sqrt{\epsilon} + \epsilon_1^p}{N_k^{s/2}} \right\},$$

for $1 \le k \le K$. To conclude Part (ii) to (v), we need to estimate the probability $P(G_1)$, $P(G_2)$, $P(G_1 \cap G_3)$, $P(G_2 \cap G_3)$, $P(H_k)$ and $P(J_k)$. By the calculations above, we have

$$\begin{aligned} P(G_1) &= \int_{|z_1| < a^{-s/2}\sqrt{\epsilon_1}} \frac{e^{-\epsilon_1^{-(1+q)} |z_1|^2 \|w\|^{-2}}}{\pi \epsilon_1^{1+q} \|w\|^2} dz_1 \\ &= \frac{2}{\epsilon_1^{1+q} \|w\|^2} \int_0^{a^{-s/2}\sqrt{\epsilon_1}} r e^{-\epsilon_1^{-(1+q)} r^2 \|w\|^{-2}} dr \\ &= \int_0^{a^{-s/2} \epsilon_1^{-q/2} \|w\|^{-1}} 2 r e^{-r^2} dr \\ &= 1 - e^{-a^{-s} \epsilon_1^{-q} \|w\|^{-2}}, \end{aligned}$$

and similarly

$$P(G_2) = \int_{|z_1| < \sqrt{\epsilon_1}} \frac{e^{-\epsilon_1^{-(1+q)} |z_1|^2 \|w\|^{-2}}}{\pi \epsilon_1^{1+q} \|w\|^2} dz_1 = 1 - e^{-\epsilon_1^{-q} \|w\|^{-2}}.$$

We are ready to summarize and conclude (ii) and (iii). If $(a,b) \in R_\delta$, then

$$|W_e(a,b) + W_f(a,b)| \ge a^{-s/2} \left(\epsilon_1^{1/2-p} + \sqrt{\epsilon} \right). \tag{3.6}$$

If $(a,b) \notin \bigcup_{1\leq k\leq K} Z_k$, then by Lemma 2.1.10,

$$|W_f(a,b)| \leq a^{-s/2}\epsilon. \tag{3.7}$$

Equation (3.6) and (3.7) lead to $|W_e(a,b)| \geq a^{-s/2}\sqrt{\epsilon_1}$. Hence,

$$P\left((a,b) \notin \bigcup_{1\leq k\leq K} Z_k\right) \leq P\left(|W_e(a,b)| \geq a^{-s/2}\sqrt{\epsilon_1}\right) = 1 - P\left(G_1\right).$$

This means that if $(a,b) \in R_\delta$, then $(a,b) \in \bigcup_{1\leq k\leq K} Z_k$ with a probability at least $P\left(G_1\right) = 1 - e^{-a^{-s}\epsilon_1^{-q}\|w\|^{-2}} = 1 - e^{-O\left(N_k^{-s}\epsilon_1^{-q}\right)}$, since $a \simeq N_k$ if $(a,b) \in Z_k$. So, (ii) is true. A similar argument applied to $(a,b) \in S_\delta$ shows that $(a,b) \in \bigcup_{1\leq k\leq K} Z_k$ with a probability at least $P\left(G_2\right) = 1 - e^{-\epsilon_1^{-q}\|w\|^{-2}}$. Hence, (iii) is proved.

Note that any rotated polydisk of radius r in $(z_1,z_2) \in \mathbb{C}^2$ contains a smaller polydisk of radius $2^{-1/2}r$ that is aligned with the z_1 and z_2 planes. If we define a transform $z' = Uz$ and introduce notations $\delta_1 = a^{-s/2}\sqrt{\epsilon_1}$, $\delta_2 = \sqrt{\epsilon_1}$, $\delta_3 = \left(a^{1-s/2} + a^{s/2}\right)\sqrt{\epsilon_1}$, $d_1 = \min\{\frac{\delta_1}{\sqrt{2}}, \frac{\delta_3}{\sqrt{2}}\}$, and $d_2 = \min\{\frac{\delta_2}{\sqrt{2}}, \frac{\delta_3}{\sqrt{2}}\}$, then

$$\begin{aligned}
P\left(G_1 \cap G_3\right) &= \int_{\{|z_1|<\delta_1, |z_2|<\delta_3\}} \frac{e^{-\epsilon_1^{-(1+q)} z^* V^{-1} z}}{\pi^2 \epsilon_1^{2(1+q)} \det V} dz_1 dz_2 \\
&= \int_{\{|z_1|<\delta_1, |z_2|<\delta_3\}} \frac{e^{-\epsilon_1^{-(1+q)}\left(D_{11}|z_1'|^2 + D_{22}|z_2'|^2\right)}}{\pi^2 \epsilon_1^{2(1+q)} \det V} dz_1' dz_2' \\
&\geq \int_{\{|z_1'|^2+|z_2'|^2<2d_1^2\}} \frac{e^{-\epsilon_1^{-(1+q)}\left(D_{11}|z_1'|^2 + D_{22}|z_2'|^2\right)}}{\pi^2 \epsilon_1^{2(1+q)} \det V} dz_1' dz_2' \\
&\geq \int_{\{|z_1'|<d_1, |z_2'|<d_1\}} \frac{e^{-\epsilon_1^{-(1+q)}\left(D_{11}|z_1'|^2 + D_{22}|z_2'|^2\right)}}{\pi^2 \epsilon_1^{2(1+q)} \det V} dz_1' dz_2' \\
&= \frac{4}{\epsilon_1^{2(1+q)} \det V} \int_0^{d_1} r_1 e^{-\frac{D_{11} r_1^2}{\epsilon_1^{1+q}}} dr_1 \int_0^{d_1} r_2 e^{-\frac{D_{22} r_2^2}{\epsilon_1^{1+q}}} dr_2 \\
&= \left(1 - e^{-\frac{D_{11} d_1^2}{\epsilon_1^{1+q}}}\right)\left(1 - e^{-\frac{D_{22} d_1^2}{\epsilon_1^{1+q}}}\right),
\end{aligned}$$

and similarly

$$P\left(G_2 \cap G_3\right) = \int_{\{|z_1|<\delta_2, |z_2|<\delta_3\}} \frac{e^{-\epsilon_1^{-(1+q)} z^* V^{-1} z}}{\pi^2 \epsilon_1^{2(1+q)} \det V} dz_1 dz_2 \geq \left(1 - e^{-\frac{D_{11} d_2^2}{\epsilon_1^{1+q}}}\right)\left(1 - e^{-\frac{D_{22} d_2^2}{\epsilon_1^{1+q}}}\right).$$

Suppose that Ξ is a real random variable with a probability density function $h(\xi) = \frac{|\widehat{w}(\xi)|^2}{\|\widehat{w}\|^2}$, then

$$\begin{aligned} D_{11}^{-1}D_{22}^{-1} &= \det(V) \\ &= 4\pi^2\|\widehat{w}\|^4\left(\int_{\mathbb{R}}(a^s\xi+a)^2\frac{|\widehat{w}(\xi)|^2}{\|\widehat{w}\|^2}d\xi - \left(\int_{\mathbb{R}}(a^s\xi+a)\frac{|\widehat{w}(\xi)|^2}{\|\widehat{w}\|^2}d\xi\right)^2\right) \\ &= 4\pi^2\|\widehat{w}\|^4\mathrm{Var}\left[a^s\Xi+a\right] \\ &\simeq a^{2s}, \end{aligned}$$

and

$$\begin{aligned} D_{11}+D_{22} &= \det\left(V^{-1}\right)\left(\|\widehat{w}\|^2+\langle 2\pi i\xi\widehat{w_{ab}}, 2\pi i\xi\widehat{w_{ab}}\rangle\right) \\ &\simeq \frac{1+4\pi^2\mathbb{E}\left[(a^s\Xi+a)^2\right]}{a^{2s}} \\ &\simeq \mathbb{E}\left[\left(\Xi+a^{1-s}\right)^2\right] \\ &\simeq a^{2(1-s)}. \end{aligned}$$

This implies $D_{11} \simeq a^{2(1-s)}$ and $D_{22} \simeq a^{-2}$. Therefore,

$$P(G_1\cap G_3) \geq \left(1-e^{-\frac{D_{11}d_1^2}{\epsilon_1^{1+q}}}\right)\left(1-e^{-\frac{D_{22}d_1^2}{\epsilon_1^{1+q}}}\right) = \left(1-e^{-O\left(a^{2-3s}\epsilon_1^{-q}\right)}\right)\left(1-e^{-O\left(a^{-s-2}\epsilon_1^{-q}\right)}\right),$$

and

$$P(G_2\cap G_3) \geq \left(1-e^{-\frac{D_{11}d_2^2}{\epsilon_1^{1+q}}}\right)\left(1-e^{-\frac{D_{22}d_2^2}{\epsilon_1^{1+q}}}\right) = \left(1-e^{-O\left(a^{2-2s}\epsilon_1^{-q}\right)}\right)\left(1-e^{-O\left(a^{-2}\epsilon_1^{-q}\right)}\right).$$

By Theorem 3.2.1, if $(a,b) \in R_\delta \cap Z_k$ for some k, then

$$P(H_k) \geq P(H_k|G_1\cap G_3)P(G_1\cap G_3) = P(G_1\cap G_3) \geq \left(1-e^{-O\left(a^{2-3s}\epsilon_1^{-q}\right)}\right)\left(1-e^{-O\left(a^{-s-2}\epsilon_1^{-q}\right)}\right).$$

Note that $a \simeq N_k$ when $a \in Z_k$, then

$$P(H_k) \geq \left(1-e^{-O\left(N_k^{2-3s}\epsilon_1^{-q}\right)}\right)\left(1-e^{-O\left(N_k^{-s-2}\epsilon_1^{-q}\right)}\right).$$

Similarly, if $(a,b) \in S_\delta \cap Z_k$ for some k, then

$$P(J_k) \geq P(J_k|G_2\cap G_3)P(G_2\cap G_3) = P(G_2\cap G_3) \geq \left(1-e^{-O\left(N_k^{2-2s}\epsilon_1^{-q}\right)}\right)\left(1-e^{-O\left(N_k^{-2}\epsilon_1^{-q}\right)}\right).$$

These arguments prove (iv) and (v).

Step 2: we go on to prove this theorem when the mother wave packet is of type (ϵ, m) with $m \geq \frac{2}{1-s}+4$. We would like to emphasize that the requirement is crucial to the following asymptotic analysis and it keeps the error caused by the non-compact support of $\widehat{w}$ reasonably small.

The sketch of the proof is similar to the first step, but $W_e(a,b)$ and $(W_e(a,b), \partial_b W_e(a,b))$ are Gaussian variables not circularly symmetric. Suppose they have covariance matrices C_1 and C_2, pseudo-covariance matrices P_1 and P_2, respectively. We can still check that they have zero mean, $C_1 = \epsilon_1^{1+q}\|w\|^2$ and $C_2 = \epsilon_1^{1+q}V$, where V is defined in the first step. By the definition of the mother wave packet of type (ϵ, m), we have

$$\begin{aligned}
& |\mathbb{E}\left[W_e(a,b)W_e(a,b)\right]| \\
\leq\ & \epsilon_1^{1+q}\int_{\mathbb{R}} |\widehat{w_{ab}}(\xi)\widehat{w_{ab}}(-\xi)|\, d\xi \\
\leq\ & \epsilon_1^{1+q}\int_{\mathbb{R}} \left|\widehat{w}\left(\xi - a^{1-s}\right)\widehat{w}\left(-\xi - a^{1-s}\right)\right| d\xi \\
\leq\ & \epsilon_1^{1+q}\left(\int_{\xi>0} \left|\widehat{w}\left(\xi - a^{1-s}\right)\widehat{w}\left(-\xi - a^{1-s}\right)\right| d\xi + \int_{\xi<0} \left|\widehat{w}\left(\xi - a^{1-s}\right)\widehat{w}\left(-\xi - a^{1-s}\right)\right| d\xi\right) \\
\leq\ & \frac{\epsilon_1^{1+q}\epsilon}{(a^{1-s}-1)^m}\left(\int_{\xi>0} \left|\widehat{w}\left(\xi - a^{1-s}\right)\right| d\xi + \int_{\xi<0} \left|\widehat{w}\left(-\xi - a^{1-s}\right)\right| d\xi\right) \\
\simeq\ & \frac{2\epsilon_1^{1+q}\epsilon}{a^{m(1-s)}}\int_{\mathbb{R}} |\widehat{w}(\xi)|\, d\xi.
\end{aligned}$$

Similarly, we know

$$\left|\mathbb{E}\left[\left(\partial_b W_e(a,b)\right)^2\right]\right| \lesssim \frac{8\pi^2\epsilon_1^{1+q}\epsilon}{a^{m(1-s)}}\int_{\mathbb{R}} \left|\xi^2\widehat{w}(\xi)\right| d\xi,$$

$$|\mathbb{E}\left[\partial_b W_e(a,b)W_e(a,b)\right]| \lesssim \frac{4\pi\epsilon_1^{1+q}\epsilon}{a^{m(1-s)}}\int_{\mathbb{R}} |\xi\widehat{w}(\xi)|\, d\xi.,$$

$$\mathbb{E}\left[\partial_b W_e(a,b)\overline{\partial_b W_e(a,b)}\right] = \epsilon_1^{1+q}\langle \partial_b w_{ab}, \partial_b w_{ab}\rangle = \epsilon_1^{1+q}\langle 2\pi i\xi\widehat{w_{ab}}, 2\pi i\xi\widehat{w_{ab}}\rangle$$

and

$$\mathbb{E}\left[W_e(a,b)\overline{\partial_b W_e(a,b)}\right] = \epsilon_1^{1+p}\langle w_{ab}, \partial_b w_{ab}\rangle = \epsilon_1^{1+q}\langle \widehat{w_{ab}}, 2\pi i\xi\widehat{w_{ab}}\rangle.$$

Hence, the magnitude of every entry in P_1 and P_2 is bounded by $O\left(\frac{\epsilon_1^{1+q}\epsilon}{a^{m(1-s)}}\right)$. Since the covariance matrix of $(W_e(a,b), W_e^*(a,b))$ is

$$V_1 = \begin{pmatrix} C_1 & P_1 \\ P_1^* & C_1^* \end{pmatrix},$$

according to Equation (27) in [134], the distribution of $W_e(a,b)$ is described by the following distribution

$$\frac{e^{-\frac{1}{2}(z_1^*, z_1)V_1^{-1}(z_1, z_1^*)^T}}{\pi\sqrt{\det V_1}},$$

which is

$$\frac{e^{-\frac{C_1|z_1|^2-\mathfrak{Re}\left(P_1^* z_1^2\right)}{C_1^2-P_1P_1^*}}}{\pi\sqrt{C_1^2-P_1P_1^*}}.$$

Notice that

$$\frac{C_1}{\sqrt{C_1^2-P_1P_1^*}} = 1+O\left(\frac{P_1P_1^*}{C_1^2}\right) = 1+O\left(\frac{\epsilon^2}{a^{2m(1-s)}}\right),$$

$$\frac{C_1|z_1|^2-\mathfrak{Re}\left(P_1^* z_1^2\right)}{C_1|z_1|^2} = 1+O\left(\frac{\epsilon}{a^{m(1-s)}}\right),$$

and

$$\frac{C_1^2}{C_1^2-P_1^*} = 1+O\left(\frac{\epsilon^2}{a^{2m(1-s)}}\right).$$

Hence,

$$\frac{e^{-\frac{1}{2}(z_1^*,z_1)V_1^{-1}(z_1,z_1^*)^T}}{\pi\sqrt{\det V_1}} = \frac{e^{-\epsilon_1^{-(1+q)}|z_1|^2\|w\|^{-2}}}{\pi\epsilon_1^{1+q}\|w\|^2}\left(1+O\left(\frac{\epsilon|z_1|^2}{\epsilon_1^{1+q}a^{m(1-s)}}\right)\right).$$

By the same argument, the covariance matrix of $(W_e(a,b),\partial_b W_e(a,b),W_e^*(a,b),\partial_b W_e^*(a,b))$ is

$$V_2 = \begin{pmatrix} C_2 & P_2 \\ P_2^* & C_2^* \end{pmatrix}.$$

Let $z=(z_1,z_2)^T$, where T and $*$ denote the transpose operator and conjugate transpose operator, respectively. Then the distribution of $(W_e(a,b),\partial_b W_e(a,b))$ is described by the joint probability density

$$\frac{e^{-\frac{1}{2}(z_1^*,z_2^*,z_1,z_2)V_2^{-1}(z_1,z_2,z_1^*,z_2^*)^T}}{\pi^2\sqrt{\det V_2}}. \tag{3.8}$$

Notice that $C_2=\epsilon_1^{1+p}V$ and V has eigenvalues of order a^2 and $a^{2(s-1)}$. Hence, C_2 has eigenvalues of order $\epsilon_1^{1+p}a^2$ and $\epsilon_1^{1+p}a^{2(s-1)}$. Recall that the magnitude of every entry in P_2 is bounded by $O\left(\frac{\epsilon_1^{1+q}\epsilon}{a^{m(1-s)}}\right)$. This means that V_2 is nearly dominated by diagonal blocks C_2 and C_2^*. Basic spectral theory for linear transforms shows that

$$V_2^{-1} = \begin{pmatrix} C_2^{-1} & \\ & (C_2^*)^{-1} \end{pmatrix} + P_\epsilon,$$

where P_ϵ is a matrix with 2-norm bounded by

$$O\left(\frac{\epsilon_1^{1+q}\epsilon}{a^{m(1-s)}}\right)O(\epsilon_1^{1+p}a^{2(s-1)})^{-2} = O\left(\epsilon_1^{-(1+q)}\epsilon a^{(m-4)(s-1)}\right).$$

$\frac{m-6}{m-4} \geq s$ ensures the above spectral analysis. Since every entry of P_2 is bounded by $O\left(\frac{\epsilon_1^{1+q}\epsilon}{a^{m(1-s)}}\right)$,

$$\det V_2 = (\det C_2)^2 + O\left(\frac{\epsilon_1^{4(1+q)}\epsilon}{a^{m-2-(m+2)s}}\right),$$

where the residual comes from the entry bound and the eigenvalues of C_2. Hence (3.8) is actually

$$\frac{e^{-\epsilon_1^{-(1+q)}z^*V^{-1}z}e^{-\frac{1}{2}(z_1^*,z_2^*,z_1,z_2)P_\epsilon(z_1,z_2,z_1^*,z_2^*)^T}}{\pi^2\epsilon_1^{2(1+q)}\sqrt{(\det V)^2 + O\left(\frac{\epsilon}{a^{m-2-(m+2)s)}}\right)}}.$$

By the same argument in the first step, we can show that there exist a diagonal matrix $D = \mathrm{diag}\{a^{2(1-s)}, a^{-2}\}$ and a unitary matrix U such that $V^{-1} = U^*DU$. Part (i) is still true by previous theorems. To conclude Part (ii) to (v), we still need to estimate the probability of those events defined in the first step, i.e., $P(G_1)$, $P(G_2)$, $P(G_1 \cap G_3)$, $P(G_2 \cap G_3)$, $P(H_k)$ and $P(J_k)$. By the calculations above, we have

$$\begin{aligned}
P(G_1) &= \int_{|z_1|<a^{-s/2}\sqrt{\epsilon_1}} \frac{e^{-\frac{1}{2}(z_1^*,z_1)V_1^{-1}(z_1,z_1^*)^T}}{\pi\sqrt{\det V_1}}dz_1 \\
&= \int_{|z_1|<a^{-s/2}\sqrt{\epsilon_1}} \frac{e^{-\epsilon_1^{-(1+q)}|z_1|^2\|w\|^{-2}}}{\pi\epsilon_1^{1+q}\|w\|^2}\left(1 + O\left(\frac{\epsilon|z_1|^2}{\epsilon_1^{1+q}a^{m(1-s)}}\right)\right)dz_1 \\
&= \frac{2}{\epsilon_1^{1+q}\|w\|^2}\int_0^{a^{-s/2}\sqrt{\epsilon_1}}\left(r + O\left(\frac{\epsilon}{\epsilon_1^{1+q}a^{m(1-s)}}\right)r^3\right)e^{-\epsilon_1^{-(1+q)}r^2\|w\|^{-2}}dr \\
&= \int_0^{a^{-s/2}\epsilon_1^{-q/2}\|w\|^{-1}} 2re^{-r^2}dr + O\left(\frac{2\epsilon\|w\|^2}{a^{m(1-s)}}\right)\int_0^{a^{-s/2}\epsilon_1^{-q/2}\|w\|^{-1}} r^3e^{-r^2}dr \\
&= 1 - e^{-a^{-s}\epsilon_1^{-q}\|w\|^{-2}} + O\left(\frac{\epsilon}{a^{m(1-s)}}\right),
\end{aligned}$$

and similarly

$$P(G_2) = 1 - e^{-\epsilon_1^{-q}\|w\|^{-2}} + O\left(\frac{\epsilon}{a^{m(1-s)}}\right).$$

Hence, we can conclude (ii) and (iii) follows the same proof in the first step. Next, we look at the last two part of this theorem.

Recall that we have defined a transform $z' = Uz$ and introduced notations $\delta_1 = a^{-s/2}\sqrt{\epsilon_1}$, $\delta_2 = \sqrt{\epsilon_1}$, $\delta_3 = \left(a^{1-s/2} + a^{s/2}\right)\sqrt{\epsilon_1}$, $d_1 = \min\{\frac{\delta_1}{\sqrt{2}}, \frac{\delta_3}{\sqrt{2}}\}$, and $d_2 = \min\{\frac{\delta_2}{\sqrt{2}}, \frac{\delta_3}{\sqrt{2}}\}$ in the first step.

Using the same notations and a similar argument, we have

$$\begin{aligned} & P\left(G_1 \cap G_3\right) \\ = & \int_{\{|z_1|<\delta_1,|z_2|<\delta_3\}} \frac{e^{-\frac{1}{2}(z_1^*,z_2^*,z_1,z_2)V_2^{-1}(z_1,z_2,z_1^*,z_2^*)^T}}{\pi^2\sqrt{\det V_2}} dz_1 dz_2 \\ = & \int_{\{|z_1|<\delta_1,|z_2|<\delta_3\}} \frac{e^{-\epsilon_1^{-(1+q)}z^*V^{-1}z} e^{-\frac{1}{2}(z_1^*,z_2^*,z_1,z_2)P_\epsilon(z_1,z_2,z_1^*,z_2^*)^T}}{\pi^2\epsilon_1^{2(1+q)}\sqrt{(\det V)^2 + O\left(\frac{\epsilon}{a^{m-2-(m+2)s)}}\right)}} dz_1 dz_2 \end{aligned} \tag{3.9}$$

Since

$$\frac{\det V}{\sqrt{(\det V)^2 + O\left(\frac{\epsilon}{a^{m-2-(m+2)s)}}\right)}} = 1 + O\left(\frac{\epsilon}{a^{(m-2)(1-s))}}\right), \tag{3.10}$$

we can drop out the term $O\left(\frac{\epsilon}{a^{m-2-(m+2)s)}}\right)$ in (3.9), which would generate an absolute error no more than $O\left(\frac{\epsilon}{a^{(m-2)(1-s))}}\right)$ in the estimate of $P(G_1 \cup G_3)$. Let

$$g(z) = -\frac{1}{2}(z_1^*, z_2^*, z_1, z_2)P_\epsilon(z_1, z_2, z_1^*, z_2^*)^T,$$

then

$$\begin{aligned} & P\left(G_1 \cap G_3\right) \\ \approx & \int_{\{|z_1|<\delta_1,|z_2|<\delta_3\}} \frac{e^{-\epsilon_1^{-(1+q)}z^*V^{-1}z} e^{g(z)}}{\pi^2\epsilon_1^{2(1+q)}\det V} dz_1 dz_2 \\ = & \int_{\{|z_1|<\delta_1,|z_2|<\delta_3\}} \frac{e^{-\epsilon_1^{-(1+q)}\left(D_{11}|z_1'|^2+D_{22}|z_2'|^2\right)} e^{g(U^*z')}}{\pi^2\epsilon_1^{2(1+q)}\det V} dz_1' dz_2' \\ \geq & \int_{\{|z_1'|^2+|z_2'|^2<2d_1^2\}} \frac{e^{-\epsilon_1^{-(1+q)}\left(D_{11}|z_1'|^2+D_{22}|z_2'|^2\right)} e^{g(U^*z')}}{\pi^2\epsilon_1^{2(1+q)}\det V} dz_1' dz_2' \\ \geq & \int_{\{|z_1'|<d_1,|z_2'|<d_1\}} \frac{e^{-\epsilon_1^{-(1+q)}\left(D_{11}|z_1'|^2+D_{22}|z_2'|^2\right)} e^{g(U^*z')}}{\pi^2\epsilon_1^{2(1+q)}\det V} dz_1' dz_2' \\ = & \frac{1}{\pi^2\epsilon_1^{2(1+q)}\det V} \int_0^{d_1}\int_0^{d_1}\int_0^{2\pi}\int_0^{2\pi} r_1 r_2 e^{-\frac{D_{11}r_1^2}{\epsilon_1^{1+q}}} e^{-\frac{D_{22}r_2^2}{\epsilon_1^{1+q}}} e^{\tilde{g}(r_1,\theta_1,r_2,\theta_2)} d\theta_1 d\theta_2 dr_1 dr_2 \\ = & \frac{1}{\pi^2\epsilon_1^{2(1+q)}\det V} \int_0^{d_1}\int_0^{d_1}\int_0^{2\pi}\int_0^{2\pi} r_1 r_2 e^{-\frac{D_{11}r_1^2}{\epsilon_1^{1+q}}} e^{-\frac{D_{22}r_2^2}{\epsilon_1^{1+q}}} \left(e^{\tilde{g}(r_1,\theta_1,r_2,\theta_2)} - 1\right) d\theta_1 d\theta_2 dr_1 dr_2 \\ & + \left(1 - e^{-\frac{D_{11}d_1^2}{\epsilon_1^{1+q}}}\right)\left(1 - e^{-\frac{D_{22}d_1^2}{\epsilon_1^{1+q}}}\right), \end{aligned} \tag{3.11}$$

where $\tilde{g}(r_1, \theta_1, r_2, \theta_2) = g(U^*z')$. Recall that the 2-norm of P_ϵ is bounded by $O\left(\epsilon_1^{-(1+q)}\epsilon a^{(m-4)(s-1)}\right)$.

Hence,

$$|\tilde{g}(r_1,\theta_1,r_2,\theta_2)| \le O\left(\epsilon_1^{-(1+q)}\epsilon a^{(m-4)(s-1)}\right)\left(|z_1|^2+|z_2|^2\right) = O\left(\epsilon_1^{-(1+q)}\epsilon a^{(m-4)(s-1)}\right)\left(r_1^2+r_2^2\right).$$

Therefore, the first term in (3.11) is bounded by

$$\begin{aligned}
&\frac{O\left(\epsilon a^{(m-4)(s-1)}\right)}{\epsilon_1^{3(1+q)}\det V}\int_0^{d_1}\int_0^{d_1} r_1 r_2 e^{-\frac{D_{11}r_1^2}{\epsilon_1^{1+q}}} e^{-\frac{D_{22}r_2^2}{\epsilon_1^{1+q}}}\left(r_1^2+r_2^2\right)dr_1dr_2\\
=\;& O\left(\frac{\epsilon}{a^{(m-4)(1-s)}}\right)\int_0^{\sqrt{\frac{D_{11}}{\epsilon_1^{1+q}}}d_1}\int_0^{\sqrt{\frac{D_{22}}{\epsilon_1^{1+q}}}d_1} r_1 r_2\left(\frac{r_1^2}{D_{11}}+\frac{r_2^2}{D_{22}}\right)e^{-r_1^2}e^{-r_2^2}dr_1dr_2\\
\le\;& O\left(\frac{\epsilon}{D_{22}a^{(m-4)(1-s)}}\right)\int_0^{\infty}\int_0^{\infty} r_1 r_2\left(r_1^2+r_2^2\right)e^{-r_1^2}e^{-r_2^2}dr_1dr_2\\
=\;& O\left(\frac{\epsilon}{a^{(m-4)(1-s)-2}}\right). \qquad (3.12)
\end{aligned}$$

The analysis in (3.10) and (3.12) implies that

$$P(G_1\cup G_3) \ge \left(1-e^{-\frac{D_{11}d_1^2}{\epsilon_1^{1+q}}}\right)\left(1-e^{-\frac{D_{22}d_1^2}{\epsilon_1^{1+q}}}\right)+O\left(\frac{\epsilon}{a^{(m-4)(1-s)-2}}\right).$$

and similarly

$$\begin{aligned}
P\left(G_2\cap G_3\right) &= \int_{\{|z_1|<\delta_2,|z_2|<\delta_3\}}\frac{e^{-\frac{1}{2}(z_1^*,z_2^*,z_1,z_2)V_2^{-1}(z_1,z_2,z_1^*,z_2^*)^T}}{\pi^2\sqrt{\det V_2}}dz_1dz_2\\
&\ge \left(1-e^{-\frac{D_{11}d_2^2}{\epsilon_1^{1+q}}}\right)\left(1-e^{-\frac{D_{22}d_2^2}{\epsilon_1^{1+q}}}\right)+O\left(\frac{\epsilon}{a^{(m-4)(1-s)-2}}\right).
\end{aligned}$$

The rest of the proof is exactly the same as the one in the first step and consequently we know this theorem is also true for a mother wave packets of type (ϵ,m) with m satisfying $m\ge\frac{2}{1-s}+4$. □

Thus far, we considered the robustness to small perturbation and Gaussian white noise. Next, we will show that Theorem 3.2.2 can be extended to a broader class of colored noise.

Theorem 3.2.3. *Suppose the mother wave packet is of type (ϵ,m), for any fixed $\epsilon\in(0,1)$ and any fixed integer $m\ge\frac{2}{1-s}+4$. Suppose $g(x)=f(x)+e$, where e is a zero mean stationary Gaussian process. Let $\widehat{e}(\xi)$ denote the spectrum of e, $\max_\xi|\widehat{e}(\xi)|\le\epsilon^{-1}$ and $M_a=\max_{|\xi|<1}\widehat{e}(a^s\xi+a)$. For any $p\in(0,\frac{1}{2}]$ and $q>0$, let $\delta_a=M_a^{(\frac{1}{2}-p)/(1+q)}+\sqrt{\epsilon}$,*

$$R_{\delta_a}=\{(a,b):|W_g(a,b)|\ge a^{-s/2}\delta_a\},$$

$$S_{\delta_a}=\{(a,b):|W_g(a,b)|\ge\delta_a\},$$

and

$$Z_k = \{(a,b) : |a - N_k\phi_k'(b)| \leq a^s\}$$

for $1 \leq k \leq K$. *For fixed* M *and* K, *there exists a constant* $N_0(M,m,K,s,\epsilon) \simeq \max\left\{\epsilon^{\frac{-1}{2s-1}}, \epsilon^{\frac{-1}{1-s}}\right\}$ *such that for any* $N > N_0(M,m,K,s,\epsilon)$ *and* $f(x) \in F(M,N,K,s)$ *the following statements hold.*

(i) $\{Z_k : 1 \leq k \leq K\}$ *are disjoint.*

(ii) If $(a,b) \in R_{\delta_a}$, *then* $(a,b) \in \bigcup_{1\leq k\leq K} Z_k$ *with a probability at least*

$$1 - e^{-O\left(N_k^{-s}M_a^{-q/(1+q)}\right)} + O\left(\frac{\epsilon}{N_k^{m(1-s)}}\right).$$

(iii) If $(a,b) \in S_{\delta_a}$, *then* $(a,b) \in \bigcup_{1\leq k\leq K} Z_k$ *with a probability at least*

$$1 - e^{-O\left(M_a^{-q/(1+q)}\right)} + O\left(\frac{\epsilon}{N_k^{m(1-s)}}\right).$$

(iv) If $(a,b) \in R_{\delta_a} \cap Z_k$ *for some* k, *then*

$$\frac{|v_g(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \lesssim \sqrt{\epsilon} + M_a^{p/(1+q)}$$

is true with a probability at least

$$\left(1 - e^{-O\left(N_k^{2-3s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-s-2}M_a^{-q/(1+q)}\right)}\right) + O\left(\frac{\epsilon}{N_k^{(m-4)(1-s)-2}}\right).$$

(v) If $(a,b) \in S_{\delta_a} \cap Z_k$, *then*

$$\frac{|v_g(a,b) - N_k\phi_k'(b)|}{|N_k\phi_k'(b)|} \lesssim N_k^{-s/2}\left(\sqrt{\epsilon} + M_a^{p/(1+q)}\right)$$

is true with a probability at least

$$\left(1 - e^{-O\left(N_k^{2-2s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-2}M_a^{-q/(1+q)}\right)}\right) + O\left(\frac{\epsilon}{N_k^{(m-4)(1-s)-2}}\right).$$

Proof. The proof of this theorem is nearly identical to Theorem 3.2.2 but for the covariance functional of the noise term which is now a general functional $\mathcal{R} : L^1 \cap C^{m-1} \to L^1 \cap C^{m-1}$.

Step 1: In a similar structure, we prove the case when the mother wave packet is of type $(0,m)$. We can still check that $W_g(a,b) = W_f(a,b) + W_e(a,b)$ and $\partial_b W_g(a,b) = \partial_b W_f(a,b) + \partial_b W_e(a,b)$

are Gaussian variables. Furthermore, $W_e(a,b)$ and $(W_e(a,b), \partial_b W_e(a,b))$ are still circularly symmetric Gaussian variables. Since the Gaussian process e is zero mean, we have $\mathbb{E}[W_e(a,b)] = 0$ and $\mathbb{E}[\partial_b W_e(a,b)] = 0$. Note that $\mathcal{R}$ can be "diagonalized" to a functional $\mathcal{D}$ by the Fourier transform denoted as $\mathcal{F}$ in the sense that

$$\langle f_1, \mathcal{R} f_2 \rangle = \langle f_1, \mathcal{F}^* \mathcal{D} \mathcal{F} f_2 \rangle = \langle \widehat{f_1}, \widehat{e}\widehat{f_2} \rangle$$

for any f_1 and f_2 in $L^1 \cap C^{m-1}$. Hence,

$$\mathbb{E}\left[W_e(a,b)\overline{W_e(a,b)}\right] = \langle w_{ab}, \mathcal{R} w_{ab} \rangle = \langle \widehat{w_{ab}}, \widehat{e w_{ab}} \rangle = \langle \widehat{w}, \widehat{e}(a^s \xi + a)\widehat{w} \rangle$$

and

$$\mathbb{E}[W_e(a,b) W_e(a,b)] = \langle w_{ab}, \overline{\mathcal{R} w_{ab}} \rangle = \langle \widehat{w_{ab}}, \overline{\widehat{e}(-\xi)\widehat{w_{ab}}(-\xi)} \rangle = \int_{\mathbb{R}} \widehat{w_{ab}}(\xi)\widehat{w_{ab}}(-\xi)\widehat{e}(-\xi) d\xi = 0.$$

If we introduce $\sigma^2 = \langle \widehat{w}, \widehat{e}(a^s \xi + a)\widehat{w} \rangle$ for simplicity and a random variable Ξ with a probability density function $\sigma^{-2}|\widehat{w}|^2 \widehat{e}(a^s \xi + a)$, then by a similar argument, we know

$$\mathbb{E}\left[(\partial_b W_e(a,b))^2\right] = \mathbb{E}[\partial_b W_e(a,b) W_e(a,b)] = 0,$$

$$\mathbb{E}\left[\partial_b W_e(a,b)\overline{\partial_b W_e(a,b)}\right] = \langle \partial_b w_{ab}, \mathcal{R} \partial_b w_{ab} \rangle = 4\pi\sigma^2 \mathbb{E}\left[(a^s \Xi + a)^2\right],$$

and

$$\mathbb{E}\left[W_e(a,b)\overline{\partial_b W_e(a,b)}\right] = \langle w_{ab}, \mathcal{R} \partial_b w_{ab} \rangle = 2\pi i \sigma^2 \mathbb{E}[a^s \Xi + a].$$

Hence, $W_e(a,b)$ and $(W_e(a,b), \partial_b W_e(a,b))$ have zero pseudo-covariance matrices and they are circularly symmetric. Therefore, the distribution of $W_e(a,b)$ is determined by its variance as follows

$$\frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^2}.$$

If we define

$$V = \begin{pmatrix} 1 & 2\pi i \mathbb{E}[a^s \Xi + a] \\ -2\pi i \mathbb{E}[a^s \Xi + a] & 4\pi^2 \mathbb{E}[(a^s \Xi + a)] \end{pmatrix},$$

then $\sigma^2 V$ is the covariance matrix of $(W_e(a,b), \partial_b W_e(a,b))$ and its distribution is described by the joint probability density

$$\frac{e^{-\sigma^{-2} z^* V^{-1} z}}{\pi^2 \sigma^4 \det V},$$

where $z = (z_1, z_2)^T$. V is an invertible and self-adjoint matrix, since $W_e(a,b)$ and $\partial_b W_e(a,b)$ are linearly independent. Hence, there exist a diagonal matrix D and a unitary matrix U such that

$V^{-1} = U^* D U$.

Part (i) is true by previous theorems. Define the following events

$$G_1 = \{|W_e(a,b)| < a^{-s/2} M_a^{1/(2+2q)}\},$$

$$G_2 = \{|W_e(a,b)| < M_a^{1/(2+2q)}\},$$

$$G_3 = \{|\partial_b W_e(a,b)| < M_a^{1/(2+2q)} \left(a^{1-s/2} + a^{s/2}\right)\},$$

$$H_k = \left\{ \frac{|v_g(a,b) - N_k \phi_k'(b)|}{|N_k \phi_k'(b)|} \lesssim \sqrt{\epsilon} + M_a^{p/(1+q)} \right\},$$

and

$$J_k = \left\{ \frac{|v_g(a,b) - N_k \phi_k'(b)|}{|N_k \phi_k'(b)|} \lesssim N_k^{-s/2} \left(\sqrt{\epsilon} + M_a^{p/(1+q)}\right) \right\},$$

for $1 \leq k \leq K$. Now we estimate the probability $P(G_1)$, $P(G_2)$, $P(G_1 \cap G_3)$, $P(G_2 \cap G_3)$, $P(H_k)$ and $P(J_k)$. By the calculations above, we have

$$P(G_1) = \int_{|z_1| < a^{-s/2} M_a^{1/(2+2q)}} \frac{e^{-\sigma^{-2}|z_1|^2}}{\pi \sigma^2} dz_1 = 1 - e^{-a^{-s} M_a^{1/(1+q)} \sigma^{-2}} \geq 1 - e^{-O\left(a^{-s} M_a^{-q/(1+q)}\right)},$$

and similarly

$$P(G_2) \geq 1 - e^{-O\left(M_a^{-q/(1+q)}\right)}.$$

We are ready to summarize and conclude (ii) and (iii). If $(a,b) \in R_{\delta_a}$, then

$$|W_e(a,b) + W_f(a,b)| \geq a^{-s/2} \left(M_a^{(\frac{1}{2}-p)/(1+q)} + \sqrt{\epsilon}\right). \tag{3.13}$$

If $(a,b) \notin \bigcup_{1 \leq k \leq K} Z_k$, then by Lemma 2.1.10,

$$|W_f(a,b)| \leq a^{-s/2} \epsilon. \tag{3.14}$$

Equation (3.13) and (3.14) lead to $|W_e(a,b)| \geq a^{-s/2} M_a^{(\frac{1}{2}-p)/(1+q)}$. Hence,

$$P\left((a,b) \notin \bigcup_{1 \leq k \leq K} Z_k\right) \leq P\left(|W_e(a,b)| \geq a^{-s/2} M_a^{(\frac{1}{2}-p)/(1+q)}\right) = 1 - P(G_1).$$

This means that if $(a,b) \in R_{\delta_a}$, then $(a,b) \in \bigcup_{1 \leq k \leq K} Z_k$ with a probability at least $P(G_1) \geq 1 - e^{-O\left(a^{-s} M_a^{-q/(1+q)}\right)} = 1 - e^{-O\left(N_k^{-s} M_a^{-q/(1+q)}\right)}$, since $a \simeq N_k$ if $(a,b) \in Z_k$. So, (ii) is true. A similar argument applied to $(a,b) \in S_{\delta_a}$ shows that $(a,b) \in \bigcup_{1 \leq k \leq K} Z_k$ with a probability at least $P(G_2) = 1 - e^{-O\left(M_a^{-q/(1+q)}\right)}$. Hence, (iii) is proved.

If we introduce notations $\delta_1 = a^{-s/2} M_a^{1/(2+2q)}$, $\delta_2 = M_a^{1/(2+2q)}$, $\delta_3 = \left(a^{1-s/2} + a^{s/2}\right) M_a^{1/(2+2q)}$, $d_1 = \min\{\frac{\delta_1}{\sqrt{2}}, \frac{\delta_3}{\sqrt{2}}\}$, and $d_2 = \min\{\frac{\delta_2}{\sqrt{2}}, \frac{\delta_3}{\sqrt{2}}\}$, then it follows from the same proof in Theorem 3.2.2 that

$$\begin{aligned} P\left(G_1 \cap G_3\right) &= \int_{\{|z_1|<\delta_1, |z_2|<\delta_3\}} \frac{e^{-\sigma^{-2} z^* V^{-1} z}}{\pi^2 \sigma^4 \det V} dz_1 dz_2 \\ &\geq \left(1 - e^{-\frac{D_{11} d_1^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22} d_1^2}{\sigma^2}}\right), \end{aligned}$$

and similarly

$$P\left(G_2 \cap G_3\right) = \int_{\{|z_1|<\delta_2, |z_2|<\delta_3\}} \frac{e^{-\sigma^{-2} z^* V^{-1} z}}{\pi^2 \sigma^4 \det V} dz_1 dz_2 \geq \left(1 - e^{-\frac{D_{11} d_2^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22} d_2^2}{\sigma^2}}\right).$$

Note that

$$D_{11}^{-1} D_{22}^{-1} = \det\left(V\right) = 4\pi^2 a^{2s} \mathrm{Var}\left[\Xi\right]$$

and

$$D_{11} + D_{22} = \frac{1 + 4\pi^2 \mathbb{E}\left[\left(a^s \Xi + a\right)^2\right]}{4\pi^2 a^{2s} \mathrm{Var}\left[\Xi\right]}$$

We assume $D_{22} \leq D_{11}$. Since $\left|\mathbb{E}\left[\Xi\right]\right| \lesssim 1$ and $\mathbb{E}\left[\Xi^2\right] \lesssim 1$, then

$$D_{22} = \frac{\det\left(V^{-1}\right)}{D_{11}} \simeq \frac{1}{\det V \left(D_{11} + D_{22}\right)} = \frac{1}{1 + 4\pi^2 \mathbb{E}\left[\left(a^s \Xi + a\right)^2\right]} \simeq a^{-2},$$

and

$$D_{11} \simeq \frac{1 + 4\pi^2 \mathbb{E}\left[\left(a^s \Xi + a\right)^2\right]}{4\pi^2 a^{2s} \mathrm{Var}\left[\Xi\right]} \gtrsim a^{2-2s}.$$

This implies

$$\begin{aligned} P\left(G_1 \cap G_3\right) &\geq \left(1 - e^{-\frac{D_{11} d_1^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22} d_1^2}{\sigma^2}}\right) \\ &\gtrsim \left(1 - e^{-O\left(a^{2-3s} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(a^{-s-2} M_a^{-q/(1+q)}\right)}\right), \end{aligned}$$

and

$$\begin{aligned} P\left(G_2 \cap G_3\right) &\geq \left(1 - e^{-\frac{D_{11} d_2^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22} d_2^2}{\sigma^2}}\right) \\ &\gtrsim \left(1 - e^{-O\left(a^{2-2s} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(a^{-2} M_a^{-q/(1+q)}\right)}\right). \end{aligned}$$

By Theorem 3.2.1, if $(a,b) \in R_{\delta_a} \cap Z_k$ for some k, then

$$P(H_k) \geq P(H_k|G_1 \cap G_3) P(G_1 \cap G_3) = P(G_1 \cap G_3).$$

Note that $a \simeq N_k$ when $a \in Z_k$, then

$$P(H_k) \geq \left(1 - e^{-O\left(N_k^{2-3s} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-s-2} M_a^{-q/(1+q)}\right)}\right).$$

Similarly, if $(a,b) \in S_{\delta_a} \cap Z_k$ for some k, then

$$P(J_k) \quad \geq \quad \left(1 - e^{-O\left(N_k^{2-2s} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-2} M_a^{-q/(1+q)}\right)}\right).$$

These arguments prove (iv) and (v).

Step 2: We discuss the case for a mother wave packet of type (ϵ, m) for $m \geq \frac{2}{1-s} + 4$.

Similar to what we have already seen in the second step of the proof of Theorem 3.2.2. $W_g(a,b) = W_f(a,b) + W_e(a,b)$ and $\partial_b W_g(a,b) = \partial_b W_f(a,b) + \partial_b W_e(a,b)$ are Gaussian. $(W_e(a,b), \partial_b W_e(a,b))$ and $W_e(a,b)$ are nearly circularly symmetric Gaussian variables. Using the same strategy in Step 2 in the proof of Theorem 3.2.2 and the notations in Step 1 in this theorem, we can still check that:

1. The distribution of $W_e(a,b)$ is well approximated by

$$\frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^2},$$

   where $\sigma^2 = \langle \widehat{w}, \widehat{e}(a^s\xi + a)\widehat{w}\rangle$.

2. The distribution of $(W_e(a,b), \partial_b W_e(a,b))$ is well approximated by

$$\frac{e^{-\sigma^{-2} z^* V^{-1} z}}{\pi^2 \sigma^4 \det V},$$

   where

$$V = \begin{pmatrix} 1 & 2\pi i \mathbb{E}\left[a^s \Xi + a\right] \\ -2\pi i \mathbb{E}\left[a^s \Xi + a\right] & 4\pi^2 \mathbb{E}\left[(a^s \Xi + a)\right] \end{pmatrix},$$

   and V has eigenvalues $D_{22}^{-1} \simeq a^2$ and $D_{11}^{-1} \lesssim a^{2(s-1)}$.

Suppose G_1, G_2, G_3 are the events defined in the first step, then the well approximation here means that:

1.

$$\begin{aligned} P(G_1) &= \int_{\{|z_1|<a^{-s/2}M_a^{1/(2+2q)}\}} \frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^2} dz_1 + O\left(\frac{\epsilon}{a^{m(1-s)}}\right) \\ &= 1 - e^{-a^{-s}M_a^{1/(1+q)} + O\left(\frac{\epsilon}{a^{m(1-s)}}\right)\sigma^{-2}} \\ &\geq 1 - e^{-O\left(a^{-s}M_a^{-q/(1+q)}\right)} + O\left(\frac{\epsilon}{a^{m(1-s)}}\right), \end{aligned}$$

similarly

$$P(G_2) \geq 1 - e^{-O\left(M_a^{-q/(1+q)}\right)} + O\left(\frac{\epsilon}{a^{m(1-s)}}\right).$$

2.

$$\begin{aligned} P(G_1 \cap G_3) &= \int_{\{|z_1|<\delta_1, |z_2|<\delta_3\}} \frac{e^{-\sigma^{-2}z^* V^{-1} z}}{\pi^2\sigma^4 \det V} dz_1 dz_2 + O\left(\frac{\epsilon}{a^{(m-4)(1-s)-2}}\right) \\ &\geq \left(1 - e^{-\frac{D_{11}d_1^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22}d_1^2}{\sigma^2}}\right) + O\left(\frac{\epsilon}{a^{(m-4)(1-s)-2}}\right) \\ &\gtrsim \left(1 - e^{-O\left(a^{2-3s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(a^{-s-2}M_a^{-q/(1+q)}\right)}\right) + O\left(\frac{\epsilon}{a^{(m-4)(1-s)-2}}\right), \end{aligned}$$

and similarly

$$\begin{aligned} P(G_2 \cap G_3) &= \int_{\{|z_1|<\delta_2, |z_2|<\delta_3\}} \frac{e^{-\sigma^{-2}z^* V^{-1} z}}{\pi^2\sigma^4 \det V} dz_1 dz_2 + O\left(\frac{\epsilon}{a^{(m-4)(1-s)-2}}\right) \\ &\geq \left(1 - e^{-\frac{D_{11}d_2^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22}d_2^2}{\sigma^2}}\right) + O\left(\frac{\epsilon}{a^{(m-4)(1-s)-2}}\right) \\ &\gtrsim \left(1 - e^{-O\left(a^{2-2s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(a^{-2}M_a^{-q/(1+q)}\right)}\right) + O\left(\frac{\epsilon}{a^{(m-4)(1-s)-2}}\right). \end{aligned}$$

Following the proof in the first step, it is straightforward to see this theorem is true for a mother wave packet of type (ϵ, m) with $m \geq \frac{2}{1-s} + 4$. □

Theorem 3.2.2 and 3.2.3 illustrate that when the sampling rate of a given signal is high enough such that the wave-like components are relatively smooth in terms of the noise, the SSWPT can estimate the instantaneous frequencies of these components accurately with a high probability. In particular, Theorem 3.2.3 says that if the noise spectrum is not overwhelming the wave packet coefficients of IMTs, the SSWPT can provide accurate estimates with a high probability. Part (ii) and (iii) in the last two theorems demonstrate that the influence of noise can be significantly reduced with a proper threshold after the wave packet transform and we could obtain useful information with a high probability. Part (iv) and (v) show that the synchrosqueezing process is able to concentrate the wave packet representation to the instantaneous frequencies with a reasonable probability after

properly thresholding. Hence, the essential support of the synchrosqueezed energy distribution helps to estimate the instantaneous frequencies statistically.

In the above discussion, we have not optimized the dependence of N on ϵ. There are two extra steps to minimize the lower bound for N. Comparing Definition 3.1.1 and Definition 2.1.6, it is clear that we have allowed fully nonlinearity to IMTs in the previous theorems. The requirement $\epsilon^{\frac{-1}{2s-1}}$ can be reduced to a constant order if we restrict to a slightly smaller class of IMTs with weaker nonlinearity. For example, if $N_k \lesssim \epsilon^{-1/s}$ or $N_k \lesssim \epsilon^{-1/(2s-1)}$, then we impose extra condition $|\alpha_k'(x)| \leq \epsilon N_k^s |\phi_k'(x)|$ or $|\phi_k''(x)| \leq \epsilon N_k^{2s-1} |\phi_k'(x)|$, respectively. A careful inspection of the proof of Lemma 2.1.9 and 2.1.10 in the Taylor expansion approximation shows that these lemmas are still true. Hence, the synchrosqueezed transforms remain accurate.

Another step is to look at $\epsilon^{\frac{-1}{1-s}}$, which comes from the decaying estimate of wave packet coefficients $W_f(a,b)$ when their scales a do not match the oscillation N_k of IMTs. If we further take advantage of the decay speed of the mother wave packet, we will see $|W_f(a,b)|$ would decay much faster when this mismatch occurs. For example, a mother wave packet in C^m satisfies that

$$\widehat{w}(\xi) \leq C_m \left(1 + |\xi|\right)^{-m}.$$

Since the mother wave packet is decaying rapidly at infinity, we can simply assume that the smooth amplitude function of the IMT has a compact support large enough and only need to bound $|W_f(a,b)|$ for b at the support center. Since $\phi \in C^\infty$, by the diffeomorphism equivalence in Lemma 2.2 in [49] by Demanet and Ying, which is also valid for the wave packet transform by careful inspection, it is sufficient to assume $\phi(x) = x$. It follows from the discussion in Lemma 2.3 in [49] that we only require the following bound for previous theorems:

$$|W_f(a,b)| \leq a^{-s/2} C_m \left(1 + |a - N|\right)^{-m/2} \leq \epsilon.$$

Thus, we see $|W_f(a,b)|$ decays rapidly when $a \notin \left[\frac{N}{2M}, 2MN\right]$ for a reasonable large N. In practice, ϵ cannot be too small for numerical purposes and the number of periods of the input data is large enough so that a nonlinear wave-like component is well defined. Hence, the above requirement is not a main issue.

We close this section with a few extra remarks. First, $s \in (1/2, 1)$ is essential in those theorems if we do not impose extra condition on the nonlinearity of IMTs. Second, as pointed out in [178], another advantage of allowing $s \in (1/2, 1)$ is that a smaller s leads to a better scale resolution to distinguish two IMTs with close instantaneous frequencies or a sequence of IMTs with instantaneous frequencies spreading out in the time-frequency domain. We refer to [178] for a detailed discussion. Finally, the theorems above provide a new insight that a smaller s yields a synchrosqueezed transform with better robustness. This new insight is especially important when designing synchrosqueezed transforms compactly supported or decaying fast in the time domain. Theorem 3.2.2 and 3.2.3 show

that the parameter m in the mother wave packet has to be large enough, satisfying $m \geq \frac{2}{1-s} + 4$. In a special case, if a compactly supported synchrosqueezed wavelet transform (corresponding to $s = 1$) is preferable in some application, its mother wavelet is better to be C^∞.

3.3 2D Synchrosqueezed Wave Packet Transform (SSWPT)

We will focus on the robustness of the 2D SSWPT illustrated in the next two theorems. Appendix A.1 is referred to for the proofs of these theorems.

Theorem 3.3.1. *Suppose the 2D mother wave packet is of type (ϵ, m), for any fixed $\epsilon \in (0,1)$ and any fixed integer $m \geq 0$. Suppose $g(x) = f(x) + e(x)$, where $e(x) \in L^\infty$ is a small error term that satisfies $\|e\|_{L^\infty} \leq \epsilon_1$ for some $\epsilon_1 > 0$. For $p \in \left(0, \frac{1}{2}\right]$, let $\delta = \sqrt{\epsilon} + \epsilon_1^{\frac{1}{2}-p}$. Define*

$$R_\delta = \{(a,b) : |W_f(a,b)| \geq |a|^{-s}\delta\},$$

$$S_\delta = \{(a,b) : |W_f(a,b)| \geq \delta\},$$

and

$$Z_k = \{(a,b) : |a - N_k \nabla\phi_k(b)| \leq |a|^s\}$$

for $1 \leq k \leq K$. For fixed M, m, s, ϵ and K, there exists a constant $N_0(M,m,K,s,\epsilon) \simeq \max\left\{\epsilon^{\frac{-2}{2s-1}}, \epsilon^{\frac{-1}{1-s}}\right\}$ such that for any $N > N_0$ and $f(x) \in F(M,N,K,s)$ the following statements hold.

(i) $\{Z_k : 1 \leq k \leq K\}$ are disjoint and $S_\delta \subset R_\delta \subset \bigcup_{1\leq k\leq K} Z_k$;

(ii) For any $(a,b) \in R_\delta \cap Z_k$,

$$\frac{|v_g(a,b) - N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|} \lesssim \sqrt{\epsilon} + \epsilon_1^p;$$

(iii) For any $(a,b) \in S_\delta \cap Z_k$,

$$\frac{|v_g(a,b) - N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|} \lesssim N_k^{-s}\left(\sqrt{\epsilon} + \epsilon_1^p\right).$$

Theorem 3.3.1 shows that the 2D SSWPT is robust to a bounded perturbation. Actually, if the threshold δ is larger, e.g., $\delta \geq \sqrt{\frac{\epsilon_1}{\epsilon}}$, the relative estimate errors in (ii) and (iii) are bounded by $\sqrt{\epsilon}$ and $\frac{\sqrt{\epsilon}}{N_k^s}$, respectively. Hence, the local wave vector estimates are better if the wave packet coefficient with the largest magnitude is selected. However, when the perturbation is overwhelming, e.g., the wave packet coefficients of a 2D IMT are below the threshold in (ii), it is difficult to estimate its local wave vector. Next, Theorem 3.3.2 will illustrate the robustness properties of the 2D SSWPT to a zero mean stationary Gaussian noise.

Theorem 3.3.2. *Suppose the 2D mother wave packet is of type (ϵ, m), for any fixed $\epsilon \in (0,1)$ and any fixed integer $m \geq \max\left\{\frac{2(1+s)}{1-s}, \frac{2}{1-s}+4\right\}$. Suppose $g(x) = f(x) + e$, where e is a zero mean stationary Gaussian process with a spectrum denoted by $\widehat{e}(\xi)$ and $\max_\xi |\widehat{e}(\xi)| \leq \epsilon^{-1}$. Define $M_a = \max_{|\xi|<1} \widehat{e}(|a|^s\xi + a)$. For any $p \in (0, \frac{1}{2}]$ and $q > 0$, let $\delta_a = M_a^{(\frac{1}{2}-p)/(1+q)} + \sqrt{\epsilon}$,*

$$R_{\delta_a} = \{(a,b) : |W_g(a,b)| \geq |a|^{-s}\delta_a\},$$

$$S_{\delta_a} = \{(a,b) : |W_g(a,b)| \geq \delta_a\},$$

and

$$Z_k = \{(a,b) : |a - N_k \nabla_b \phi_k(b)| \leq |a|^s\}$$

for $1 \leq k \leq K$. For fixed M, m, s, ϵ and K, there exists a constant $N_0(M,m,K,s,\epsilon) \simeq \max\left\{\epsilon^{\frac{-2}{2s-1}}, \epsilon^{\frac{-1}{1-s}}\right\}$ such that for any $N > N_0$ and $f(x) \in F(M,N,K,s)$ the following statements hold.

(i) $\{Z_k : 1 \leq k \leq K\}$ are disjoint.

(ii) If $(a,b) \in R_{\delta_a}$, then $(a,b) \in \bigcup_{1\leq k\leq K} Z_k$ with a probability at least

$$1 - e^{-O\left(N_k^{-2s} M_a^{-q/(1+q)}\right)} + O\left(\frac{\epsilon}{N_k^{m(1-s)}}\right).$$

(iii) If $(a,b) \in S_{\delta_a}$, then $(a,b) \in \bigcup_{1\leq k\leq K} Z_k$ with a probability at least

$$1 - e^{-O\left(M_a^{-q/(1+q)}\right)} + O\left(\frac{\epsilon}{N_k^{m(1-s)}}\right).$$

(iv) If $(a,b) \in R_{\delta_a} \cap Z_k$ for some k, then

$$\frac{|v_g(a,b) - N_k \nabla \phi_k(b)|}{|N_k \nabla \phi_k(b)|} \lesssim \sqrt{\epsilon} + M_a^{p/(1+q)}$$

is true with a probability at least

$$\left(1 - e^{-O\left(N_k^{2-4s} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-4s} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-2-2s} M_a^{-q/(1+q)}\right)}\right)$$
$$+O\left(\frac{\epsilon}{N_k^{(m-4)(1-s)-2}}\right) + O\left(\frac{\epsilon}{N_k^{m-2-(m+2)s}}\right).$$

(v) If $(a,b) \in S_{\delta_a} \cap Z_k$ for some k, then

$$\frac{|v_g(a,b) - N_k \nabla \phi_k(b)|}{|N_k \nabla \phi_k(b)|} \lesssim N_k^{-s}\left(\sqrt{\epsilon} + M_a^{p/(1+q)}\right)$$

is true with a probability at least

$$\left(1-e^{-O\left(N_k^{2-2s}M_a^{-q/(1+q)}\right)}\right)\left(1-e^{-O\left(N_k^{-2s}M_a^{-q/(1+q)}\right)}\right)\left(1-e^{-O\left(N_k^{-2}M_a^{-q/(1+q)}\right)}\right)$$
$$+O\left(\frac{\epsilon}{N_k^{(m-4)(1-s)-2}}\right)+O\left(\frac{\epsilon}{N_k^{m-2-(m+2)s}}\right).$$

We would like to remark that the requirement of $N_0\left(M,K,s,\epsilon\right)\simeq\max\left\{\epsilon^{\frac{-2}{2s-1}},\epsilon^{\frac{-1}{1-s}}\right\}$ can be relieved if we impose a weak assumption on the nonlinearity of IMTs, as discussed in the end of Section 3.2. For example,

$$|\nabla\alpha_k(x)|\leq\epsilon N_k^s|\nabla\phi_k(x)| \text{ and } |\nabla^2\phi_k(x)|\leq\epsilon N_k^{2s-1}|\nabla\phi_k(x)|.$$

Hence, the theorems introduced in this section are multiscale indeed.

Theorem 3.3.1 and 3.3.2 show that the local wave vector estimates by the 2D SSWPT are robust against bounded perturbation and additive Gaussian random noise, if a threshold is properly chosen after the wave packet transform. First, the robustness becomes stronger as s gets smaller. Second, similar to the 1D case, as we increase the sampling rate of the signal to make IMTs relatively smoother compared to Gaussian random noise, the SSWPT can estimate local wave vectors accurately with a high probability.

3.4 2D Synchrosqueezed Curvelet Transform (SSCT)

In some applications such as the wave field separation problem [143, 163] and the ground roll removal problem [17, 72, 185] in geophysics, the IMTs to be analyzed and decomposed would have bounded supports in space, sometimes even banded supports. This motivates the design of the SSCT as a better tool to estimate local wave vectors of banded IMTs with close propagating directions in [184]. As we shall see in the following theorems, the geometric scaling of the SSCT is crucial to obtaining an accurate estimate of local wave vectors.

To model a wave-like component with a band-shape support, we are going to analyze components of the form

$$f(x)=e^{-(\phi(x)-c)^2/\sigma^2}\alpha(x)e^{2\pi iN\phi(x)},$$

where $\alpha(x)$ is a smooth amplitude function, $\phi(x)$ a smooth phase function, and σ is a band parameter that controls the width of the signal.

To understand how large the bandwidth should be in order to obtain accurate local wave vector estimates by the SSCT, we assume $\sigma=\Theta\left(N^{-\eta}\right)$ and show that the SSCT gives good estimates when $\eta<t$ and N is sufficiently large. In the space domain, a generalized curvelet at the scale $a=O\left(N\right)$ has a width $O\left(N^{-t}\right)$. $\sigma\geq N^{-\eta}$ with $\eta<t$ indicates that the bandwidth σ of

$e^{-(\phi(x)-c)^2/\sigma^2}\alpha(x)e^{2\pi iN\phi(x)}$ can be almost as narrow as the width of a generalized curvelet that sharing the same wave number $O(N)$, when N is sufficiently large.

Definition 3.4.1. *For any $c \in \mathbb{R}$, $N > 0$ and $M > 0$, a function $f(x) = e^{-(\phi(x)-c)^2/\sigma^2}\alpha(x)e^{2\pi iN\phi(x)}$ is a banded intrinsic mode function of type (M, N, η), if $\alpha(x)$ and $\phi(x)$ satisfy*

$$\begin{aligned} &\alpha(x) \in C^\infty, \qquad |\nabla\alpha(x)| \le M, \qquad 1/M \le \alpha(x) \le M, \\ &\phi(x) \in C^\infty, \quad 1/M \le |\nabla\phi(x)| \le M, \quad |\nabla^2\phi(x)| \le M, \\ &\qquad\qquad \text{and} \quad \sigma \ge N^{-\eta}. \end{aligned}$$

Definition 3.4.2. *A function $f(x)$ is a well-separated superposition of type (M, N, η, s, t, K) if*

$$f(x) = \sum_{k=1}^{K} f_k(x),$$

where each $f_k(x) = e^{-(\phi_k(x)-c_k)^2/\sigma_k^2}\alpha_k(x)e^{2\pi iN\phi_k(x)}$ is a banded intrinsic mode function (IMT) of type (M, N_k, η) with $N_k \ge N$ and they satisfy the separation condition: $\forall a \in [1, \infty)$ and $\forall\theta \in [0, 2\pi)$, there is at most one banded intrinsic mode function f_k satisfying that

$$|A_a^{-1}R_\theta^{-1}(a \cdot u_\theta - N_k\nabla\phi_k(b))| \le 1.$$

We denote by $F(M, N, \eta, s, t, K)$ the set of all such functions.

We are ready to discuss the robustness of the 2D SSCT illustrated in the next two theorems. We refer to their proofs in Appendix B.

Theorem 3.4.3. *Suppose the 2D mother wave packet is of type (ϵ, m), for any fixed $\epsilon \in (0, 1)$ and any fixed integer $m \ge 0$. Suppose $g(x) = f(x) + e(x)$, where $e(x) \in L^\infty$ is a small error term that satisfies $\|e\|_{L^\infty} \le \epsilon_1$ for some $\epsilon_1 \ge 0$. For any $p \in (0, \frac{1}{2}]$, let $\delta = \sqrt{\epsilon} + \epsilon_1^{\frac{1}{2}-p}$. Define*

$$R_\delta = \left\{(a, \theta, b) : |W_f(a, \theta, b)| \ge a^{-\frac{s+t}{2}}\delta\right\},$$

$$S_\delta = \{(a, \theta, b) : |W_f(a, \theta, b)| \ge \delta\},$$

and

$$Z_k = \left\{(a, \theta, b) : |A_a^{-1}R_\theta^{-1}(a \cdot u_\theta - N_k\nabla\phi_k(b))| \le 1\right\}$$

for $1 \le k \le K$. For fixed M, m, s, t, η, ϵ and K, there exists

$$N_0(M, m, s, t, \eta, \epsilon, K) \simeq \max\left\{\epsilon^{\frac{-1}{1-t}}, \epsilon^{\frac{-2}{t-\eta}}, \epsilon^{\frac{-2}{2s-1}}\right\}$$

such that for any $N > N_0$ and $f(x) \in F(M, N, \eta, s, t, K)$ the following statements hold.

(i) $\{Z_k : 1 \le k \le K\}$ *are disjoint and* $S_\delta \subset R_\delta \subset \bigcup_{1\le k\le K} Z_k$.

(ii) For any $(a,\theta,b) \in R_\delta \cap Z_k$,

$$\frac{|v_g(a,\theta,b) - N_k \nabla\phi_k(b)|}{|N_k \nabla\phi_k(b)|} \lesssim \sqrt{\epsilon} + \epsilon_1^p.$$

(iii) For any $(a,\theta,b) \in S_\delta \cap Z_k$,

$$\frac{|v_g(a,\theta,b) - N_k \nabla\phi_k(b)|}{|N_k \nabla\phi_k(b)|} \lesssim N_k^{-\frac{s+t}{2}} (\sqrt{\epsilon} + \epsilon_1^p).$$

Theorem 3.4.3 justifies the robustness of the 2D SSCT to a bounded perturbation. Next theorem below will illustrate its robustness against additive zero mean stationary Gaussian noise.

Theorem 3.4.4. *Suppose the 2D mother wave packet is of type* (ϵ, m), *for any fixed* $\epsilon \in (0,1)$ *and any fixed integer* $m \ge \max\left\{\frac{2(1+s)}{1-t}, \frac{2}{1-t} + 4\right\}$. *Suppose* $g(x) = f(x) + e$, *where* e *is a zero mean stationary Gaussian process with a spectrum denoted by* $\widehat{e}(\xi)$ *and* $\max_\xi |\widehat{e}(\xi)| \le \epsilon^{-1}$. *Define* $M_a = \max_{|\xi|<1} \widehat{e}\,(R_\theta A_a \xi + a \cdot u_\theta)$. *For any* $p \in (0, \frac{1}{2}]$ *and* $q > 0$, *let* $\delta_a = M_a^{(\frac{1}{2}-p)/(1+q)} + \sqrt{\epsilon}$,

$$R_{\delta_a} = \left\{(a,\theta,b) : |W_f(a,\theta,b)| \ge a^{-\frac{s+t}{2}} \delta_a\right\},$$

$$S_{\delta_a} = \left\{(a,\theta,b) : |W_f(a,\theta,b)| \ge \delta_a\right\},$$

and

$$Z_k = \left\{(a,\theta,b) : |A_a^{-1} R_\theta^{-1} \left(a \cdot u_\theta - N_k \nabla\phi_k(b)\right)| \le 1\right\}$$

for $1 \le k \le K$. *For fixed* M, m, s, t, η, ϵ *and* K, *there exists*

$$N_0\left(M, m, s, t, \eta, \epsilon, K\right) \simeq \max\left\{\epsilon^{\frac{-1}{1-t}}, \epsilon^{\frac{-2}{t-\eta}}, \epsilon^{\frac{-2}{2s-1}}\right\}$$

such that for any $N > N_0$ *and* $f(x) \in F\left(M, N, \eta, s, t, K\right)$ *the following statements hold.*

(i) $\{Z_k : 1 \le k \le K\}$ *are disjoint.*

(ii) If $(a,\theta,b) \in R_{\delta_a}$, *then* $(a,\theta,b) \in \bigcup_{1\le k\le K} Z_k$ *with a probability at least*

$$1 - e^{-O\left(N_k^{-(s+t)} M_a^{-q/(1+q)}\right)} + O\left(\frac{\epsilon}{N_k^{m(1-t)}}\right).$$

(iii) If $(a,\theta,b) \in S_{\delta_a}$, *then* $(a,\theta,b) \in \bigcup_{1\le k\le K} Z_k$ *with a probability at least*

$$1 - e^{-M_a^{-q/(1+q)}} + O\left(\frac{\epsilon}{N_k^{m(1-t)}}\right).$$

(iv) If $(a,\theta,b) \in R_{\delta_a} \cap Z_k$ *for some* k*, then*

$$\frac{|v_g(a,\theta,b) - N_k \nabla\phi_k(b)|}{|N_k \nabla\phi_k(b)|} \lesssim \sqrt{\epsilon} + M_a^{p/(1+q)}$$

is true with a probability at least

$$\left(1 - e^{-O\left(N_k^{2-s-3t} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-(3s+t)} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-2-s-t} M_a^{-q/(1+q)}\right)}\right)$$
$$+O\left(\frac{\epsilon}{N_k^{(m-4)(1-t)-2}}\right) + O\left(\frac{\epsilon}{N_k^{m-2-mt-2s}}\right).$$

(v) If $(a,\theta,b) \in S_{\delta_a} \cap Z_k$ *for some* k*, then*

$$\frac{|v_g(a,\theta,b) - N_k \nabla\phi_k(b)|}{|N_k \nabla\phi_k(b)|} \lesssim N_k^{-(s+t)/2}\left(\sqrt{\epsilon} + M_a^{p/(1+q)}\right)$$

is true with a probability at least

$$\left(1 - e^{-O\left(N_k^{2-2t} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-2s} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-2} M_a^{-q/(1+q)}\right)}\right)$$
$$+O\left(\frac{\epsilon}{N_k^{(m-4)(1-t)-2}}\right) + O\left(\frac{\epsilon}{N_k^{m-2-mt-2s}}\right).$$

Similar to the discussion in previous sections, the requirement for $N_0 = \max\left\{\epsilon^{\frac{-1}{1-t}}, \epsilon^{\frac{-2}{t-\eta}}, \epsilon^{\frac{-2}{2s-1}}\right\}$ can be further optimized if we impose extra conditions on the nonlinearity of IMTs and consider the polynomial decaying of the mother curvelet in the frequency domain.

Up to now, we have proved that the SSCT is able to accurately and robustly estimate the local wave vectors of banded IMTs, if their essential supports can be modeled by a Gaussian function with an essential support larger than the width of a curvelet sharing the same order of oscillation. Before closing this section, we would like to emphasize a new understanding of the results obtained in those theorems in this section: if the amplitude function of an IMT has a vanishing boundary, then the vanishing rate can be almost as fast as the oscillation. If an IMT has a sharp boundary, the estimates provided by synchrosqueezed transforms are reliable at the locations almost $O(1)$ wave lengths away from the boundary (see Figure 3.1 right). As a corollary in 1D cases, a similar conclusion is true as illustrated in Figure 3.1 left.

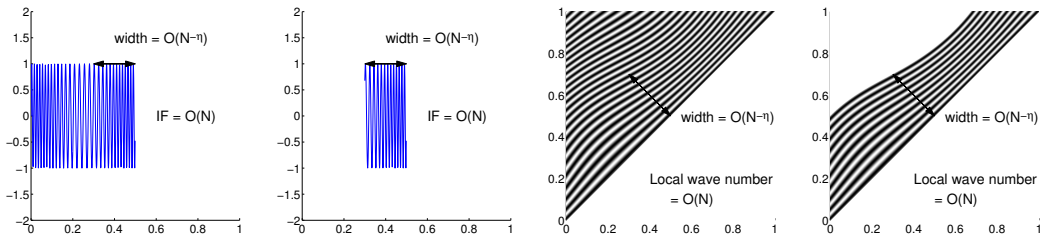


Figure 3.1: The reliable estimation area of an IMT with boundaries. Left and middle right: synchrosqueezed transforms can provide accurate estimates at the locations $O(N^{-\eta})$ away from a boundary. Middle left and right: if the width of an IMT is less than $O(N^{-\eta})$, the accuracy of synchrosqueezed transforms is still an open problem.

Chapter 4

Discrete Synchrosqueezed Transforms

4.1 Fast Discrete SSWPT and Mode Decomposition

4.1.1 Implementation

In this section, we describe in detail the discrete synchrosqueezed wave packet transform proposed mainly following the work in [182, 178] with Lexing Ying. Let us first recall the continuous setting. For a given superposition $f(x)$ of several well-separated components in $\mathbb{R}^d$, the synchrosqueezed wave packet transform consists of the following steps:

(i) Apply the wave packet transform to obtain $W_f(a,b)$ and the gradient $\nabla_b W_f(a,b)$;

(ii) Compute the approximate instantaneous frequency or local wavevector $v_f(a,b)$ and perform synchrosqueezing to get $T_f(v,b)$;

(iii) Use a clustering algorithm to identify the support of the new representation $T_f(v,b)$ of different intrinsic mode type functions;

(iv) Reconstruct each intrinsic mode type function using the dual frame.

In order to realize these steps in the discrete setting, we first introduce a discrete implementation of the wave packet transform in the first part of Section 4.1.1. The full discrete algorithm will then be discussed in the second part of Section 4.1.1. A few numerical examples will be provided in 4.1.2 to demonstrate the efficiency of these algorithms.

Multi-dimensional discrete wave packet transform

For simplicity, we consider functions that are periodic over the unit square $[0,1)^d$ in a d-dimensional space. Let

$$X = \{(n/L : n \in \mathbb{Z}^d, \quad 0 \le n_j < L, \text{ for } 1 \le j \le d\}$$

be the spatial grid of size L in each dimension at which these functions are sampled. The corresponding Fourier grid is

$$\Xi = \{\xi \in \mathbb{Z}^d : -L/2 \le \xi_j < L/2, \text{ for } 1 \le j \le d\}.$$

For a function $f(x) \in \ell^d(X)$, the discrete forward Fourier transform is defined by

$$\hat{f}(\xi) = \frac{1}{L^{d/2}} \sum_{x \in X} e^{-2\pi i x \cdot \xi} f(x),$$

while the discrete inverse Fourier transform of $g(\xi) \in \ell^d(\Xi)$ is

$$\check{g}(x) = \frac{1}{L^{d/2}} \sum_{\xi \in \Xi} e^{2\pi i x \cdot \xi} g(\xi).$$

In both transforms, the factor $\frac{1}{L^{d/2}}$ ensures that these discrete transforms are isometries between $\ell_d(X)$ and $\ell_d(\Xi)$.

A filterbank-based time-frequency transform is a natural choice to design the discrete wave packet transform due to the localization requirement of wave packets in the frequency domain. It also enjoys fast implementation. In order to design a discrete wave packet transform using the filterbank-based method, we need to specify how to decimate the momentum space and the position space. Let us focus on the 1D case and first consider the momentum space. In the continuous setting, the Fourier transform $\widehat{w_{ab}}(\xi)$ of the wave packets for a fixed a value have the profile

$$|a|^{-sd/2} \widehat{w}(|a|^{-s}(\xi - a)), \tag{4.1}$$

modulo complex modulation. In 1D transform, we sample the Fourier domain $[-L/2, L/2]$ with a set of points a (as shown in Figure 4.1 marked in blue) and associate with each $a \in A$ a window function $g_a(\xi)$ (see Figure 4.1 in black) that behaves qualitatively as $\widehat{w}(|a|^{-s}(\xi - a))$ essentially centered at a. More precisely, $g_a(\xi)$ is required to satisfy the following conditions:

- $g_a(\xi)$ is non-negative and centered at a with an essentially compact support of width $L_a = O(|a|^s)$; $g_a(\xi)$ decays sufficiently fast outside this essential support;
- $g_a(|a|^s \tau + a)$ is a sufficiently smooth function of τ, so that the discrete wave packets decay rapidly in the spatial domain;

- $C_1 \leq \int |g_a(|a|^s\tau + a)|^2 d\tau \leq C_2$ for constants $C_1, C_2 > 0$ which are independent of a;
- In addition, for any $\xi \in [-L/2, L/2)$, $\sum_{a\in A} |g_a(\xi)|^2 = 1$.

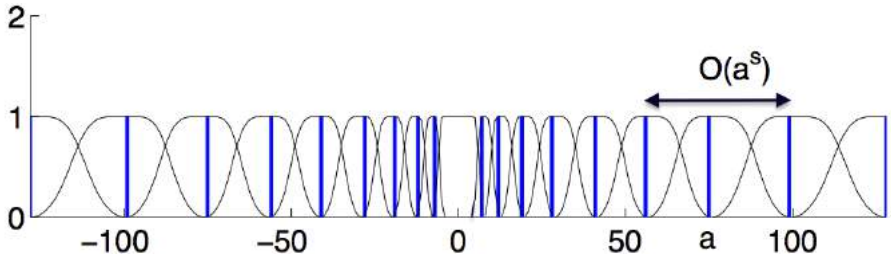


Figure 4.1: The 1D sample set A in blue. Each stick represents a point a in A. Each a is associated with a 1D window function $g_a(\xi)$ in black of size $O(a^s)$ in the frequency domain.

In higher dimensional cases, the set A and functions $\{g_a(\xi), a \in A\}$ can be generated by tensor product using the results in 1D.

One possible way to specify the set A and the functions $\{g_a(\xi), a \in A\}$ is to follow the constructions of the wave atom frame in [49] or the Gaussian wave packets of [142]. In both constructions, the parabolic scaling $s = 1/2$ is used in order to represent the oscillatory patterns efficiently. However, in the current setting, the proposed wave packet transform requires $s \in (1/2, 1)$ and hence one needs to increase the support of $g_a(\xi)$ accordingly. We refer to [49, 142] for more detailed discussions. The above conditions for $g_a(\xi), a \in A$ also impose a constraint on the sampling density of the set A. In the frequency plane, the set A becomes denser near the origin and sparser for large ξ. A straightforward calculation shows that the total number of samples in A is of order $O(L^{(1-s)d})$.

The decimation of the position space is much easier; we simply discretize it with a uniform grid of size L_B in each dimension as follows:

$$B = \{n/L_B : n \in \mathbb{Z}^d, \quad 0 \leq n_j < L_B, \text{ for } 1 \leq j \leq d\}.$$

As we shall see, the only requirement is that $L_B \geq \max_{a\in A} L_a$ so that the discrete wave packets can form a frame.

For each fixed $a \in A$ and $b \in B$, the discrete wave packet, still denoted by $w_{ab}(x)$ without causing much confusion, is defined through its Fourier transform as

$$\widehat{w_{ab}}(\xi) = \frac{1}{L_B^{d/2}} e^{-2\pi i b\cdot\xi} g_a(\xi) \tag{4.2}$$

for $\xi \in \Xi$. In fact, to match the quantity in (4.1), one should define

$$\widehat{w_{ab}}(\xi) = \frac{1}{L_a^{d/2}} e^{-2\pi i b\cdot\xi} g_a(\xi) \tag{4.3}$$

as it was defined in [182]. However, (4.3) would lead to weak spectral energy of in the high frequency domain. A high frequency wave-like component becomes hardly visible after synchrosqueezed transform. Hence, in practice, we adopt the definition in (4.2) instead of (4.3). Since $g_a(\xi)$ is centered at a and has a support of width $L_a = O(|a|^s)$, this function fits into the scaling of wave packets. Applying the discrete inverse Fourier transform provides its spatial description

$$w_{ab}(x) = \frac{1}{(L\cdot L_B)^{d/2}} \sum_{\xi\in\Xi} e^{2\pi i(x-b)\cdot\xi} g_a(\xi).$$

For a function $f(x)$ defined on $x \in X$, the discrete wave packet transform is a map from $\ell_2(X)$ to $\ell_2(A\times B)$, defined by

$$W_f(a,b) = \langle w_{ab}, f\rangle = \langle \widehat{w_{ab}}, \hat{f}\rangle = \int \overline{\widehat{w_{ab}(\xi)}}\hat{f}(\xi)d\xi = \frac{1}{L_B^{d/2}} \sum_{\xi\in\Xi} e^{2\pi i b\cdot\xi} g_a(\xi)\hat{f}(\xi). \tag{4.4}$$

The following result shows that $\{w_{ab}, (a,b)\in A\times B\}$ forms a tight frame.

Proposition 4.1.1. *For any function $f(x)$ in $\ell_2(X)$, we have*

$$\sum_{a\in A, b\in B} |W_f(a,b)|^2 = \|f\|_2^2.$$

Proof. From the definition of the wave packet transform, we have

$$\begin{aligned}\sum_{a\in A,b\in B} |W_f(a,b)|^2 &= \sum_{a\in A,b\in B} \left|\sum_{\xi\in\Xi} \frac{1}{L_B^{d/2}} e^{2\pi i b\cdot\xi} g_a(\xi)\hat{f}(\xi)\right|^2 \\ &= \sum_{a\in A}\sum_{\xi\in\Xi} \left|g_a(\xi)\hat{f}(\xi)\right|^2 \\ &= \sum_{\xi\in\Xi} |\hat{f}(\xi)|^2.\end{aligned}$$

□

For a function $h(a,b)$ in $\ell_2(A\times B)$, the transpose of the wave packet transform is given by

$$W_h^t(x) := \sum_{a\in A,b\in B} h(a,b) w_{ab}(x). \tag{4.5}$$

The next result shows that this transpose operator allows us to reconstruct $f(x), x \in X$ from its wave packet transform $W_f(a,b), (a,b) \in A \times B$.

Proposition 4.1.2. *For any function $f(x)$ with $x \in X$,*

$$f(x) = \sum_{a\in A, b\in B} W_f(a,b) w_{ab}(x).$$

Proof. Let us consider the Fourier transform of the right hand side. It is equal to

$$\begin{aligned}
&\sum_{a\in A,b\in B} \left(\sum_{\eta\in\Xi} \frac{1}{L_B^{d/2}} e^{2\pi i b\cdot\eta} g_a(\eta)\hat{f}(\eta) \right) \cdot \frac{1}{L_B^{d/2}} e^{-2\pi i b\cdot\xi} g_a(\xi) \\
=&\sum_{a\in A} \left(\sum_{\eta\in\Xi} \frac{1}{L_B^{d}} \left(\sum_{b\in B} e^{2\pi i b\cdot(\eta-\xi)} g_a(\eta)\hat{f}(\eta) \right) \right) g_a(\xi) \\
=&\sum_{a\in A} (g_a(\xi))^2 \hat{f}(\xi) = \hat{f}(\xi),
\end{aligned}$$

where the second step uses the fact that in the η sum only the term with $\eta = \xi$ yields a non-zero contribution. □

Let us now turn to the discrete approximation of $\nabla_b W_f(a,b)$. From the continuous definition, we have

$$\nabla_b W_f(a,b) = \nabla_b \langle \widehat{w_{ab}}, \hat{f} \rangle = \langle -2\pi i \xi \widehat{w_{ab}}(\xi), \hat{f}(\xi) \rangle.$$

Therefore, we define the discrete gradient $\nabla_b W_f(a,b)$ in a similar way

$$\nabla_b W_f(a,b) = \sum_{\xi\in\Xi} \frac{1}{L_B^{d/2}} 2\pi i \xi e^{2\pi i b\cdot\xi} g_a(\xi)\hat{f}(\xi). \tag{4.6}$$

The above definitions give rise to fast algorithms for computing the forward wave packet transform, its transpose, and the discrete gradient operator. All three algorithms heavily rely on the fast Fourier transform (FFT). For the forward transform

$$W_f(a,b) = \left(\frac{1}{L_B^{d/2}} \sum_{\xi\in\Xi} e^{2\pi i b\cdot\xi} g_a(\xi)\hat{f}(\xi) \right),$$

we have the following algorithm

Algorithm 4.1.3. *Forward transform from $f(x)$ to $W_f(a,b)$*

1: *Compute $\hat{f}(\xi)$ with $\xi \in \Xi$ from $f(x)$ with $x \in X$ using a forward FFT of size L.*
2: ***for** each $a \in A$ **do***
3: *Form $g_a(\xi)\hat{f}(\xi)$ on the support of $g_a(\xi)$*

4: *Wrap the result modulo L_B onto the domain $[-L_B/2, L_B/2)^d$*
5: *Apply an inverse FFT of size L_B to the wrapped result to get $W_f(a,b)$ for all $b \in B$*
6: ***end for***

The transpose operator (4.5) can be written equivalently in the Fourier domain as

$$\hat{W_h^t}(\xi) = \sum_{a\in A, b\in B} h(a,b)\frac{1}{L_B^{d/2}} e^{-2\pi i b\cdot\xi} g_a(\xi) = \sum_{a\in A}\left(\sum_{b\in B}\frac{1}{L_B^{d/2}} h(a,b)e^{-2\pi i b\cdot\xi}\right) g_a(\xi),$$

which suggests the following algorithm for the transpose operator:

Algorithm 4.1.4. *Transpose operator from $h(a,b)$ to $W_h^t(x)$*

1: ***for*** *each $a \in A$* ***do***
2: *Apply a forward FFT of size L_B to $h(a,b)$*
3: *Unwrap the result modulo L_B onto the support of $g_a(\xi)$*
4: *Multiply the unwrapped data with $g_a(\xi)$ and add the product to get $\hat{f}(\xi)$*
5: ***end for***
6: *Compute $f(x)$ with $x \in X$ from $\hat{f}(\xi)$ with $\xi \in \Xi$ using an inverse FFT of size L.*

To implement the discrete gradient operator in (4.6), we have the following algorithm.

Algorithm 4.1.5. *Discrete gradient operator from $f(x)$ to $\nabla_b W_f(a,b)$*

1: *Compute $\hat{f}(\xi)$ with $\xi \in \Xi$ from $f(x)$ with $x \in X$ using a forward FFT of size L.*
2: ***for*** *each $a \in A$* ***do***
3: *Form $2\pi i\xi g_a(\xi)\hat{f}(\xi)$ on the support of $g_a(\xi)$*
4: *Wrap the result modulo L_B onto the domain $[-L_B/2, L_B/2)^d$*
5: *Apply an inverse FFT of size L_B to each component of the wrapped result to get $\nabla_b W_f(a,b)$ for all $b \in B$*
6: ***end for***

As we mentioned earlier, the conditions on $\{g_a(\xi), a \in A\}$ imply that there are $O(L^{d(1-s)})$ samples in set A. A straightforward calculation shows that the computational cost of all three algorithms is $O(L^d \log L + L^{d(1-s)} L_B^d \log L_B)$ with $L_B \geq \max_{a\in A} L_a = O(L^s)$. If we choose L_B to be of the same order as L^s, the complexity of these algorithms is $O(L^d \log L)$, which is the cost of an FFT on a Cartesian grid with L grid points in each dimension.

Description of the full algorithm

With the discrete transforms and their fast algorithms available, we now go through the steps of the synchrosqueezed wave packet transform.

For a given function $f(x)$ defined on $x \in X$, we apply Algorithm 4.1.3 to compute $W_f(a,b)$ and Algorithm 4.1.5 to compute $\nabla_b W_f(a,b)$. The approximate local wavevector $v_f(a,b)$ is then

estimated by

$$v_f(a,b) = \frac{\nabla_b W_f(a,b)}{2\pi i W_f(a,b)}$$

for $a \in A, b \in B$ with $W_f(a,b) \neq 0$. In view of Theorem 2.2.7, a threshold $|W_f(a,b)| \geq |a|^{-ds/2}\sqrt{\epsilon}$ ($|a| \geq 1$) is necessary. Since we adopt (4.2) instead of (4.3) in the numerical implementation, we only need a uniform threshold independent of the scale a. Following Theorem 2.2.7, we define a discrete set R_ϵ with

$$R_\epsilon = \{(a,b) : a \in A, b \in B, |W_f(a,b)| \geq \sqrt{\epsilon}\}$$

and $v_f(a,b)$ provides an approximate estimate for the local wavevector only for $(a,b) \in R_\epsilon$.

To specify the synchrosqueezed energy distribution $T_f(v,b)$, we first place in the Fourier domain a d-dimensional Cartesian grid of step-size Δ:

$$V = \{n\Delta : n \in \mathbb{Z}^d\}.$$

At each $v = n\Delta \in V$, we associate a cell D_v centered at v

$$D_v = \prod_{j=1}^{d} \left[(n_j - \frac{1}{2})\Delta, (n_j + \frac{1}{2})\Delta\right).$$

Then the discrete synchrosqueezed energy distribution is defined as

$$T_f(v,b) = \sum_{(a,b)\in R_\epsilon : \Re\mathfrak{e} v_f(a,b) \in D_v} |W_f(a,b)|^2.$$

It is straightforward to check that

$$\sum_{v\in V, b\in B} T_f(v,b) = \sum_{(a,b)\in R_\epsilon} |W_f(a,b)|^2 \leq \|f\|_2^2$$

where the last inequality comes from Proposition 4.1.1 and the fact that R_ϵ is a subset of $A \times B$.

Suppose that $f(x)$ is a superposition of K well-separated intrinsic mode functions:

$$f(x) = \sum_{k=1}^{K} f_k(x) = \sum_{k=1}^{K} \alpha_k(x) e^{2\pi i N_k \phi_k(x)}.$$

From the previous discussion, we know that, for each $b \in B$, $v_f(a,b)$ points approximately to one of $N_k \nabla\phi_k(b)$, depending on a. Therefore, after synchrosqueezing, $T_f(v,b)$ is essentially supported in the phase space near the K "discrete" surfaces $\{(N_k\nabla\phi_k(b), b), b \in B\}$. The next step is to decompose the essential support of $T_f(v,b)$ into K clusters, one for each intrinsic mode type function, through a suitable clustering method. The resulting clusters are defined to be $U_1, \ldots, U_K$. In many cases,

the number of components K is not known a priori and needs to be discovered from the function $T_f(v, b)$.

We estimate the instantaneous frequencies or local wave vectors $N_k \nabla \phi_k(b)$ efficiently by identifying the energy peaks in U_k. To obtain finer estimates, we can compute the weighted average of $v_f(a, b)$ over each cluster U_k as follows

$$N_k \nabla \phi_k(b) = \frac{\sum_{(a,b):\mathfrak{Re} v_f(a,b) \in U_k} |W_f(a, b)|^2 \mathfrak{Re} v_f(a, b)}{\sum_{(v,b) \in U_k} T_f(v, b)}. \tag{4.7}$$

In the final step, we recover each intrinsic mode function by computing.

$$f_k(x) = \sum_{(a,b):\mathfrak{Re} v_f(a,b) \in U_k} W_f(a, b) w_{ab}(x).$$

This step can be carried out efficiently by restricting $W_f(a, b)$ to the set $\{(a, b) : \mathfrak{Re} v_f(a, b) \in U_k\}$ and applying Algorithm 4.1.4 to the restriction for each k.

4.1.2 Numerical Examples

This section presents several numerical examples to illustrate the proposed synchrosqueezed wave packet transforms. Throughout all examples, the threshold value ϵ is 10^{-4} and the size L of the Cartesian grid X of the discrete algorithm is 512. In the implementation of the discrete wave packet transforms, the scaling parameter s is equal to $2/3$, which is a good balance as discussed previously. We will only show 2D examples because the results in other dimensions are similar.

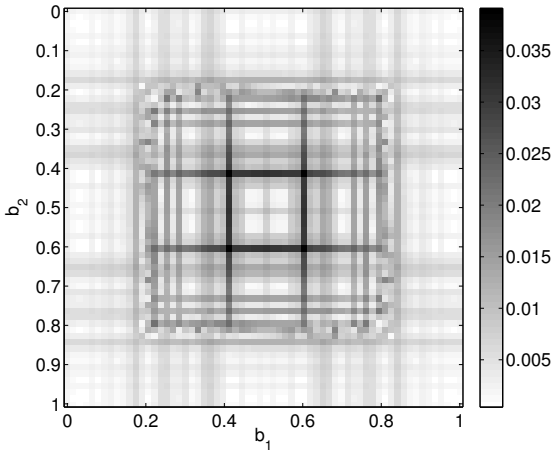


Figure 4.2: Relative error $R(b)$ of local wavevector estimation.

Local wavevector estimation

We first test the accuracy of the local wavevector information function $v_f(a,b)$. Let $f(x)$ be a deformed plane wave

$$f(x) = \alpha(x)e^{2\pi i N\phi(x)}.$$

Theorem 2.2.7 shows that, for each fixed point b in space, the estimate $v_f(a,b)$ approximates the local wavevector at b for any a that satisfies the condition $(a,b) \in R_\epsilon$. Though $v_f(a,b)$ for any such a provides an estimate of the local wavevector at b, it is more useful to combine them together to obtain a unique local wavevector estimate for each fixed b. More precisely, we compute the weighted average as in (4.7) to estimate local wave vectors. Denoting the weighted average as $v_f^m(b)$, we can define the (discrete) relative error $R(b)$ between $v_f^m(b)$ and the exact local frequency $N\nabla\phi(b)$ as

$$R(b) = \frac{|v_f^m(b) - N\nabla\phi(b)|}{|N\nabla\phi(b)|}.$$

We perform the above test on a deformed plane wave $f(x)$ with $\alpha(x) = 1$, $\phi(x) = \phi(x_1, x_2) = x_1 + x_2 + \beta\sin(2\pi x_1) + \beta\sin(2\pi x_2)$ with $\beta = 0.1$, and $N = 135$. The relative error $R(b)$ shown in Figure 4.2 is of order 10^{-2}, which agrees with Theorem 2.2.7 on that the relative approximation error is $O(\sqrt{\epsilon})$.

Intrinsic mode decomposition

Here $f(x)$ is a sum of two deformed plane waves

$$\begin{aligned} f(x) &= e^{2\pi i N\phi_1(x)} + e^{2\pi i N\phi_2(x)}, \\ \phi_1(x) &= \phi_1(x_1, x_2) = x_1 + x_2 + \beta\sin(2\pi x_1) + \beta\sin(2\pi x_2), \\ \phi_2(x) &= \phi_2(x_1, x_2) = -x_1 + x_2 - \beta\sin(2\pi x_1) + \beta\sin(2\pi x_2) \end{aligned}$$

with $N = 135$ and $\beta = 0.1$. The algorithm described in Section 4.1.1 is applied to $f(x)$ to extract these two components. Figure 4.3 summarizes the results of this test. The first row shows the superposition $f(x)$ (left) and the synchrosqueezed energy distribution $T_f(v,b)$ with b_1 fixed at 1 (right). For a fixed b_1 value $T_f(v,b)$ concentrates near two curves. More generally, in phase space $T_f(v,b)$ concentrates near two 2D surfaces. The second row shows the two sets U_1 and U_2 after the clustering steps. Finally, the third row plots the two reconstructed components.

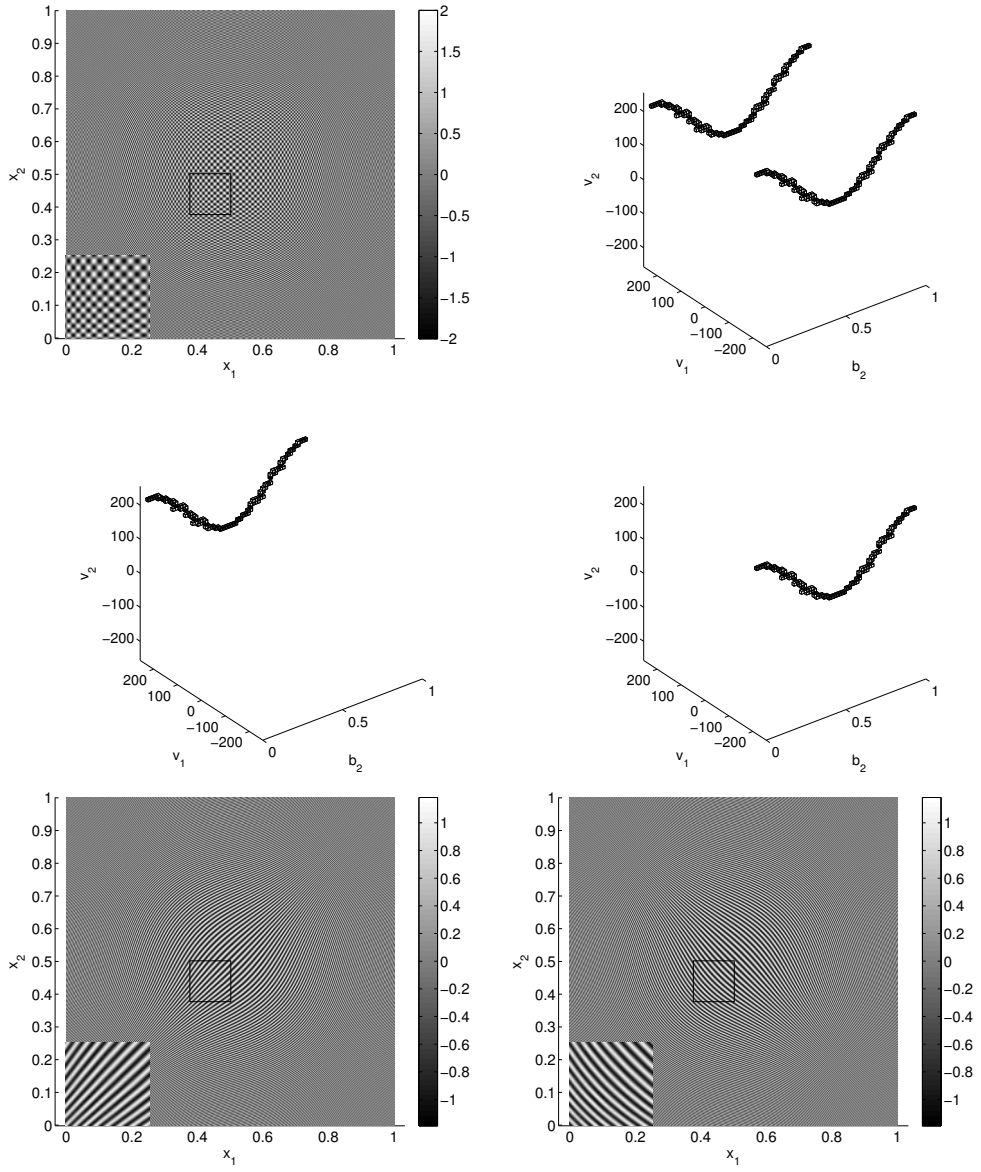


Figure 4.3: A mode decomposition without noise. Top-left: A superposition of two deformed plane waves with the bottom-left corner showing a zoomed-in view of the highlighted rectangle. Top-right: Synchrosqueezed energy distribution $T_f(v, b)$ at $b_1 = 1$. Second row: The support of $T_f(v, b)$ is clustered into two subsets. Third row: The two reconstructed components.

4.2 Fast Discrete SSCT and Mode Decomposition

4.2.1 Implementation

In this section, we describe the 2D discrete synchrosqueezed curvelet transform and the mode decomposition in detail. The description mainly follows the work in [184] with Lexing Ying. Let us first recall the continuous setting. Suppose $f(x)$ is a superposition of several well-separated components, the mode decomposition by the SSCT consists of the following steps:

(i) Apply the generalized curvelet transform to obtain $W_f(a,\theta,b)$ and the gradient $\nabla_b W_f(a,\theta,b)$;

(ii) Compute the local wave vector estimate $v_f(a,\theta,b)$ and concentrate the energy around it to get $T_f(v,b)$;

(iii) Separate the essential supports of the concentrated phase space energy distribution $T_f(v,b)$ into several components by clustering techniques;

(iv) Restrict $W_f(a,\theta,b)$ to each resulting component and reconstruct corresponding intrinsic mode functions using the dual frame.

We first introduce a discrete implementation of the generalized curvelet transform for Step (i) and Step (iv). The full discrete algorithm will then be summarized later.

2D Discrete generalized curvelet transforms

For simplicity, we consider periodic functions over the unit square $[0,1)^2$ in 2D. If it is not the case, the functions will be periodized by multiplying a smooth decaying function near the boundary of $[0,1)^2$. We follow basic discrete setting in Section 4.1. Recall that

$$X = \{(n_1/L, n_2/L) : 0 \le n_1, n_2, < L, n_1, n_2 \in \mathbb{Z}\}$$

is the $L \times L$ spatial grid at which these functions are sampled. The corresponding $L \times L$ Fourier grid is

$$\Xi = \{(\xi_1, \xi_2) : -L/2 \le \xi_1, \xi_2 < L/2, \xi_1, \xi_2 \in \mathbb{Z}\}.$$

For a function $f(x) \in \ell^2(X)$, the discrete forward Fourier transform is defined by

$$\widehat{f}(\xi) = \frac{1}{L} \sum_{x \in X} e^{-2\pi i x \cdot \xi} f(x).$$

For a function $g(\xi) \in \ell^2(\Xi)$, the discrete inverse Fourier transform is

$$\check{g}(x) = \frac{1}{L} \sum_{\xi \in \Xi} e^{2\pi i x \cdot \xi} g(\xi).$$

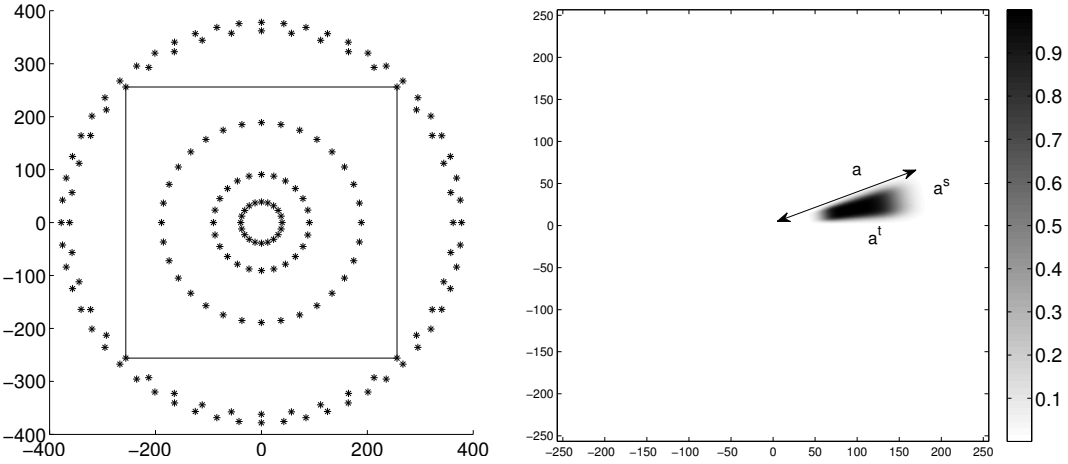


Figure 4.4: Left: Sampled point set A in Fourier domain for an image of size 512×512. Each point represents the center of the support of a window function. The window function centered at the origin is supported on a disk and is not indicated in this picture. The size of finest scale is set to be small (e.g. 16) in order to save memory. Right: An example of a fan-shaped window function $g_{a,\theta}(\xi)$.

In order to design a discrete curvelet transform, we need to specify how to decimate the Fourier domain in (a, θ) and the position space in b. Let us first consider the Fourier domain (a, θ). In the continuous setting, the Fourier transform $\widehat{w_{a\theta b}}(\xi)$ for fixed (a, θ) have the profile

$$a^{-\frac{s+t}{2}} \widehat{w}(A_a^{-1} R_\theta^{-1}(\xi - a \cdot u_\theta)), \tag{4.8}$$

modulo complex modulation. In the discrete setting, we sample the Fourier domain $[-L/2, L/2)^2$ with a set of points A (Figure 4.4 left) and associate with each $(a, \theta) \in A$ a window function $g_{a,\theta}(\xi)$ (Figure 4.4 right) that behaves qualitatively as $\widehat{w}(A_a^{-1} R_\theta^{-1}(\xi - a \cdot u_\theta))$. More precisely, $g_{a,\theta}(\xi)$ is required to satisfy the following conditions:

- $g_{a,\theta}(\xi)$ is non-negative and centered at $a \cdot u_\theta$ with a compact fan-shaped support of length $O(a^t)$ and width $O(a^s)$, which is approximately a directional elliptical support $\{\xi : |A_a^{-1} R_\theta^{-1}(\xi - a \cdot u_\theta)| \leq 1\}$.
- $g_{a,\theta}(R_\theta A_a \tau + a \cdot u_\theta)$ is a sufficiently smooth function of τ, thus making the discrete curvelets to decay rapidly in the spatial domain;
- $C_1 \leq \int |g_{a,\theta}(R_\theta A_a \tau + a \cdot u_\theta)|^2 d\tau \leq C_2$ for positive constants C_1 and C_2, independent of (a, θ);
- In addition, for any $\xi \in [-L/2, L/2)^2$, $\sum_{(a,\theta)\in A} |g_{a,\theta}(\xi)|^2 = 1$.

We follow the discretization and construction of frames in [23] to specify the set A and window functions, and refer to [19] for detail implementation. The difference here is that, we do not restrict

angular scaling parameter to $s = \frac{1}{2}$ and radial scaling parameter to $t = 1$. This allows us to adaptively adjust the size of tiles according to data structure. In the construction of the tiling in this article, the scaling parameters s and t remain constant as the scale changes.

The decimation of the position space b is much easier; we simply discretize it with an $L_B \times L_B$ uniform grid as follows:

$$B = \{(n_1/L_B, n_2/L_B) : 0 \le n_1, n_2 < L_B, n_1, n_2 \in \mathbb{Z}\}.$$

The only requirement is that L_B is large enough so that a sampling grid of size $L_B \times L_B$ can cover the supports of all window functions.

For each fixed $(a, \theta) \in A$ and $b \in B$, the discrete curvelet, still denoted by $w_{a\theta b}(x)$ without causing much confusion, is defined through its Fourier transform as

$$\widehat{w_{a\theta b}}(\xi) = \frac{1}{L_B} e^{-2\pi i b\cdot\xi} g_{a,\theta}(\xi) \tag{4.9}$$

for $\xi \in \Xi$. In fact, to match the quantity in (4.8), one should define

$$\widehat{w_{a\theta b}}(\xi) = \frac{1}{L_a} e^{-2\pi i b\cdot\xi} g_{a,\theta}(\xi) \tag{4.10}$$

with $L_a = a^{\frac{s+t}{2}}$. However, (4.10) would lead to weak spectral energy of in the high frequency domain. A high frequency wave-like component becomes hardly visible after synchrosqueezed transform. Hence, in practice, we adopt the definition in (4.9) instead of (4.10). Applying the discrete inverse Fourier transform provides its spatial description

$$w_{a\theta b}(x) = \frac{1}{L \cdot L_B} \sum_{\xi\in\Xi} e^{2\pi i (x-b)\cdot\xi} g_{a,\theta}(\xi).$$

For a function $f(x)$ defined on $x \in X$, the discrete curvelet transform is a map from $\ell_2(X)$ to $\ell_2(A \times B)$, defined by

$$W_f(a, \theta, b) = \langle w_{a\theta b}, f\rangle = \langle \widehat{w_{a\theta b}}, \widehat{f}\rangle = \frac{1}{L_B} \sum_{\xi\in\Xi} e^{2\pi i b\cdot\xi} g_{a,\theta}(\xi)\widehat{f}(\xi). \tag{4.11}$$

We can introduce an inner product on the space $\ell_2(A\times B)$ as follows: for any two functions $g(a, \theta, b)$ and $h(a, \theta, b)$,

$$\langle g, h\rangle = \sum_{(a,\theta)\in A, b\in B} \overline{g(a, \theta, b)} h(a, \theta, b).$$

The following result shows that $\{w_{a\theta b} : (a, \theta, b) \in A \times B\}$ forms a tight frame when equipped with this inner product.

Proposition 4.2.1. *For any function $f(x)$ for $x \in X$, we have*

$$\sum_{(a,\theta)\in A, b\in B} |W_f(a,\theta,b)|^2 = \|f\|_2^2.$$

Proof. From the definition of the curvelet transform, we have

$$\begin{aligned}\sum_{(a,\theta)\in A, b\in B} |W_f(a,\theta,b)|^2 &= \sum_{(a,\theta)\in A, b\in B} \left| \sum_{\xi\in\Xi} \frac{1}{L_B} e^{2\pi i b\cdot\xi} g_{a,\theta}(\xi)\widehat{f}(\xi) \right|^2 \\ &= \sum_{(a,\theta)\in A} \sum_{\xi\in\Xi} \left| g_{a,\theta}(\xi)\widehat{f}(\xi) \right|^2 \\ &= \sum_{\xi\in\Xi} |\widehat{f}(\xi)|^2.\end{aligned}$$

□

For a function $h(a,\theta,b)$ in $\ell_2(A\times B)$, the transpose of the curvelet transform is given by

$$W_h^t(x) := \sum_{(a,\theta)\in A, b\in B} h(a,\theta,b) w_{a\theta b}(x). \tag{4.12}$$

The next result shows that this transpose operator allows us to reconstruct $f(x), x \in X$ from its curvelet transform $W_f(a,\theta,b), (a,\theta,b) \in A\times B$.

Proposition 4.2.2. *For any function $f(x)$ with $x \in X$,*

$$f(x) = \sum_{(a,\theta)\in A, b\in B} W_f(a,\theta,b) w_{a\theta b}(x).$$

Proof. Let us consider the Fourier transform of the right hand side. It is equal to

$$\begin{aligned}&\sum_{(a,\theta)\in A, b\in B} \left(\sum_{\eta\in\Xi} \frac{1}{L_B} e^{2\pi i b\cdot\eta} g_{a,\theta}(\eta)\widehat{f}(\eta) \right) \cdot \frac{1}{L_B} e^{-2\pi i b\cdot\xi} g_{a,\theta}(\xi) \\ &= \sum_{(a,\theta)\in A} \left(\sum_{\eta\in\Xi} \frac{1}{L_B^2} \left(\sum_{b\in B} e^{2\pi i b\cdot(\eta-\xi)} g_{a,\theta}(\eta)\widehat{f}(\eta) \right) \right) g_{a,\theta}(\xi) \\ &= \sum_{(a,\theta)\in A} (g_{a,\theta}(\xi))^2 \widehat{f}(\xi) = \widehat{f}(\xi),\end{aligned}$$

where the second step uses the fact that in the η sum only the term with $\eta = \xi$ yields a nonzero contribution. □

Let us now turn to the discrete approximation of $\nabla_b W_f(a,\theta,b)$. From its continuous definition,

we have

$$\nabla_b W_f(a,\theta,b) = \nabla_b \langle \widehat{w}_{a\theta b}, \widehat{f} \rangle = \langle -2\pi i \xi \widehat{w}_{a\theta b}(\xi), \widehat{f}(\xi) \rangle.$$

Therefore, we define the discrete gradient $\nabla_b W_f(a,\theta,b)$ in a similar way

$$\nabla_b W_f(a,\theta,b) = \sum_{\xi \in \Xi} \frac{1}{L_B} 2\pi i \xi e^{2\pi i b\cdot\xi} g_{a,\theta}(\xi) \widehat{f}(\xi). \tag{4.13}$$

The above definitions give rise to fast algorithms for computing the forward generalized curvelet transform, its transpose, and the discrete gradient operator. All three algorithms heavily rely on the fast Fourier transform (FFT). The detailed implementation of these fast algorithms is similar to Algorithm 4.1.3, 4.1.4 and 4.1.5. The computational cost of all three algorithms is $O(L^2 \log L + L^{2-s-t} L_B^2 \log L_B)$ with L_B large enough so that a grid of size $L_B \times L_B$ can cover the supports of all window functions. If we choose L_B to be of the same order as L^t, the complexity of these algorithms is $O(L^{2+t-s} \log L)$.

Description of the full algorithm

We now go through the steps of the discrete synchrosqueezed curvelet transform.

For a given function $f(x)$ defined on $x \in X$, we apply fast algorithms to compute $W_f(a,\theta,b)$ and $\nabla_b W_f(a,\theta,b)$. Then the local wave vector estimate $v_f(a,\theta,b)$ is computed by

$$v_f(a,\theta,b) = \frac{\nabla_b W_f(a,\theta,b)}{2\pi i W_f(a,\theta,b)}$$

for $(a,\theta) \in A, b \in B$ with $W_f(a,\theta,b) \neq 0$.

In view of Theorem 2.3.7, a threshold $|W_f(a,\theta,b)| \geq |a|^{-(s+t)/2}\sqrt{\epsilon}$ $(a \geq 1)$ is necessary. Since we adopt (4.9) instead of (4.10) in the numerical implementation, we only need a uniform threshold independent of the scale a. Following Theorem 2.3.7, we define a discrete set R_ϵ with

$$R_\epsilon = \{(a,b) : a \in A, b \in B, |W_f(a,b)| \geq \sqrt{\epsilon}\}$$

and $v_f(a,b)$ provides an approximate estimate for the local wavevector only for $(a,b) \in R_\epsilon$.

The energy resulting in $\mathfrak{Re} v_f(a,\theta,b)$ should be stacked up to obtain $T_f(\mathfrak{Re} v_f(a,\theta,b),b)$. To realize this step, a two dimensional Cartesian grid of step size Δ is generated to discretize the Fourier domain of $T_f(v,b)$ in variable v as follows:

$$V = \{(n_1\Delta, n_2\Delta) : n_1, n_2 \in \mathbb{Z}\}.$$

At each $v = (n_1\Delta, n_2\Delta) \in V$, we associate a cell D_v centered at v

$$D_v = \left[(n_1 - \frac{1}{2})\Delta, (n_1 + \frac{1}{2})\Delta\right) \times \left[(n_2 - \frac{1}{2})\Delta, (n_2 + \frac{1}{2})\Delta\right).$$

Then $T_f(v, b)$ is estimated by

$$T_f(v, b) = \sum_{(a,\theta,b):\mathfrak{Re}v_f(a,\theta,b)\in D_v} |W_f(a, \theta, b)|^2.$$

It is straightforward to check that

$$\sum_{v\in V, b\in B} T_f(v, b) = \sum_{(a,b)\in R_\epsilon} |W_f(a, \theta, b)|^2 \leq \|f\|_2^2$$

where the last inequality comes from Proposition 4.2.1 and the fact that R_ϵ is a subset of $A \times B$.

Suppose that $f(x)$ is a superposition of K well-separated banded intrinsic mode type functions:

$$f(x) = \sum_{k=1}^{K} f_k(x) = \sum_{k=1}^{K} e^{-(\phi_k(x)-c_k)^2/\sigma_k^2} \alpha_k(x) e^{2\pi i N_k \phi_k(x)}.$$

In the discrete implementation, we choose a threshold parameter $\delta > 0$ and define the set S to be

$$\{(v, b) : v \in V, b \in B, T_f(v, b) \geq \delta\}.$$

After synchrosqueezing, $T_f(v, b)$ is essentially supported in the phase space near K "discrete" surfaces $\{(N_k\nabla\phi_k(b), b), b \in B\}$. Hence, under the separation condition given by Theorem 2.3.7, S will have K well-separated clusters $U_1, \ldots, U_K$, and they would be identified by a suitable clustering method.

Once we discover $U_1, \ldots, U_K$, we can define $W_{f_k}(a, \theta, b)$ by restricting $W_f(a, \theta, b)$ to the set $\{(a, \theta, b) : \mathfrak{Re}v_f(a, \theta, b) \in U_k\}$. Then, we can recover each intrinsic mode type function efficiently using the fast algorithm discussed to compute

$$f_k(x) = \sum_{(a,\theta)\in A, b\in B} W_{f_k}(a, \theta, b) w_{a\theta b}(x).$$

A weighted average of $v_f(a, \theta, b)$ similar to (4.7) gives good estimates of local wave vectors $N_k\nabla\phi_k(b)$.

4.2.2 Numerical Examples

In this section, we start with error analysis of local wave vector estimation using synchrosqueezed curvelet transform, and compare it with synchrosqueezed wave packet transform. Afterward, some mode decomposition examples of synthetic data will be presented to illustrate the efficiency of

proposed synchrosqueezed curvelet transform. For all the synthetic examples in this section, the size L of the Cartesian grid X of the discrete algorithm is 512, the threshold value $\epsilon = 10^{-4}$ for $W_f(a, \theta, b)$. The scaling parameters of synchrosqueezed curvelet transform are $t = 1 - \frac{1}{8}$ and $s = \frac{1}{2} + \frac{1}{8}$, as an appropriate balance as discussed previously. In the meantime, we chose $t = s = \frac{1}{2} + \frac{1}{8}$ to construct discrete synchrosqueezed wave packet transform for a reasonable comparison.

Local wave vector estimation

In Theorem 2.3.7, we have seen that the estimate $v_f(a, \theta, b)$ approximates the local wave vector at b, if $(a, b) \in R_\epsilon$. In such region, though $v_f(a, \theta, b)$ provides an accurate estimate of the local wave vector at each b, it is more rational to average them up to obtain a unique local wave vector estimate for each fixed b. By the definition of synchrosqueezed energy distribution, $T_f(\mathfrak{Re} v_f(a, \theta, b), b)$ truly reflects a natural weight of $v_f(a, \theta, b)$ in variables a and θ. More precisely, we compute similar weighted average as in (4.7) to estimate local wave vectors. Denoting the weighted average as $v_f^m(b)$, we can define the (discrete) relative error $R(b)$ between $v_f^m(b)$ and the exact local frequency $N\nabla\phi(b)$ as

$$R(b) = \frac{|v_f^m(b) - N\nabla\phi(b)|}{|N\nabla\phi(b)|}.$$

We test the accuracy for a noise free deformed plane wave $f(x) = \alpha(x)e^{2\pi i N\phi(x)}$ with $\alpha(x) = 1$, $\phi(x) = \phi(x_1, x_2) = x_1 + (1 - x_2) + 0.1\sin(2\pi x_1) + 0.1\sin(2\pi(1 - x_2))$, and $N = 135$ (see Figure 4.5 left). The relative error $R(b)$ of SSCT shown in Figure 4.5 (middle) is of order 10^{-2}, which agrees with Theorem 2.3.7 on that the relative approximation error is of order $O(\sqrt{\epsilon})$. The synchrosqueezed wave packet transform and the synchrosqueezed curvelet transform share the same accuracy in this case shown by Figure 4.5 middle and right.

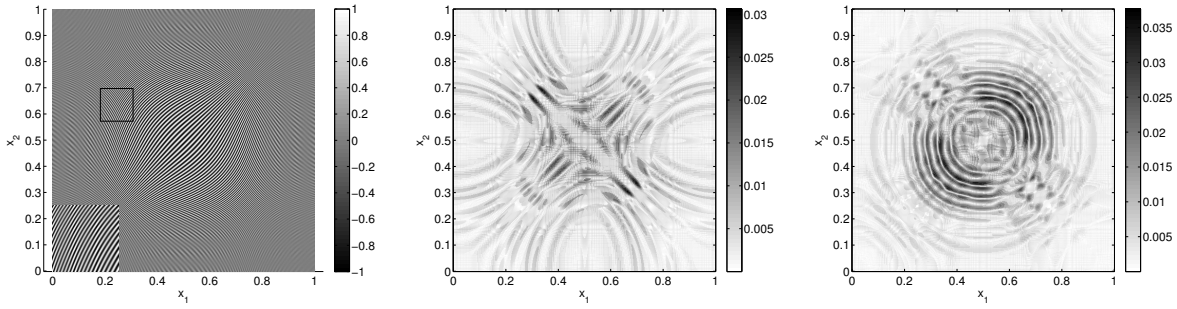


Figure 4.5: Left: A deformed plane wave propagating in the full space with zoomed-in data indicated by a rectangle. Middle: Relative error $R(b)$ of local wave vector estimation using SSCT. Right: Relative error $R(b)$ of local wave vector estimation given by SSWPT.

We compare the efficiency of SSCT and SSWPT in a noiseless case of a banded deformed plane wave $f(x) = e^{-(\phi(x)-c)^2/\sigma^2}\alpha(x)e^{2\pi i N\phi(x)}$ with the same parameters in last example and two more

parameters $c = 0.7$ and $\sigma = \frac{4}{135}$. As we discussed at the beginning of this subsection, $v_f(a, \theta, b)$ is only computed in the relevant region $|W_f(a, \theta, b)| \geq \sqrt{\epsilon}$. So, the relative error will be set to be zero at the position b such that $|W_f(a, \theta, b)| < \sqrt{\epsilon}$ for all (a, θ). The numerical result matches well with our theoretical prediction, showing that SSCT estimates local wave vectors of this banded wave-like component within a relative error of order $O(\sqrt{\epsilon})$. However, SSWPT fails the truth as we discussed in the section of introduction.

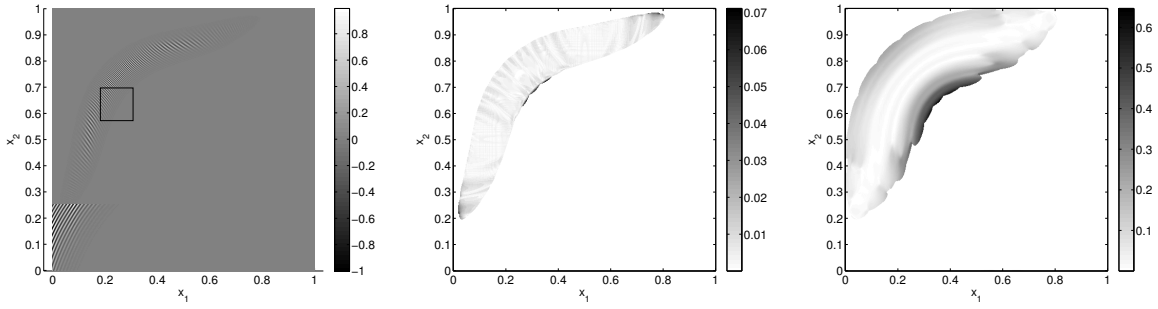


Figure 4.6: Left: A banded deformed plane wave. The zoomed-in data comes from the small rectangle. Middle: Relative error $R(b)$ of local wave vector estimation using SSCT. Right: Relative error $R(b)$ of local wave vector estimation given by SSWPT.

4.2.3 Intrinsic Mode Decomposition for Synthetic Data

In many applications, it is required to extract each component from a superposition. To show that our algorithm may provide a solution, we present some numerical examples of mode decompositions for highly oscillatory synthetic seismic data in noiseless and noisy cases (see Figure 4.7 top). Figure 4.7 shows the results of the application of our algorithm described in Section 4.2.1. On the left is a noiseless example and the example on the right has additive noise. Each mode of given data is accurately recovered in the noiseless case. In the noisy case, different modes with different propagation characters are completely separated. Each recovered mode practically reflects the curvature of corresponding mode in the original data, though there is some energy loss due to the threshold to remove noise.

In some other applications, one component might be disrupted (e.g. randomly shifted in this example), and it is required to remove such component and recover others. Here we randomly shift the first mode in the previous example in the vertical direction and apply our algorithm to recover the second mode. The numerical results summarized in Figure 4.8 show the capability of our algorithm to solve such a problem with or without noise. In this problem, the disrupted component can be considered as noise with high energy, i.e., this is a problem with very small signal-to-noise ratio. It is even more problematic that random shifting may create some texture similar to the mode to be

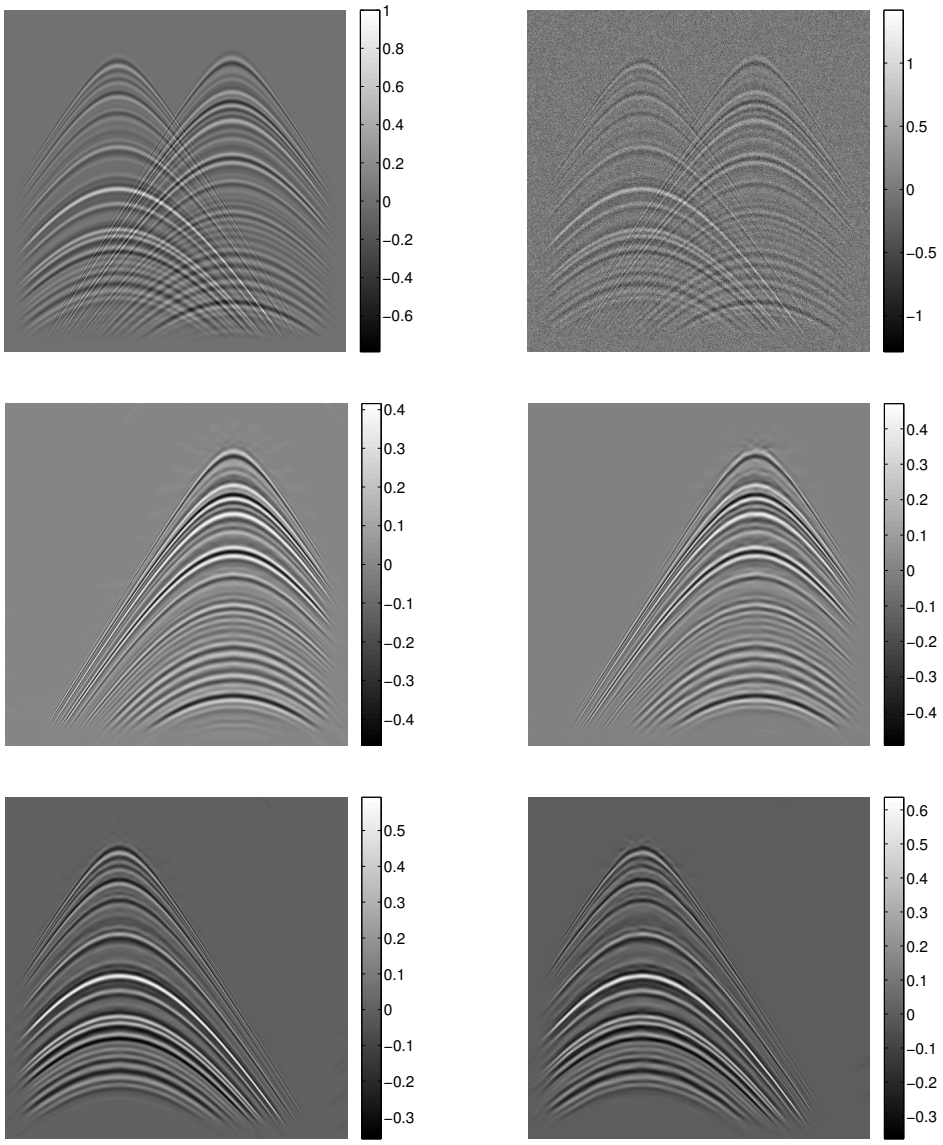


Figure 4.7: Example 2. Left: A mode decomposition without noise. Right: A mode decomposition with noise. Top: A superposition of two components. Second row: The first recovered relevant mode. Third row: The second recovered relevant mode.

recovered in some region. Fortunately, the synchrosqueezed representation is so concentrated that the resolution is still good enough to separate the mode from such similar texture by appropriately thresholding $T_f(a, \theta, b)$.

The left example in Figure 4.8 shows the result of noiseless data. The recovered mode looks

almost the same as the one recovered in noiseless example in Figure 4.7 bottom left, except some energy loss due to thresholding. It is of interest to add some background noise to see how well our algorithm is performing. Figure 4.8 right shows the result of the noisy case. The result (see Figure 4.8 bottom right) is almost identical with the recovered mode in Figure 4.7 bottom left.

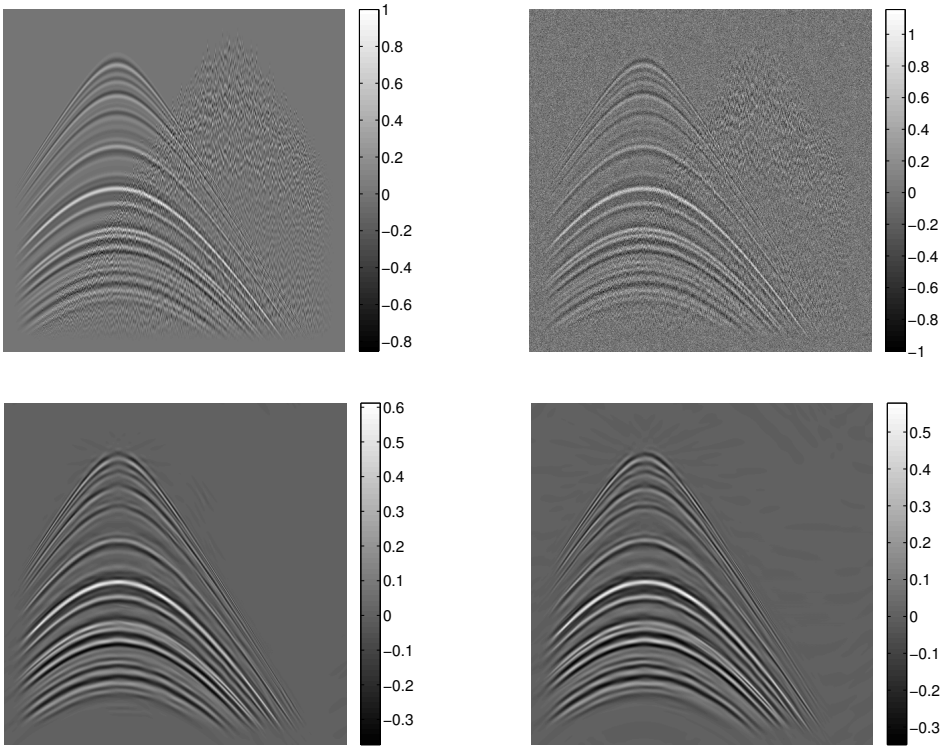


Figure 4.8: Example 3. Left: Mode identification without noise. Right: Mode identification with noise Top: A superposition of two components, one of which is disrupted by random shifting and need to be removed. Second row: The recovered relevant mode.

4.3 Numerical Robustness Analysis

In this section, we provide numerical examples to demonstrate the conclusions of those theorems in the robustness analysis in Chapter 3 and explain several ideas to obtain reliable instantaneous frequency or local wave vector information from extremely noisy data.

In all examples, we assume the given data $g(x) = f(x) + e(x)$ is defined in $[0,1]^n$, where $f(x)$ is the target signal, $e(x)$ is Gaussian white noise with a distribution $\sigma^2\mathcal{N}(0,1)$, and n is the number of dimensions. We would only focus on testing the robust performance of the SSWPT, since the

SSCT has similar properties. Detailed implementations of these transforms have been discussed in [178, 180, 182, 184], Section 4.1 and Section 4.2. As we have seen in the theorems in Chapter 3, a proper thresholding adaptive to the noise level after the wave packet/generalized curvelet transform is important to obtain an accurate instantaneous frequency/local wave vector estimate. We refer to [55, 56] for estimating noise level and [155] for designing thresholds for the SSWT. The generalization of these techniques for the SSWPT and the SSCT is straightforward.

Our main purpose in this section is to show the robustness properties of synchrosqueezed transforms with various scaling parameters. We compare the performance of the SSWPT with $s = 1/2 + k/8$, where $k = 1$, 2 and 3, in both noiseless cases and highly noisy cases. For the purpose of showing how the synchrosqueezing process is affected by heavy noise, we are using a small uniform threshold $\delta = 10^{-2}$ (rather than a threshold adaptive to noise level) and setting σ^2 such that the noise is overwhelming the original signal in all of our synthetic toy models. The accuracy tolerance in the theorems $\epsilon = 10^{-4}$.

4.3.1 Robustness Tests for 1D SST

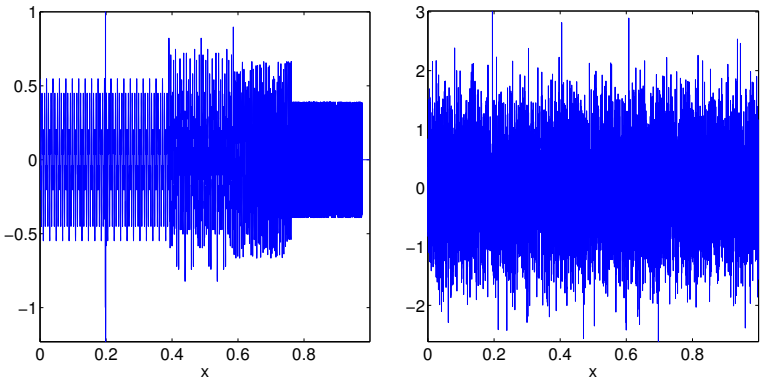


Figure 4.9: Left: A 1D synthetic benchmark signal. It is normalized using L^∞ norm. Right: A noisy signal generated with Gaussian white noise $0.75\mathcal{N}(0, 1)$.

We start from testing the 1D SSWPT. In some real applications, IMTs are only supported in a bounded domain or they have sharp changes in instantaneous frequencies. Hence, we would like to test a benchmark signal in which there is a component with a bounded support and an oscillatory instantaneous frequency, and a component with an exponential instantaneous frequency (see Figure 4.9). Of a special interest to test the performance of synchrosqueezed transforms for impulsive waves, a wavelet component is added in this signal at $x = 0.2$. The synthetic benchmark signal[1] is

[1]Prepared by Professor Mirko van der Baan and available at [153, 162].

generated using the example functions:

$$\begin{aligned} f_1(x) &= 0.6\cos(700\pi x); \\ f_2(x) &= 0.8\cos(300\pi x); \\ f_3(x) &= 0.7\cos(1300\pi x + 5\sin(20\pi x)); \\ f_4(x) &= \sin\left(\frac{80\pi 100^{5x/4}}{\ln(100)}\right); \\ f_5(x) &= 3e^{-50x^2}\cos(50x). \end{aligned}$$

The sampling rate of this signal is 8192 Hz and the instantaneous frequency range is $150 - 1600$ Hz. The Gaussian white noise in this example is $0.75\mathcal{N}(0,1)$. To make a fair comparison, we have tuned the size of the essential support of mother wave packets to obtain a good result for each kind of synchrosqueezed transforms.

Although we have not identified the optimal value of the scaling parameter s, it is clear from Theorem 3.2.2 and 3.2.3 that the synchrosqueezed transform with a smaller s is more robust. As shown in the second and the third rows in Figure 4.10, in the noisy cases, the synchrosqueezed energy distribution with $s = 0.625$ (in the first column) is better than the one with $s = 0.75$ (in the second column), which is better than the one with $s = 0.875$ (in the last column). This agrees with the conclusion in Theorem 3.2.2 and 3.2.3 that a smaller s results in a higher probability to obtain a better instantaneous frequency estimate.

Another key point of Theorem 3.2.2 and 3.2.3 is that a wave packet coefficient with a larger magnitude gives a better instantaneous frequency estimate with a higher probability. A highly redundant wave packet transform with a denser translation grid in space and scaling grid in frequency would have wave packets better fitting local oscillation of IMTs. In another word, there would be more coefficients with larger magnitudes. The resulting synchrosqueezed energy distribution has higher non-zero energy concentrating around instantaneous frequencies. This is also validated in Figure 4.10. The synchrosqueezed energy distributions in the third row are obtained by a SSWPT with a 16 times denser grid in frequency than the grid used in the second row. Hence, instantaneous frequencies are much more visible if a SSWPT with a higher redundancy is applied.

It is also interesting to observe that the synchrosqueezed transform with a smaller s is better to capture the component boundaries, e.g. at $x = 0.39$, 0.59 and 0.77 and is more robust to an impulsive perturbation (see Figure 4.9 and 4.10 at $x = 0.2$ and an example of α stable noise in Figure 4.11 and 4.12). Boundaries and impulse perturbation would produce frequency aliasing. The SSWPT with a smaller s has wave packets with a smaller support in frequency and a larger support in space. Hence, it is more robust to frequency aliasing in the sense that the influence of impulsive perturbation is smoothed out and the synchrosqueezed energy of the target components might not get dispersed when it meets the frequency aliasing, as shown in Figure 4.12.

However, if s is small, the instantaneous frequency estimate might be smoothed out and it is difficult to observe detailed information of instantaneous frequencies. As shown in the first row of Figure 4.10, when the input signal is noiseless, the synchrosqueezed transforms with $s = 0.75$ and 0.875 have better accuracy than the one with $s = 0.625$. In short, it is important to have tunable scaling parameters to design problem dependent synchrosqueezed transforms, which has been implemented in the SynLab toolbox.

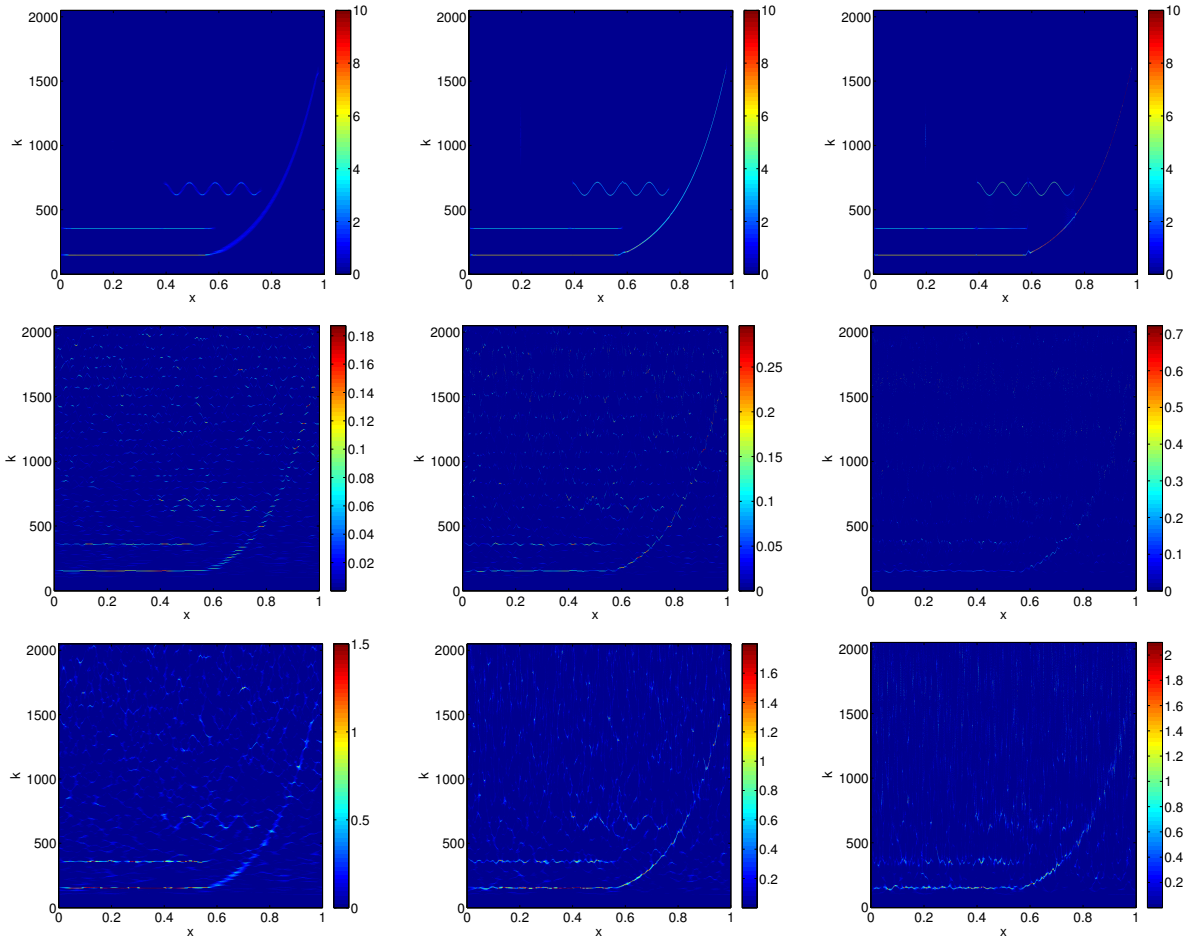


Figure 4.10: Synchrosqueezed energy distributions with $s = 0.625$ (left column), $s = 0.75$ (middle column) and $s = 0.875$ (right column). In the first row, we apply the SSWPT to clean data. In the second row, the SSWPT with a smaller redundancy is applied to the noisy data with $0.75\mathcal{N}(0, 1)$ noise in Figure 4.9. In the last row, a highly redundant SSWPT is applied to the same noisy data.

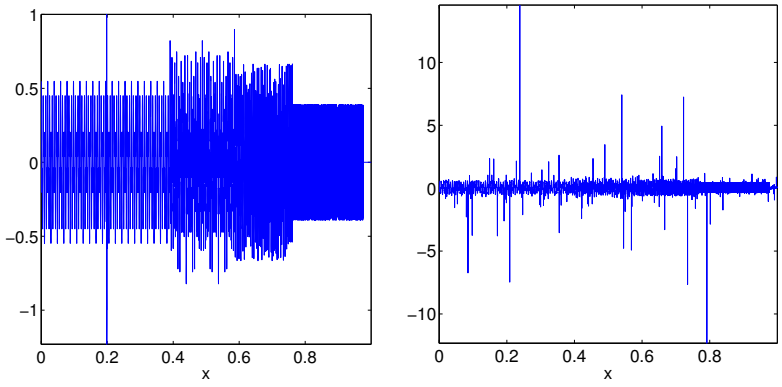


Figure 4.11: Left: A 1D synthetic benchmark signal. Right: A signal contaminated by an α stable random noise [1] with parameters $\alpha = 1$, dispersion= 0.9, $\delta = 1$, $N = 8192$. The noise is rescaled to have a 15 L^∞-norm by dividing a constant factor.

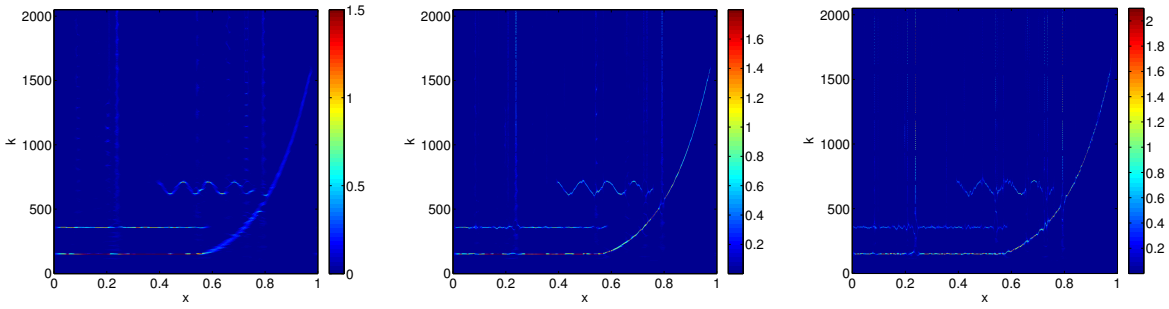


Figure 4.12: Synchrosqueezed energy distributions with $s = 0.625$ (left), $s = 0.75$ (middle) and $s = 0.875$ (right) using highly redundant SSWPTs. The synchrosqueezed energy with a smaller s is smoother and the influence of impulsive noise is weaker.

4.3.2 Robustness Tests for 2D SST

We now explore the performance of the 2D SSWPT using a single IMT in Figure 4.13. The function

$$f(x) = e^{2\pi i(60(x_1+0.05\sin(2\pi x_1))+60(x_2+0.05\sin(2\pi x_2)))} \tag{4.14}$$

is uniformly sampled in $[0,1]^2$ with a sampling rate 512 Hz and is disturbed by additive Gaussian white noise $5\mathcal{N}(0,1)$. The 2D SSWPTs with $s = 0.625$, 0.75 and 0.875 are applied to this noisy example and their results are shown in Figure 4.14. Since the synchrosqueezed energy distribution $T_f(x_1, x_2, k_1, k_2)$ of an image is a function in $\mathbb{R}^4$, we fix $x_2 = 0$, stack the results in k_2, and visualize $\int_{\mathbb{R}} T_f(x_1, 0, k_1, k_2)dk_2$.

The results in Figure 4.14 again validate the theoretical conclusion in Theorem 3.3.2 that a

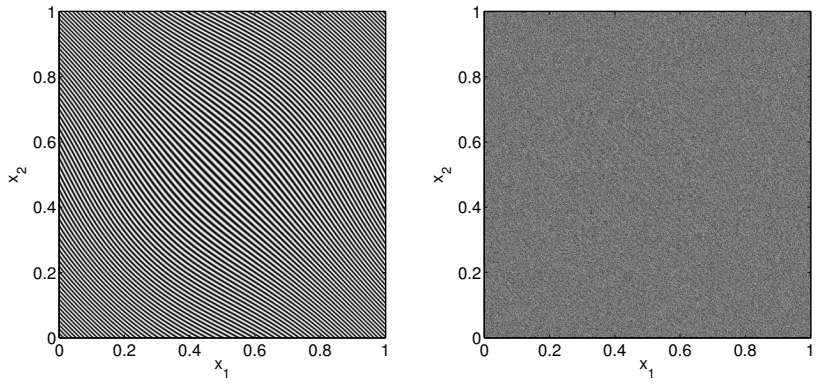


Figure 4.13: Left: A 2D noiseless IMT. Right: A noisy IMT generated with Gaussian white noise $5\mathcal{N}(0,1)$.

smaller scaling parameter s and a higher redundancy yield to a better robustness. A new idea here to achieve a better robustness is to design adaptive synchrosqueezed transforms tracing the possible frequency band of IMTs. A band-limited synchrosqueezed transform is designed in [180] for an efficient tool to analyze atomic crystal images. Numerical experiments will show that this method is strongly robust to noise. This inspires the idea of adaptive synchrosqueezed transforms above. We will do a simple experiment to justify this idea. In this experiment, we apply the band-limited SSWPT to the same 2D noisy image and present the results in the last row of Figure 4.14. Comparing to the results in the second row of Figure 4.14, the band-limited SSWPT clearly outperforms the original SSWPT.

4.3.3 Component Test

We will present the last example to validate the results of those theorems in the robustness analysis. Suppose we look at a region in the time-frequency or phase space domain and we know there might be only one IMT in this region. This assumption is reasonable because, after the synchrosqueezed transform, one might be interested in the synchrosqueezed energy in a particular region: is this corresponding to a component or just heavy noise? A straightforward solution is that, at each time or space grid point, we only reassign those coefficients with the largest magnitude. By Theorem 3.3.2, if there is an IMT, we can obtain a sketch of its instantaneous frequency or local wave vector with a high probability. If there is only noise, we would obtain random reassigned energy with a high probability. Using this idea, we apply the band-limited SSWPT with $s = 0.625$ and 10 times redundancy to a noisy version of the image in Figure 4.13 left. From left to right, Figure 4.15 shows the results of a noisy image (4.14) with $5\mathcal{N}(0,1)$ noise, a noisy image (4.14) with $10\mathcal{N}(0,1)$ noise, and an image with only noise, respectively. A reliable sketch of the local wave vector is still visible

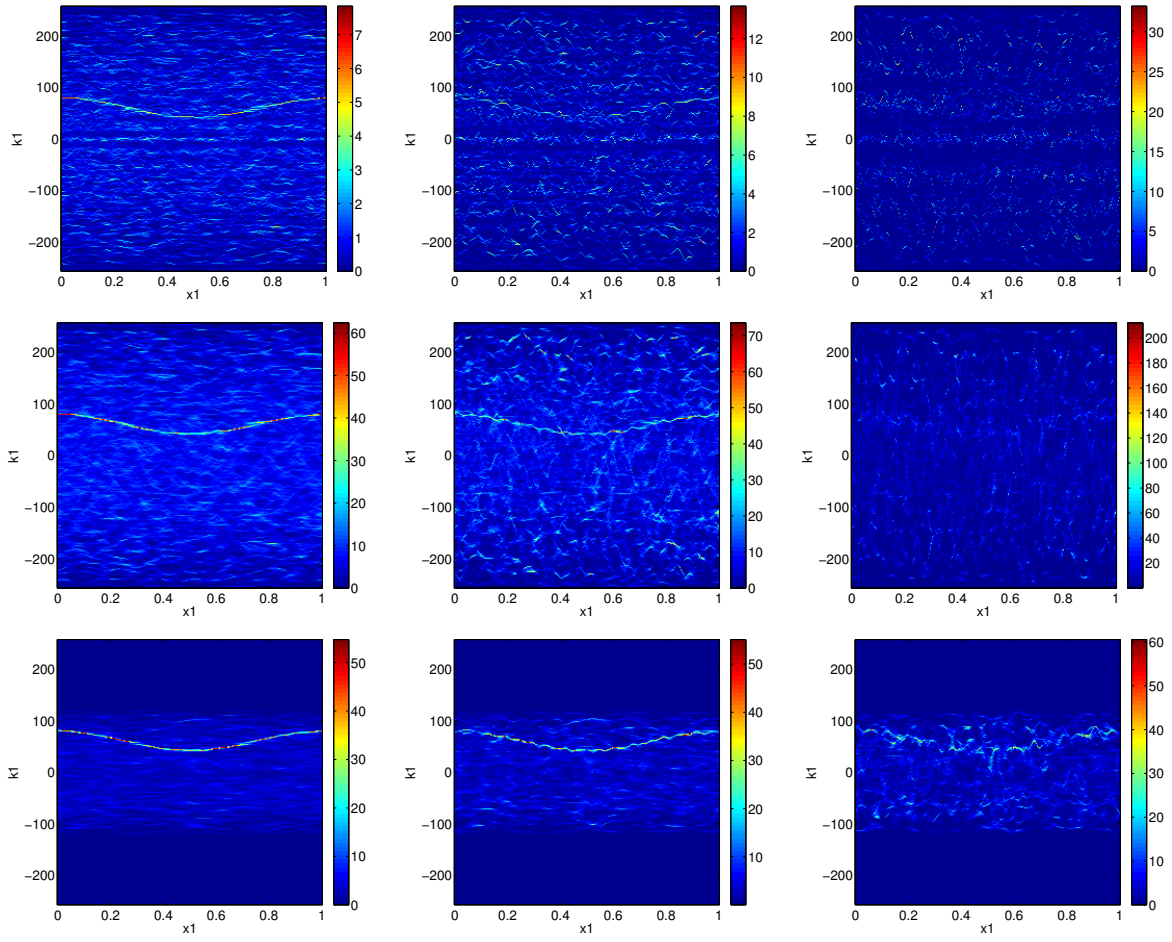


Figure 4.14: Stacked synchrosqueezed energy distribution $\int_{\mathbb{R}} T_f(x_1, 0, k_1, k_2)dk_2$ of the noisy 2D signal in Figure 4.13. From left to right, $s = 0.625$, 0.75 and 0.875. From top to bottom: standard redundancy, 10 times redundancy and 10 times redundancy with a SSWPT restricted to a frequency band from 20 to 120 Hz.

even if the input image is highly noisy.

4.3.4 Real Examples

In the last part of this section, we introduce a newly developed atomic crystal image analysis method based on synchrosqueezed transforms [180] to demonstrate the robustness of synchrosqueezed transforms in real applications. We will further study this application later in Chapter 6. In materials science, the information hidden in an atomic crystal image, e.g., grain boundaries, isolated defects, deformation field of each grain, is important for better understanding the properties of materials. As

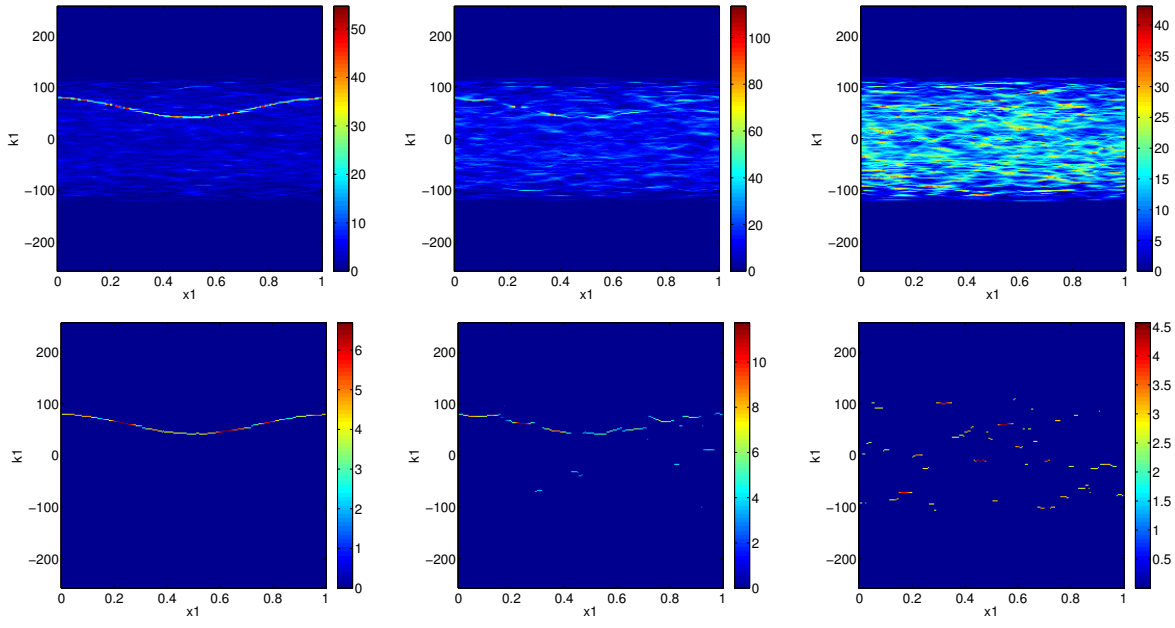


Figure 4.15: Top row: The synchrosqueezed energy distribution of the highly redundant band-limited SSWPT with a frequency band 20 to 120 Hz. Bottom row: reassigned wave packet coefficients with the largest magnitude at a space location. Left column: $5\mathcal{N}(0,1)$ noise. Middle column: $10\mathcal{N}(0,1)$ noise. Right column: noise only.

seen in Figure 4.16 (a), an atomic crystal image can be considered as an assemble of several general IMTs [180], where a general IMT is a superposition of a few IMTs with similar local wave vectors, e.g., with local wave vectors $\{(n_k\partial_{b_1}\phi(b), m_k\partial_{b_2}\phi(b))\}$ for some $\phi(b)$ and a few pairs (n_k, m_k). The method in [180] automatically determines a frequency band of the input image and applies a band-limited SSWPT to estimate the synchrosqueezed energy of each IMT. The location of the essential synchrosqueezed energy reveals grain boundaries, isolated defects and deformation fields (denoted by $\nabla F(x_1, x_2) \in \mathbb{R}^{2\times 2}$). Integrating $\nabla F(x_1, x_2)$ around a defect region can estimate the Burgers vector corresponding to the defect region. The distortion volume of $\nabla F(x_1, x_2)$, i.e., $\det(\nabla F(x_1, x_2)) - 1$ can reflect the strain stress on the grains (e.g. see Figure 4.16 (c)).

We apply the method in [180] to a phase field crystal image (Figure 4.16 (a)) and show the detected grain boundaries and isolated defects in Figure 4.16 (b), and the distortion volume in Figure 4.16 (c). To demonstrate the robustness, we generate additive Gaussian white noise with a distribution $0.5\mathcal{N}(0,1)$ and $1.4\mathcal{N}(0,1)$, respectively and present the noisy results in the second and the third rows of Figure 4.16. In the results of extremely noisy cases, even if no crystal structure visible by human eyes, the SSWPT method is still able to reveal grain boundaries and isolated defects with a reasonable accuracy. The distortion volume in Figure 4.16 (f) and (i) still roughly

reflects the strain stress encoded by color.

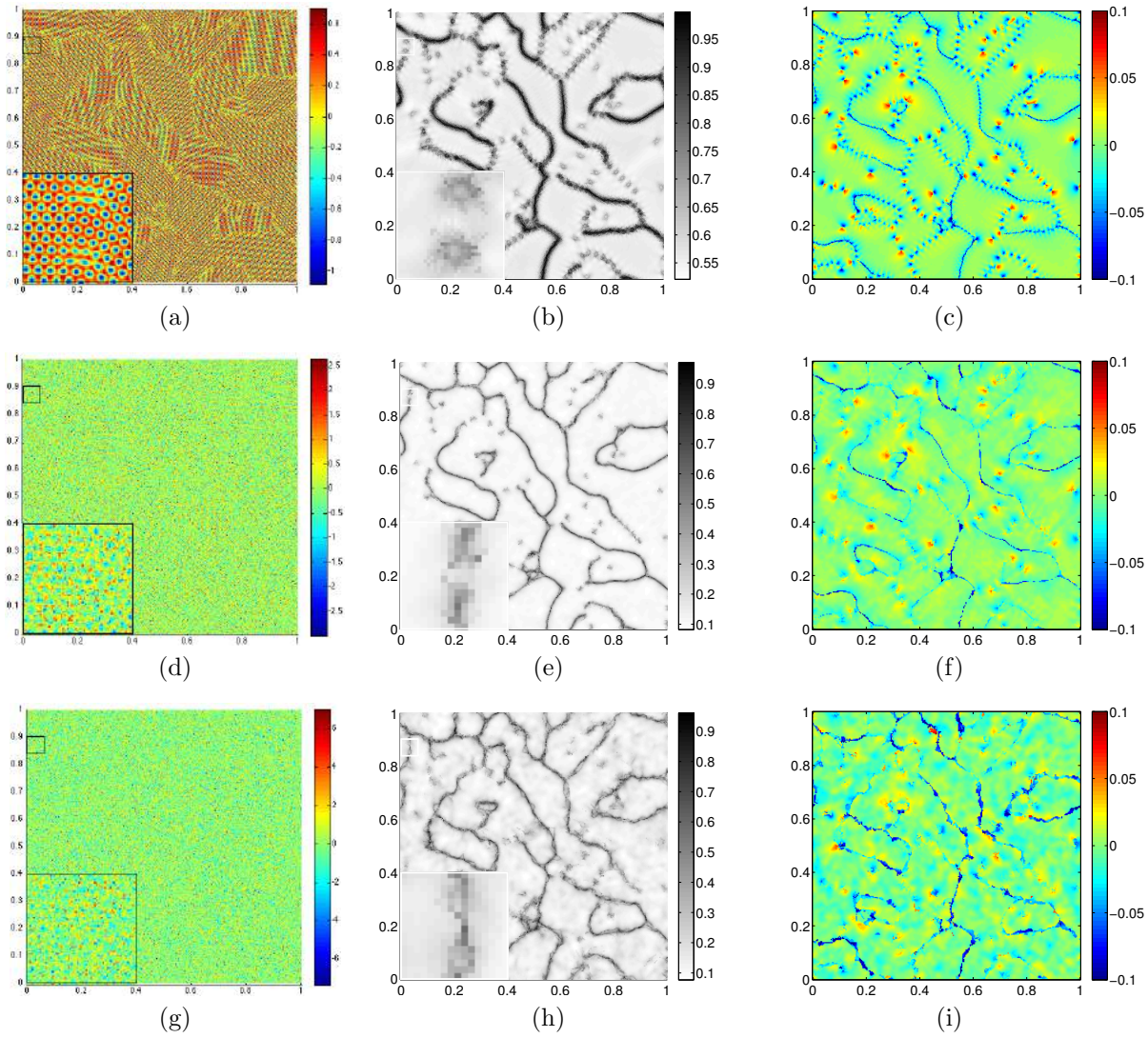


Figure 4.16: Atomic crystal image analysis using 2D synchrosqueezed transforms. First row: Results of noiseless data. Second row: Results of noisy data with Gaussian white noise $0.5\mathcal{N}(0,1)$. Third row: Results of noisy data with Gaussian white noise $1.4\mathcal{N}(0,1)$. First column: input images. Second column: detected grain boundaries and isolated defects. Third column: distortion volume. Zoomed-in images show that our method can still identify isolated defects even if noise is heavy.

Chapter 5

Diffeomorphism Based Spectral Analysis

5.1 Introduction

In Section 2.1, we have introduced the 1D synchrosqueezed transform to analyze the instantaneous properties of a complex signal of the form

$$f(x) = \sum_{k=1}^{K} f_k(x) = \sum_{k=1}^{K} \alpha_k(x) e^{2\pi i N_k \phi_k(x)}, \tag{5.1}$$

where $\alpha_k(x)$ is the instantaneous amplitude, $2\pi N_k \phi_k(x)$ is the instantaneous phase and $N_k \phi_k'(x)$ is the instantaneous frequency. One wishes to decompose the signal $f(x)$ to obtain each component $f_k(x)$ and its corresponding instantaneous properties. This is referred to as the mode decomposition problem.

In spite of considerable successes of analyzing signals by decomposing them in the form of (5.1), a superposition of a few wave-like components belongs to a very limited class of oscillatory patterns. Most of all, a decomposition in the form of (5.1) would lose important physical information in some cases as detailed in [170, 176]. To be more concrete, we take the daily atmospheric CO_2 concentration data in [176] as an example (provided by National Oceanic and Atmospheric Administration at Mauna Loa (MLO)). The method based on wavelet transforms is capable of decomposing data in the form of (5.1), providing one annual cycle, one semiannual cycle and a growing trend (see Figure 5.1). However, each component alone cannot reflect the true nonlinear evolution pattern: the CO_2 concentration slowly increased in a longer period and quickly decreased in a shorter period. This special pattern is a result of seasonal photosynthetic drawdown and respiratory release of CO_2 by terrestrial ecosystems [176]. Fortunately, such a nonlinear evolution pattern can be recovered by

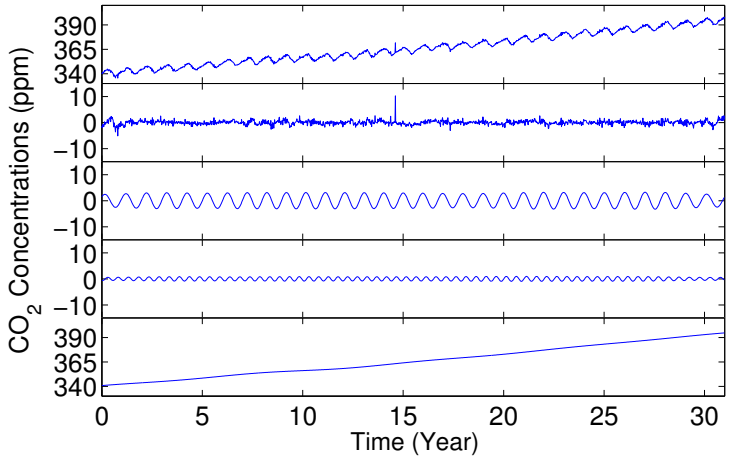


Figure 5.1: The top signal is the observed CO_2 concentration of recent 31 years (1981-2011) at MLO. Below the original signal are the components provided by the wavelet transform. Only relevant components are separated and presented.

summing up the annual cycle and the semiannual cycle as shown in Figure 5.2. This motivates the study of a more general decomposition of the form

$$f(x) = \sum_{k=1}^{K} f_k(x) = \sum_{k=1}^{K} \alpha_k(x) s_k(2\pi N_k \phi_k(x)), \tag{5.2}$$

where $\{s_k(x)\}_{1\le k\le K}$ are 2π-periodic general shape functions. By applying the Fourier expansion of general shape functions, the form of (5.2) is informally similar to the form of (5.1) with a superposition of infinite terms, i.e.,

$$f(x) = \sum_{k=1}^{K} \alpha_k(x) s_k(2\pi N_k \phi_k(x)) = \sum_{k=1}^{K} \sum_{n=-\infty}^{\infty} \widehat{s_k}(n) \alpha_k(x) e^{2\pi i n N_k \phi_k(x)}. \tag{5.3}$$

One could combine terms with similar oscillatory patterns in the form of (5.1) to obtain a more efficient and more meaningful decomposition in the form of (5.2). This is the general mode decomposition problem discussed in this chapter.

Although there have been well-established methods for mode decompositions, there is relative little literature for solving general mode decomposition problems due to the complex time-frequency geometry of (5.2). In the analysis of existing methods [43, 90, 170], they require a certain well-separation condition of $\widehat{s_k}(n)\alpha_k(x)e^{2\pi i n N_k \phi_k(x)}$ (e.g., see Definition 2.1.7). However, the superposition of two nearby Fourier expansion terms $\widehat{s_k}(n)\alpha_k(x)e^{2\pi i n N_k \phi_k(x)}$ and $\widehat{s_k}(n+1)\alpha_k(x)e^{2\pi i (n+1) N_k \phi_k(x)}$ are not well separated when n is large. For two different instantaneous frequencies $N_k\phi_k'(x)$ and

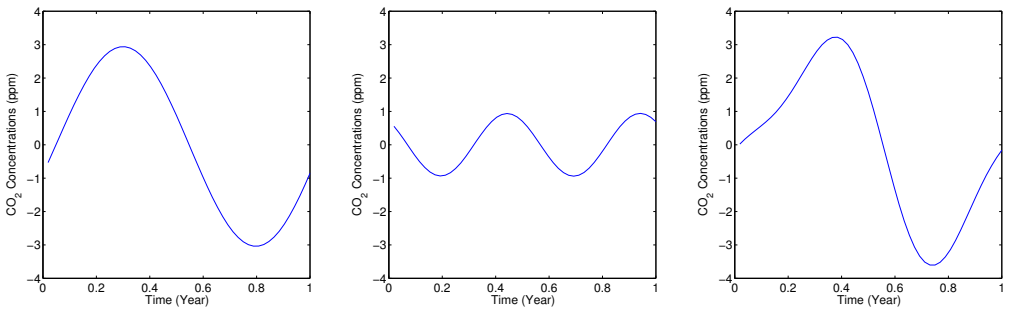


Figure 5.2: Wave shapes of relevant components provided by wavelet transform. Left: Annual wave shape. Middle: Semiannual wave shape. Right: Summation of the annual wave shape and semiannual wave shape.

$N_j\phi_j'(x)$, their multiples may have crossover frequencies with high probability. Although [89] shows that two different components $\widehat{s_k}(n)\alpha_k(x)e^{2\pi i n N_k\phi_k(x)}$ and $\widehat{s_j}(m)\alpha_j(x)e^{2\pi i m N_j\phi_j(x)}$ can be separated if their instantaneous frequencies $nN_k\phi_k'(x)$ and $mN_j\phi_j'(x)$ intersects at only a few points, in general, there is no existing method for the general mode decomposition with many instantaneous frequencies intersecting at many points.

This chapter introduces the diffeomorphism based spectral analysis method (DSA) in [178] as the first attempt to tackle the general mode decomposition problem with complex time-frequency geometry. The DSA method consists of diffeomorphisms and a short-time Fourier transform (in practice, the Fourier transform is applied if $f(x)$ is defined only in a bounded interval). Note that the wave-like components $\widehat{s_k}(n)\alpha_k(x)e^{2\pi i n N_k\phi_k(x)}$ with small n are relatively well separated in the sense that they would only intersect at a few points. Hence, we assume that the basic instantaneous frequencies $N_k\phi_k'(x)$ and instantaneous amplitudes $|\widehat{s_k}(1)|\alpha_k(x)$ can be estimated by existing methods. With this information available, it is shown that the DSA method is capable of decomposing a wide class of general superpositions accurately.

5.2 Diffeomorphism Based Spectral Analysis (DSA)

5.2.1 Implementation of the DSA

As discussed above, we assume that the basic instantaneous frequencies $N_k\phi_k'(x)$ and instantaneous amplitudes $|\widehat{s_k}(1)|\alpha_k(x)$ are known in this section. In practice, they are estimated by existing mode decomposition methods, e.g. the synchrosqueezed wave packet transform (SSWPT). Detailed description of searching for this basic information can be found in [178]. In what follows, the DSA

is introduced to identify all the nonzero Fourier expansion terms in

$$f(x) = \sum_{k=1}^{K} \alpha_k(x) s_k(2\pi N_k \phi_k(x)) = \sum_{k=1}^{K} \sum_{n=-\infty}^{\infty} \widehat{s_k}(n) \alpha_k(x) e^{2\pi i n N_k \phi_k(x)}, \tag{5.4}$$

assuming $\{N_k \phi_k'(x)\}_{k=1}^K$ and $\{|\widehat{s_k}(1)|\alpha_k(x)\}_{k=1}^K$ are known.

Without loss of generality, let us assume we are analyzing a signal $f(x)$ for $x \in [0, T_0]$ with $T_0 > 0$ sufficiently large and $f(x)$ is periodic over this interval. For non-periodic signals, introducing mirror extended signals can reduce the boundary effect. Notice that the smooth function $\phi_k(x)$ has the interpretations of a warping in each general mode via a diffeomorphism $\phi_k : \mathbb{R} \to \mathbb{R}$. With the instantaneous frequencies $\{N_k \phi_k'(x)\}_{k=1}^K$ available, we can therefore define the instantaneous phase profile by

$$p_k(x) = \frac{1}{m_k} \int_0^t N_k \phi_k'(x) dx,$$

where $m_k = \frac{1}{2}\left(\max_t N_k \phi_k'(x) + \min_t N_k \phi_k'(x)\right)$. Because $p_k(x)$ is a smooth monotonous function, we can define the inverse-warping profile in $[0, 1]$ by

$$\begin{aligned} h_k(x) &= \frac{f \circ p_k^{-1}(x)}{|\widehat{s_k}(1)| \alpha_k \circ p_k^{-1}(x)} \\ &= \sum_{n=-\infty}^{\infty} \frac{\widehat{s_k}(n)}{|\widehat{s_k}(1)|} e^{2\pi i (n m_k t + n N_k \phi_k(0))} \\ &\quad + \sum_{j \neq k} \sum_{n=-\infty}^{\infty} \frac{\widehat{s_j}(n)}{|\widehat{s_k}(1)|} \frac{\alpha_j \circ p_k^{-1}(x)}{\alpha_k \circ p_k^{-1}(x)} e^{2\pi i n N_j \phi_j \circ p_k^{-1}(x)}. \end{aligned}$$

If the diffeomorphisms $\phi_k : \mathbb{R} \to \mathbb{R}$ are significantly different, which will be defined later in Definition 5.2.4, and the phases $2\pi N_k \phi_k(x)$ are sufficiently steep in $[0, T_0]$, which will be clarified later, the Fourier transform of each inverse-warping profile $\widehat{h_k}(\xi)$ will have sheer peaks at $\xi = n m_k$ and will be relative small elsewhere. This motivates the design of the DSA method as follows.

Step 1: Input: A signal $f(x)$, its instantaneous phase profiles $\{p_k(x)\}_{k=1}^K$ and instantaneous amplitudes $\{|\widehat{s_k}(1)|\alpha_k(x)\}_{k=1}^K$.

Step 2: Initialize: Set up the initial residual $r(x) = f(x)$ and the tolerance ϵ. Let $f_k(x) = 0$ be the initial guess of the kth general mode and denote $S_k = \emptyset$ as the initial guess of the spectrum information of the kth general shape function s_k for $k = 1, \ldots, K$.

Step 3: For $k = 1, \ldots, K$, compute the inverse-warping profiles in $[0, 1]$ by

$$h_k(x) = \frac{r \circ p_k^{-1}(x)}{|\widehat{s_k}(1)| \alpha_k \circ p_k^{-1}(x)}.$$

Step 4: Apply the Fourier transform on $h_k(x)$ in $[0, 1]$ to obtain $\widehat{h_k}(\xi)$ for $k = 1, \ldots, K$ and

solve the following optimization problem,

$$(\tau, j) = \arg\max_{(\xi,k)} |\widehat{h_k}(\xi)|. \tag{5.5}$$

Then $\tau \approx nm_j$ for some n such that $\widehat{s_j}(n) \neq 0$.

Step 5: Let $g(x) = e^{2\pi i \tau t}$. Warp the harmonic $g(x)$ with the jth instantaneous phase profile $p_j(x)$ and multiply the warped harmonic by the jth instantaneous amplitude $|\widehat{s_j}(1)|\alpha_j(x)$ to obtain

$$\begin{aligned} |\widehat{s_j}(1)|\alpha_j(x) g \circ p_j(x) &\approx |\widehat{s_j}(1)|\alpha_j(x) e^{2\pi i n m_j p_j(x)} \\ &= |\widehat{s_j}(1)| e^{-2\pi i n N_j \phi_j(0)} \alpha_j(x) e^{2\pi i n N_j \phi_j(x)}. \end{aligned}$$

Step 6: Solve the L^2 minimization problem for a complex factor $\beta \in \mathbb{C}$ such that

$$\beta = \arg\min_{\beta \in \mathbb{C}} \|r(x) - \beta |\widehat{s_j}(1)|\alpha_j(x) g \circ p_j(x)\|_{L^2}.$$

Then

$$\beta |\widehat{s_j}(1)|\alpha_j(x) g \circ p_j(x) \approx \widehat{s_j}(n)\alpha_j(x) e^{2\pi i n N_j \phi_j(x)},$$

which implies

$$|\beta| \approx \frac{|\widehat{s_j}(n)|}{|\widehat{s_j}(1)|}.$$

Step 7: Update: Compute the new residual

$$r(x) = r(x) - \beta |\widehat{s_j}(1)|\alpha_j(x) g \circ p_j(x).$$

Update the jth recovered general mode

$$f_j(x) = f_j(x) + \beta |\widehat{s_j}(1)|\alpha_j(x) g \circ p_j(x),$$

and the jth spectrum information set

$$S_j = S_j \cup \{(\tau, |\beta|)\}.$$

Step 8: If $\|r(x)\|_{L^2} > \epsilon$, repeat step 3-7. Otherwise, stop iterating and export the general mode estimates f_k and the spectrum information S_k for $k = 1, \ldots, K$.

Note that in general there is no guarantee to obtain a constant β such that we have exact equality in Step 6, since many components are overlapping in the Fourier domain. However, as long as the phase functions are significantly different, the interference of other components is small. Moreover, as long as Step 4 is accurate, the approximation of $\widehat{s_j}(n)\alpha_j(x) e^{2\pi i n N_j \phi_j(x)}$ can be retrieved in later

iteration process, since the error

$$\widehat{s_j}(n)\alpha_j(x)e^{2\pi i n N_j\phi_j(x)} - \beta|\widehat{s_j}(1)|\alpha_j(x)g\circ p_j(x)$$

is a new wave-like component with the same phase function $nN_j\phi_j(x)$ and smaller spectral energy to be recovered.

5.2.2 Analysis of the DSA

As discussed above, the key step of the DSA is the estimation in (5.5). Theorem 5.2.5 in this section proves that (5.5) can provide precise spectral analysis, if phase functions are significantly different and steep enough. We consider the following short-time Fourier transform with real-valued, non-negative and smooth window function $w_1(x)$ compactly supported in $(-1,1)$ such that $|\widehat{w_1}|$ has a sheer peak around the origin and rapidly decays elsewhere.

Definition 5.2.1. *Given the window function $w_1(x)$ and a parameter $T>1$, the short-time Fourier transform of a function $f(x)$ with a parameter T is a function*

$$\mathcal{F}_T(f)(a,b) = \int_{\mathbb{R}} f(x)w_T(x-b)e^{-2\pi i a x}dx$$

for $a,b\in\mathbb{R}$, where $w_T(x)=w_1(x/T)$ and $\mathcal{F}_T$ denote the short-time Fourier transform operator with the parameter T.

Next, we introduce the model of wave-like components in the general mode decomposition.

Definition 5.2.2. *General shape functions:*

The general shape function class $\mathcal{S}_M$ consists of 2π-periodic functions $s(x)$ in the Wiener Algebra with a unit $L^2([-\pi,\pi])$-norm and a L^∞-norm bounded by M satisfying the following spectral conditions:

1. *The Fourier series of $s(x)$ is uniformly convergent;*
2. *$\sum_{n=-\infty}^{\infty}|\widehat{s}(n)|\le M$ and $\widehat{s}(0)=0$;*
3. *Let Λ be the set of integers $\{|n|:\widehat{s}(n)\neq 0\}$. The greatest common divisor $\gcd(s)$ of all the elements in Λ is 1.*

In fact, if $\gcd(s)>1$, then the general mode $s(2\pi N\phi(x))$ can be considered as a more oscillatory mode $\tilde{s}(2\pi\gcd(s)N\phi(x))$ with $\gcd(\tilde{s})=1$ and the Fourier coefficients $\widehat{\tilde{s}}(n)=\widehat{s}(\gcd(s)n)$. The requirement that $\widehat{s}(0)=0$ and s has a unite $L^2([-\pi,\pi])$-norm is to normalize the general shape function.

Definition 5.2.3. *A function $f(x) = \alpha(x)s(2\pi N\phi(x))$ is a general intrinsic mode type function (GIMT) of type (M, N), if $s(x) \in \mathcal{S}_M$ and $\alpha(x)$ and $\phi(x)$ satisfy the conditions below.*

$$\alpha(x) \in C^\infty, \quad |\alpha'| \le M, \quad 1/M \le \alpha \le M$$
$$\phi(x) \in C^\infty, \quad 1/M \le |\phi'| \le M, \quad |\phi''| \le M.$$

Definition 5.2.4. *For $M > 0$ and $K > 0$, the phase functions $\{\phi_k(x)\}_{1\le k\le K}$ are significantly different of type (M, K) at $b \in \mathbb{R}$, if they satisfy the following conditions.*

1. *For any $T > 0$, the number of extrema of $\phi_k \circ \phi_j^{-1}(x)$ in $(b - T, b + T)$ is at most TM for $k \ne j$.*
2. *For any $T > 0$ there exists $\eta_0 > 0$, $\eta_1 > 0$ and $N_0(M, K, T, b)$ such that $\forall a \in (\frac{1}{2M^2}, 2M^2)$ and $\forall N > N_0(M, K, T, b)$*

$$\lambda^*\left(\left\{x : \left|\partial_x\left(\phi_k \circ \phi_j^{-1}(x)\right) - a\right| \le \frac{1}{N^{1-\eta_0}}\right\} \cap \{x : b - T \le x \le b + T\}\right) \lesssim O(\frac{1}{N^{\eta_1}})$$

for $k \ne j$, where $\lambda^*(\cdot)$ denotes the Lebesgue measure and $\lesssim$ means the implicit constant may depend on M, K, T and b.

The first condition in Definition 5.2.4 assumes that the instantaneous frequencies are not oscillating fast, while the second condition requires that $\phi_k \circ \phi_j^{-1}(x)$ is far from a constant function. The definition of significantly different phase functions is crucial to general mode decompositions. The difference of phase functions is the key feature for grouping the Fourier expansion terms of the general modes. If two phase functions are similar, their corresponding general modes would have similar evolution patterns. It is reasonable to combine them as one general mode. On the other hand, the significant-difference of phase functions guarantees that the key idea of the DSA method can provide accurate spectral information of general shape functions, as proved in the following theorem.

Theorem 5.2.5. *Suppose $f(x) = \sum_{k=1}^K f_k(x)$, where $f_k(x) = \alpha_k(x)s_k(2\pi N_k\phi_k(x))$ is a GIMT of type (M, N_k) with $N_k \ge N$ and the phase functions $\{\phi_k(x)\}_{1\le k\le K}$ are significantly different of type (M, K) at b. Let $s_0 = \max_{(k,n)} |\widehat{s_k}(n)|$. Define*

$$h_k(x) = \frac{f \circ \phi_k^{-1}(x)}{\alpha_k \circ \phi_k^{-1}(x)}$$

for $1 \le k \le K$. For fixed M, K, b, s_0 and $\delta > 0$, $\exists T_0(M, K, s_0, \delta, b)$, $\forall T > T_0$, $\exists N_0(M, K, s_0, T, b) > 0$ such that $\forall N > N_0$ the solution of the following optimization problem

$$(a_0, k_0) = \arg\max_{(a,k)} |\mathcal{F}_T(h_k)(a, b)|$$

satisfies $|a_0 - nN_{k_0}| < \delta$ *for some* n *such that* $\widehat{s_{k_0}}(n) \neq 0$.

In what follows, when we write $O(\cdot)$, $\lesssim$, or $\gtrsim$, the implicit constants may depend on M, K, T and b.

Proof. Notice that

$$h_k(x) = \frac{f \circ \phi_k^{-1}(x)}{\alpha_k \circ \phi_k^{-1}(x)} = \sum_{n=-\infty}^{\infty} \widehat{s_k}(n) e^{2\pi i n N_k x} + \sum_{j \neq k} \sum_{n=-\infty}^{\infty} \widehat{s_j}(n) \frac{\alpha_j \circ \phi_k^{-1}(x)}{\alpha_k \circ \phi_k^{-1}(x)} e^{2\pi i n N_j \phi_j \circ \phi_k^{-1}(x)},$$

then

$$\begin{aligned} \mathcal{F}_T(h_k)(a,b) &= \sum_{n=-\infty}^{\infty} \widehat{s_k}(n) \int_{\mathbb{R}} w_T(x-b) e^{2\pi i (nN_k - a)x} dx \\ &\quad + \sum_{j \neq k} \sum_{n=-\infty}^{\infty} \widehat{s_j}(n) \int_{\mathbb{R}} \frac{\alpha_j \circ \phi_k^{-1}(x)}{\alpha_k \circ \phi_k^{-1}(x)} w_T(x-b) e^{2\pi i (n N_j \phi_j \circ \phi_k^{-1}(x) - ax)} dx \end{aligned}$$

by the uniform convergence of the Fourier series of $s_k(x)$. The first part of $\mathcal{F}_T(h_k)(a,b)$ is

$$\begin{aligned} I_1(a,k) &= \sum_{n=-\infty}^{\infty} \widehat{s_k}(n) \int_{\mathbb{R}} w_T(x-b) e^{2\pi i (nN_k - a)x} dx \\ &= \sum_{n=-\infty}^{\infty} T \widehat{s_k}(n) e^{2\pi i b (nN_k - a)} \int_{\mathbb{R}} w_1(x) e^{2\pi i T (nN_k - a)x} dx \\ &= \sum_{n=-\infty}^{\infty} T \widehat{s_k}(n) e^{2\pi i b (nN_k - a)} \widehat{w_1}\left(T(a - nN_k)\right). \end{aligned}$$

Hence, $\exists T_0(M, K, s_0, \delta, b)$ such that, if $T > T_0$, then $|I_1(a,k)|$ has well-separated sheer energy peaks at $a = nN_k$ of order $T\left|\widehat{s_k(n)}\right|$ and $|I_1(a,k)| < \frac{Ts_0}{3}$ if $|a - nN_k| \geq \delta$ for all n. The estimate of the second part

$$I_2(a,k) = \sum_{j \neq k} \sum_{n=-\infty}^{\infty} \widehat{s_j}(n) \int_{\mathbb{R}} \frac{\alpha_j \circ \phi_k^{-1}(x)}{\alpha_k \circ \phi_k^{-1}(x)} w_T(x-b) e^{2\pi i (n N_j \phi_j \circ \phi_k^{-1}(x) - ax)} dx$$

relies on the estimate of each term

$$I_{jn} = \widehat{s_j}(n) \int_{\mathbb{R}} \frac{\alpha_j \circ \phi_k^{-1}(x)}{\alpha_k \circ \phi_k^{-1}(x)} w_T(x-b) e^{2\pi i (n N_j \phi_j \circ \phi_k^{-1}(x) - ax)} dx.$$

Notice that $\frac{\alpha_j \circ \phi_k^{-1}(x)}{\alpha_k \circ \phi_k^{-1}(x)} w_T(x-b)$ and $2\pi(nN_j \phi_j \circ \phi_k^{-1}(x) - ax)$ are real smooth functions and $w_T(x-b)$ has a compact support in $(b-T, b+T)$. If $\partial_x(nN_j \phi_j \circ \phi_k^{-1}(x) - ax) \neq 0$ in $(b-T, b+T)$, a similar

argument of the integration by parts in Lemma 2.1.9 shows that

$$|I_{jn}| \lesssim |\widehat{s_j}(n)| \frac{1}{\left|nN_j \partial_x(\phi_j \circ \phi_k^{-1})(x) - a\right|}.$$

Therefore, the order of $|I_{jn}|$ is determined by points x such that $\left|nN_j \partial_x(\phi_j \circ \phi_k^{-1})(x) - a\right|$ is vanishing or relatively small.

If $a \notin (\frac{nN_j}{2M^2}, 2nN_jM^2)$, then by the fact that $\partial_x(\phi_j \circ \phi_k^{-1})(x) \in [\frac{1}{M^2}, M^2]$, we have

$$\left|nN_j \partial_x(\phi_j \circ \phi_k^{-1})(x) - a\right| \gtrsim nN_j$$

, which implies

$$|I_{jn}| \lesssim \frac{|\widehat{s_j}(n)|}{nN_j} \lesssim \frac{1}{N}. \tag{5.6}$$

If $a \in (\frac{nN_j}{2M^2}, 2nN_jM^2)$, then $\frac{a}{nN_j} \in (\frac{1}{2M^2}, 2M^2)$. Let

$$A = \left\{ x : \left| \partial_x \left(\phi_j \circ \phi_k^{-1}(x) \right) - \frac{a}{nN_j} \right| \leq \frac{1}{(nN_j)^{1-\eta_0}} \right\} \cap \{ x : b - T \leq x \leq b + T \}.$$

Because the phase functions are significantly different of type (M, K) at b, for fixed T there exists $\eta_0 > 0$, $\eta_1 > 0$ and $N_1(M, K, T, b)$ such that for $\frac{a}{nN_j} \in (\frac{1}{2M^2}, 2M^2)$ and $nN_j > N_1(M, K, T, b)$, we have $\lambda^*(A) \lesssim O(\frac{1}{(nN_j)^{\eta_1}})$. This gives

$$\left| \widehat{s_j}(n) \int_A \frac{\alpha_j \circ \phi_k^{-1}(x)}{\alpha_k \circ \phi_k^{-1}(x)} w_T(x - b) e^{2\pi i (nN_j \phi_j \circ \phi_k^{-1}(x) - ax)} dx \right| \lesssim O(\frac{|\widehat{s_j}(n)|}{(nN_j)^{\eta_1}}).$$

By the definition of significant-difference of type (M, K), $(\mathbb{R} \setminus A) \cap (b - T, b + T)$ is a union of at most $O(TM)$ intervals. Hence, similar to the method of stationary phase, we have

$$\left| \widehat{s_j}(n) \int_{\mathbb{R} \setminus A} \frac{\alpha_j \circ \phi_k^{-1}(x)}{\alpha_k \circ \phi_k^{-1}(x)} w_T(x - b) e^{2\pi i (nN_j \phi_j \circ \phi_k^{-1}(x) - ax)} dx \right| \lesssim O(\frac{|\widehat{s_j}(n)|}{(nN_j)^{\eta_0}}).$$

In sum,

$$\begin{aligned} |I_{jn}| &\leq \left| \widehat{s_j}(n) \int_{\mathbb{R} \setminus A} \frac{\alpha_j \circ \phi_k^{-1}(x)}{\alpha_k \circ \phi_k^{-1}(x)} w_T(x - b) e^{2\pi i (nN_j \phi_j \circ \phi_k^{-1}(x) - ax)} dx \right| \\ &\quad + \left| \widehat{s_j}(n) \int_A \frac{\alpha_j \circ \phi_k^{-1}(x)}{\alpha_k \circ \phi_k^{-1}(x)} w_T(x - b) e^{2\pi i (nN_j \phi_j \circ \phi_k^{-1}(x) - ax)} dx \right| \\ &\lesssim O(\frac{|\widehat{s_j}(n)|}{(nN_j)^{\eta_1}}) + O(\frac{|\widehat{s_j}(n)|}{(nN_j)^{\eta_0}}). \end{aligned}$$

Recall that $N_k \geq N$ and $\sum_{n=-\infty}^{\infty} |\widehat{s_k}(n)| \leq M$ for $1 \leq k \leq K$. So, if $N > N_1(M, K, T, b)$

$$|I_2(a,k)| \lesssim \sum_{j \neq k} \sum_{n=-\infty}^{\infty} \left(O(\frac{|\widehat{s_j}(n)|}{(nN_j)^{\eta_1}}) + O(\frac{|\widehat{s_j}(n)|}{(nN_j)^{\eta_0}}) \right) \lesssim O(\frac{(K-1)M}{N^{\eta}}) \lesssim O(\frac{1}{N^{\eta}}), \tag{5.7}$$

where $\eta = \min\{\eta_0, \eta_1\}$.

By (5.6) and (5.7), $\exists N_0 = \max\left\{ N_1(M,K,T,b), \left(\frac{3}{Ts_0}\right)^{1/\eta}, \frac{3}{Ts_0} \right\}$ such that $\forall N > N_0$, we have $|I_2(a,k)| < \frac{Ts_0}{3}$.

Let Ξ_k be the index set $\{n : \widehat{s_k}(n) \neq 0\}$ and $(\tilde{n}, \tilde{k}) = \arg\max_{(n,k)} |\widehat{s_k}(n)|$. Now suppose $N > N_0$. Let $|\mathcal{F}_T(h_k)(a,b)|$ take the maximum value at the pair (a_0, k_0). If there is no $n \in \Xi_{k_0}$ such that $|a_0 - nN_{k_0}| < \delta$, then

$$|\mathcal{F}_T(h_{k_0})(a_0,b)| \leq |I_1(a_0,k_0)| + |I_2(a_0,k_0)| < \frac{2Ts_0}{3}.$$

However, for the pair $(\tilde{n}, \tilde{k})$, we have

$$\left|\mathcal{F}_T(h_{\tilde{k}})(\tilde{n},b)\right| \geq \left|I_1(\tilde{n},\tilde{k})\right| - \left|I_2(\tilde{n},\tilde{k})\right| > Ts_0 - \frac{Ts_0}{3} > \frac{2Ts_0}{3}.$$

This conflicts with the fact that $|\mathcal{F}_T(h_k)(a,b)|$ takes the maximum value less than $\frac{2Ts_0}{3}$ at the pair (a_0, k_0). Hence, there exists $n \in \Xi_{k_0}$ satisfying that $|a_0 - nN_{k_0}| < \delta$. This completes the proof. □

In practice, the signal $f(x)$ is defined in a bounded interval, e.g., $[0, T_0]$ without loss of generality. Applying the Fourier transform on $f(x)$ in $[0, T_0]$ is equivalent to applying the short-time Fourier transform on $f(x)$ with a rectangle window function centered at $t = T_0/2$. In this sense, Theorem 5.2.5 implies that the DSA method can accurately decompose $f(x)$ into GIMTs $\{\alpha_k(x) s_k(2\pi N_k \phi_k(x))\}_{k=1}^K$ and analyzes the spectra of general shape functions $\{\alpha_k(x)\}_{k=1}^K$ by extracting the Fourier expansion terms $\widehat{s_k}(n)\alpha_k(x)e^{2\pi i n N_k \phi_k(x)}$ one by one from the one with highest energy.

5.3 Numerical Examples

In this section, some numerical examples of synthetic and real data are provided to demonstrate the properties of the proposed DSA method. In all of these examples, the 1D SSWPT is applied to provide basic instantaneous frequencies and instantaneous amplitudes as input of the DSA method. The mother wave packet $w(x)$ of the SSWPT is constructed using the same method in [49] with a support parameter $d = 1$. The scaling parameter s is equal to 2/3. For the purpose of convenience, the synthetic data is defined in $[0, 1]$ and the number of samples is between 2^{13} and 2^{15}.

5.3.1 Synthetic Examples

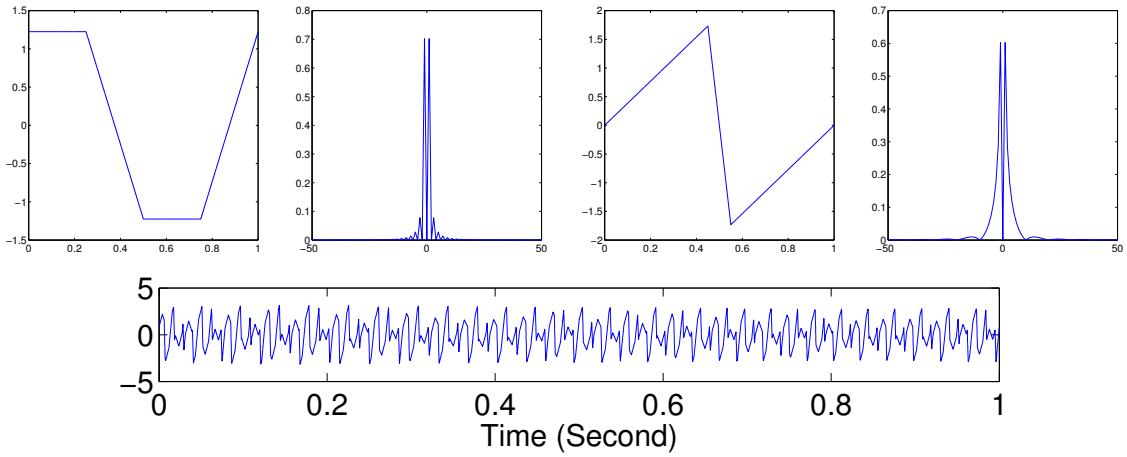


Figure 5.3: Top left: The general shape function $s_1(x)$ and its spectral energy $|\widehat{s_1}(\xi)|$. Top right: The general shape function $s_2(x)$ and its spectral energy $|\widehat{s_2}(\xi)|$. Bottom: A superposition of general modes generated by using $s_1(x)$ and $s_2(x)$.

Example 1: In the first example, we illustrate the performance of the SSWPT and the DSA step by step for general mode decompositions. Let us consider a toy model in which there are two general modes

$$f_1(x) = \alpha_1(x)s_1(2\pi N_1\phi_1(x)) = (1 + 0.05\sin(4\pi x))s_1\left(120\pi(x + 0.01\sin(2\pi x))\right)$$

and

$$f_2(x) = \alpha_2(x)s_2(2\pi N_2\phi_2(x)) = (1 + 0.1\sin(2\pi x))s_2\left(180\pi(x + 0.01\cos(2\pi x))\right),$$

where $s_1(x)$ and $s_2(x)$ are periodic general shape functions defined in $[0, 1]$ as shown in Figure 5.3. Let $f(x) = f_1(x) + f_2(x)$ (see Figure 5.3 bottom) and we try to recover $f_1(x)$ and $f_2(x)$ from $f(x)$.

As proved in Chapter 2 and Chapter 4, the SSWPT is able to provide a sharpened time-frequency representation of $f(x)$, the synchrosqueezed energy distribution $T_f(v, x)$, with essential supports concentrating around the instantaneous frequencies of $f(x)$ (see Figure 5.4 left). By a proper curve extraction and classification method in [178], we can identify well-separated instantaneous frequencies of $f_1(x)$ and $f_2(x)$ in the low frequency part (see Figure 5.4 middle) and their basic instantaneous frequencies $N_1\phi_1'(x)$ and $N_2\phi_2'(x)$ (see Figure 5.4 right). The inverse SSWPT on the synchrosqueezed energy distribution restricted to the each essential support recovers $\widehat{s_1}(n_1)e^{2\pi i n_1 N_1\phi_1(x)}$ and $\widehat{s_2}(n_2)e^{2\pi i n_2 N_2\phi_2(x)}$ for some n_1 and n_2. Hence, the instantaneous amplitudes are identified by taking the absolute value of them (see Figure 5.5 left).

As we can see in the this example, the SSWPT can provide accurate estimates of instantaneous

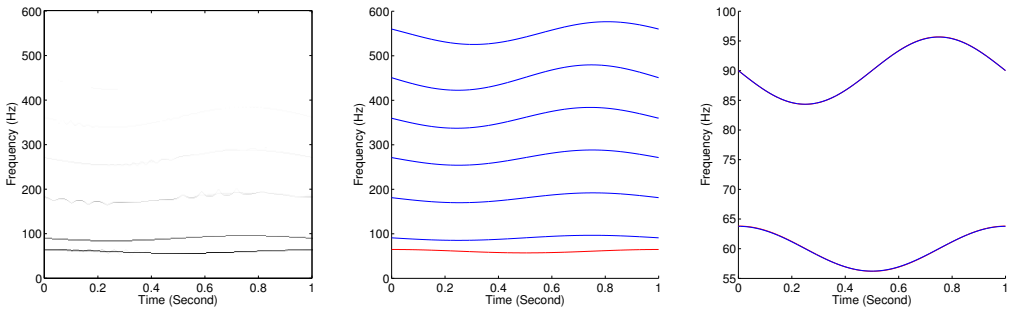


Figure 5.4: Left: The synchrosqueezed energy distribution of $f(x)$. Middle: The instantaneous frequency estimates and the result of curve classification as indicated by different colors. Right: The red curves are the estimates of basic instantaneous frequencies and the blue curves are the real basic instantaneous frequencies $N_k\phi_k'(b)$.

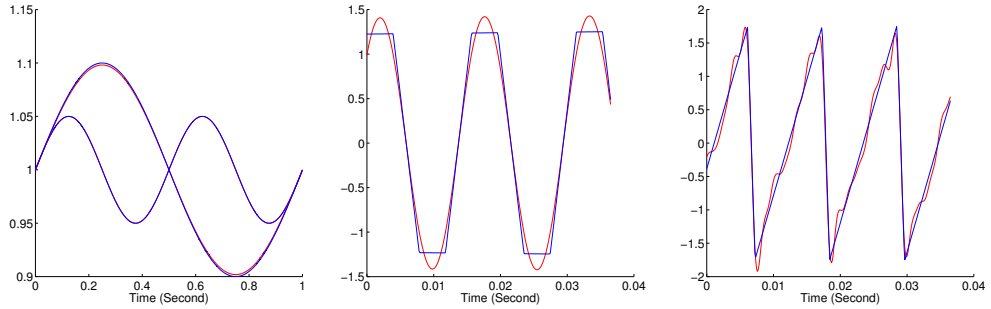


Figure 5.5: Blue: Real signals. Red: Reconstructed results. Left: Estimated normalized instantaneous amplitude and real normalized instantaneous amplitudes. Middle and right: The real general modes and the recovered general modes by simply summing up the identified components with well-separated instantaneous frequencies.

frequencies and instantaneous amplitudes from the well-separated essential supports of the synchrosqueezed energy distribution. However, by simply summing up the reconstructed modes cannot recover satisfactory $f_1(x)$ and $f_2(x)$ (see Figure 5.5 middle and right). As Figure 5.6 shows, considering only the well-separated essential supports of the synchrosqueezed energy distribution would ignore modes with weak energy and crossover frequencies, the information of which is indispensable to reconstruct exact general modes. This desires the DSA method for exact reconstructions of general modes.

With the basic instantaneous frequencies and instantaneous amplitudes provided by the SSWPT, the DSA is able to recover the general modes $f_1(x)$ and $f_2(x)$ as shown in Figure 5.7.

Example 2: In what follows, we would study the robustness against noise. The shapes of general modes are determined by all the Fourier expansion terms, including those weak energy terms that

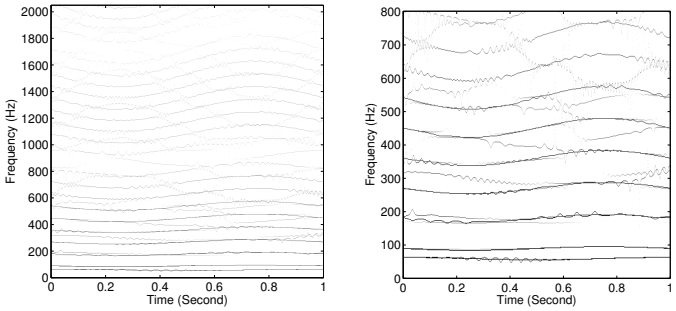


Figure 5.6: Left: $\log_{10}(T_f(v,x))$ in the visible time-frequency domain. Right: $\log_{10}(T_f(v,x))$ in the low frequency part of the time-frequency domain. Some components with weak energy are interfering other terms. Only a few components are well separated.

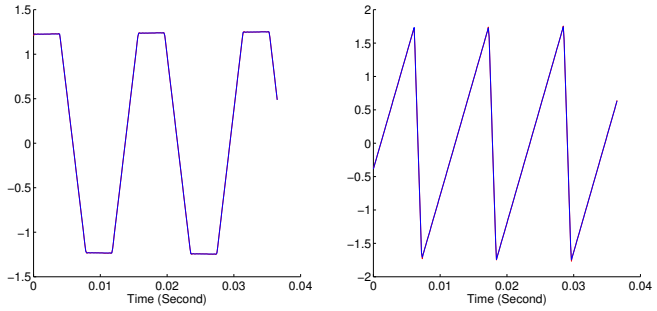


Figure 5.7: Blue: Real signals. Red: Reconstructed results. Two recovered general modes provided by the DSA method.

have been concealed by noise. The noise used here is a Gaussian random noise $n(x)$ with zero mean and variance σ^2. To quantify the influence of the noise on each general mode, we introduce the following Signal-to-Noise Ratio (SNR)

$$\text{SNR}[dB] = \min\left\{10\log_{10}\left(\frac{\|f_i\|_{L^2}}{\sigma^2}\right), 1 \le i \le K\right\},$$

where $\{f_i\}_{i=1}^K$ are the general modes contained in the original signal $f(x)$.

Let us revisit Example 1 in Figure 5.4 and study its noisy case,

$$f(x) = \alpha_1(x)s_1(2\pi N_1\phi_1(x)) + \alpha_2(x)s_2(2\pi N_2\phi_2(x)) + n(x).$$

Figure 5.8 shows two superpositions with different noise levels. As the reconstructed results show in Figure 5.9 and Figure 5.10 left, the instantaneous frequencies are accurately estimated, even if the signal is disturbed by severe noise. The essential feature of the general modes are recovered.

When the noise is overwhelming the general modes, additional denoising procedure is application dependent, as we will show in the next example.

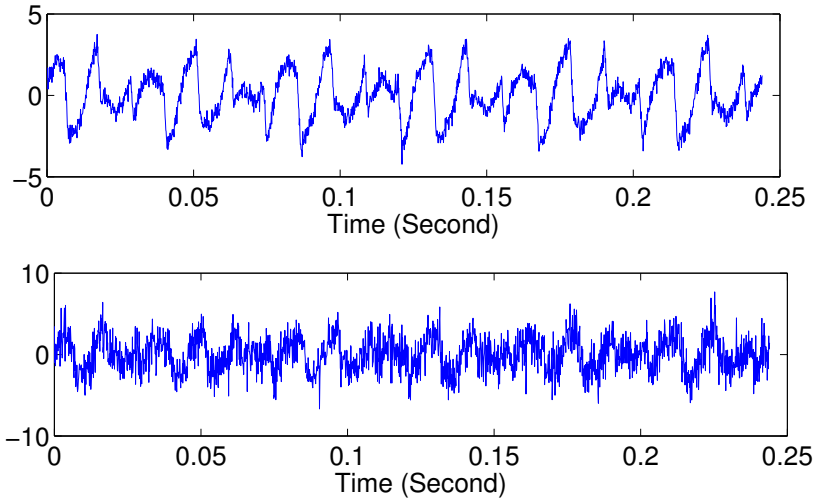


Figure 5.8: Noisy signals of Example 1 and their SNRs are 6 and -3, respectively.

As a comparison, the EEMD method in [174] is applied to the noisy data with $SNR = 6$ in this example. In the EEMD, we set up the ratio of the standard deviation of the added noise and that of the original data as 0.2. The ensemble number is 50 and the expected number of modes is 5. This method is able to provide several modes in different frequency scales shown in Figure 5.10 (right). However, they are not those modes expected.

Example 3: It is worth pointing out that suitable denoising according to the feature of recovered modes can significantly improve the results. Combining the DSA with some post processing techniques can detect general shape functions in a wider class than the one defined in Definition 5.2.2. For example, we test piecewise constant shape functions s_3 and s_4 as shown in Figure 5.11. A noisy superposition of general modes is generated as follows.

$$f(x) = \alpha_3(x)s_3(2\pi N_3\phi_3(x)) + \alpha_4(x)s_4(2\pi N_4\phi_4(x)) + n(x),$$

where $\alpha_3(x) = 1 + 0.4\sin(4\pi t)$, $\alpha_4(x) = 1 - 0.3\sin(2\pi t)$, $N_3 = 120$, $N_4 = 185$, $\phi_3(x) = t + 0.005\sin(2\pi t)$, and $\phi_4(x) = t + 0.01\cos(4\pi t)$. In this example, the SSWPT is applied to estimate the instantaneous information first and then the DSA method is applied to decompose $f(x)$ into two general modes. Finally, a TV norm minimization is applied to obtain the final results shown in Figure 5.11. The DSA method is able to detect the basic feature of these general modes and the post processing TV norm minimization helps to reduce the noise.

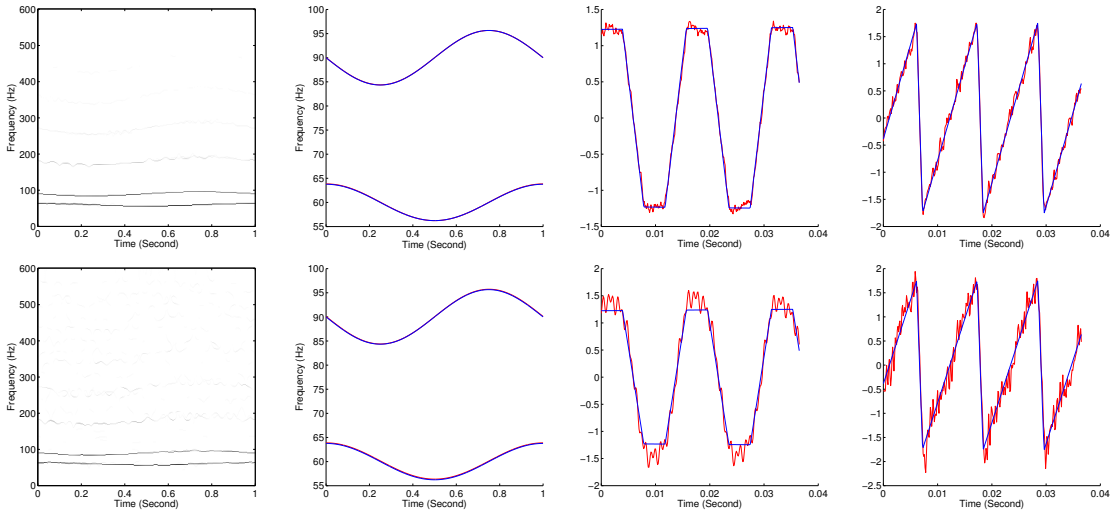


Figure 5.9: Noisy Example 1. Top: SNR = 6. Bottom: SNR = −3. Left: The synchrosqueezed energy distributions of signals. Middle left: The real instantaneous frequencies (blue) and the estimated instantaneous frequencies (red). Middle right and right: Recovered general modes.

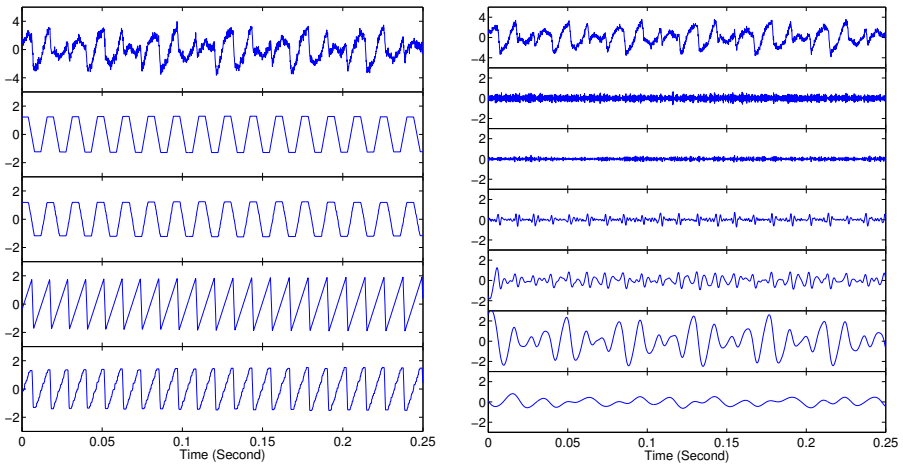


Figure 5.10: Left: Noisy signals of Example 1 with SNR = 6. The first row: original data. The second and the third row: the first noiseless general mode and its recovered result. The fourth and the fifth row: The second noiseless general mode and its recovered result. Right: The first row: original data. Rows below the first one: identified modes provided by the EEMD method.

5.3.2 Real Applications

Example 4**:** In the first example of real applications, we study ECG signals. Two real ECG general shape functions $s_5(x)$ and $s_6(x)$ (see Figure 5.12) are cut out from real ECG signals in [79] and [170]

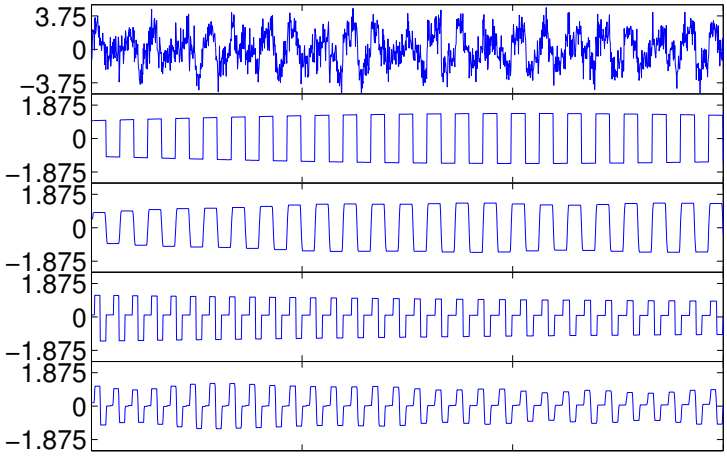


Figure 5.11: The first row: A noisy superposition with SNR = 0. The second and the third row: The first noiseless general mode and its recovered result. The fourth and the fifth row: The second noiseless general mode and its recovered result.

. A noisy superposition with SNR = 0 is generated as follows:

$$f(x) = \alpha_5(x)s_5(2\pi N_5\phi_5(x)) + \alpha_6(x)s_6(2\pi N_6\phi_6(x)) + n(x),$$

where $\alpha_5(x) = 1 + 0.05\sin(2\pi x)$, $\alpha_6(x) = 1 + 0.05\cos(2\pi x)$, $N_5 = 150$, $N_6 = 220$, $\phi_5(x) = x + 0.006\sin(2\pi x)$, and $\phi_6(x) = x + 0.006\cos(2\pi x)$. As shown in Figure 5.12, the synchrosqueezed energy distribution is well concentrated around the real instantaneous frequencies and the instantaneous frequencies are accurately estimated. Most importantly, the main spikes of real ECG shape functions are precisely recovered, even if the SNR is small. The decomposition results are plotted in Figure 5.13 left.

As a comparison, the EEMD method with the same parameters in previous examples is also applied to the same data. As shown in Figure 5.13 right, the EEMD method cannot provide useful mode decomposition results.

Example 5: Let us revisit the example shown in Figure 5.1 in the introduction. The original data $f_0(x)$ has a slowly growing trend linear in time. Suppose $f_r(x)$ is the linear regression of $f_0(x)$ and let $f(x) = f_0(x) - f_r(x)$. The synchrosqueezed energy distribution of $f(x)$ shown in Figure 5.14 left has three essential supports corresponding to three wave-like components. By weighting the locations of theses supports, we obtain the instantaneous frequency estimates of each component as shown in Figure 5.14. According to the evolutive pattern of the intrinsic frequencies, there are only two general modes contained in the superposition. The curve classification step in [178] automatically groups the annual estimate and the semiannual estimate together. Hence, the decomposition result

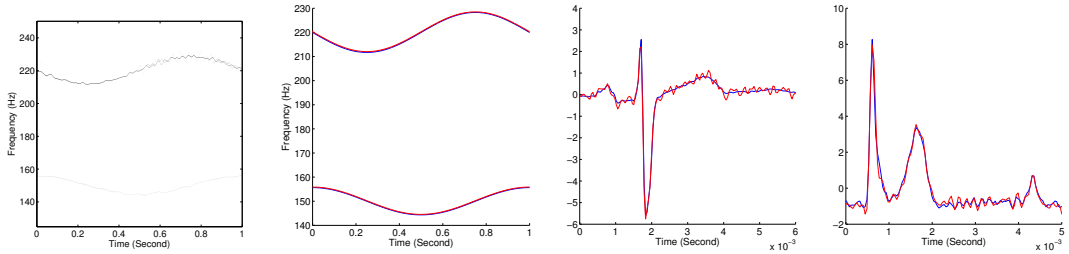


Figure 5.12: Left: The synchrosqueezed energy distribution of Example 5. Middle left: Real instantaneous frequencies (red) and instantaneous frequency estimates (blue). Middle right: Real ECG shape function $s_5(x)$ (blue) and its estimate (red). Right: Real ECG shape function $s_6(x)$ (blue) and its estimate (red).

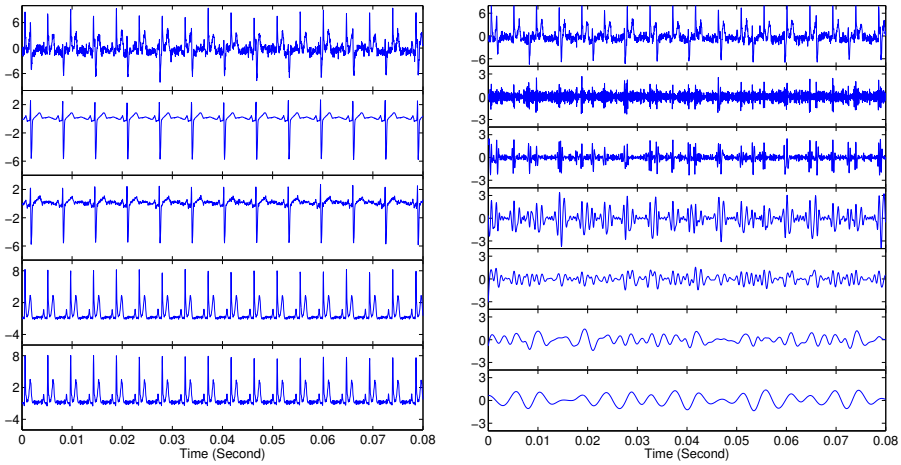


Figure 5.13: Left: The first row is a noisy superposition of two synthetic ECG signals with SNR = 0. The second and the third row: the first noiseless ECG component and its recovered result. The fourth and the fifth row: the second noiseless ECG component and its recovered result. Right: The first row is the same superposition of two ECG signals. Rows below the first one are identified modes provided by the EEMD method.

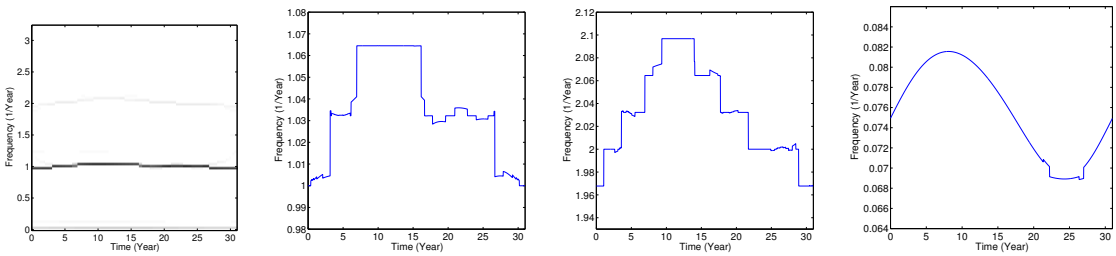


Figure 5.14: Left: $T_f(a,b)$ of the 31 years CO_2 concentration data. Middle left: The instantaneous frequency of the annual cycle. Middle right: The instantaneous frequency of the semiannual cycle. Right: the instantaneous frequency of the low frequency cycle. The curve classification algorithm groups the annual and semiannual cycle together.

contains a general mode that is the sum of the annual cycle and the semiannual cycle shown in Figure 5.15. Because of the low frequency of the third term, it is reasonable to combine it with $f_r(x)$ to obtain a slowly varying growing trend shown in Figure 5.15.

5.4 Conclusion

This chapter introduces the diffeomorphism based spectral analysis (DSA) method to solve the general mode decomposition problem. Given the instantaneous information, the DSA method is able to decompose a general superposition accurately. There are many future directions for the general mode decomposition problem. The most important work is to estimate the instantaneous information of each general mode without any well-separation condition. Another work of importance is the rigorous noise analysis of these methods. Although the numerical results have shown robustness against random Gaussian noise, a theoretical analysis is still missing. It is also of interest to study other types of noise and to explore the effects of noise on the reconstruction. Finally, it would be appealing to weaken the significant-difference condition of phase functions in Theorem 5.2.5 and to classify the class of significantly different phase functions.

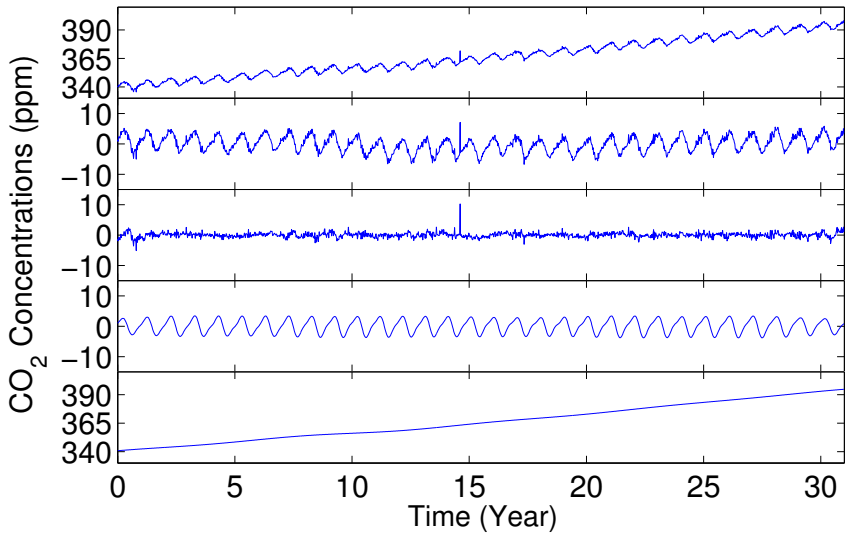


Figure 5.15: Decomposition results of the 31 years CO_2 concentration data. Top: The original data $f_0(x)$. Second row: The signal $f(x) = f_0(x) - f_r(x)$. Third row: The remaining noise term. Fourth row: The annual general mode. Last row: The slowly growing mode, which is the sum of $f_r(x)$ and the low frequency component.

Chapter 6

Applications

Oscillatory signals with nonlinear and non-stationary wave-like patterns are ubiquitous in science and engineering, e.g., clinical data [171, 173], seismic data [73, 86, 153], climate data [155, 176], astronomical data [20, 30], art forensics [179], and materials science [147, 180]. Analyzing instantaneous properties (e.g., instantaneous frequencies, instantaneous amplitudes and instantaneous phases [13, 139]) or local properties (concepts for 2D signals similar to "instantaneous" in 1D) of signals has been an important topic for over two decades. In this chapter, we introduce the application of synchrosqueezed transforms to two real problems in materials science (joint work with Jianfeng Lu, Benedikt Wirth, and Lexing Ying in [126, 180]) and art forensics (joint work with Jianfeng Lu, William P. Brown, Ingrid Daubechies, and Lexing Ying in [179]).

6.1 Atomic Crystal Analysis

6.1.1 Introduction

In materials science, crystal image analysis on a microscopic length scale has become an important research topic recently [189, 147, 12, 14, 150, 62, 64, 151, 152]. The development of image acquisition techniques (such as high resolution transmission electron microscopy (HR-TEM) [109]) and the advancement of atomic simulation of molecular dynamics [2] or mean field models like phase field crystals [61, 63] create data of large scale crystalline solids with defects at an atomic resolution. This provides unprecedented opportunities in understanding materials properties at a microscopic level.

Defects, like dislocations, grain boundaries, vacancies, etc., play a fundamental role in polycrystalline materials. They greatly change the material behavior from a perfect crystal and affect the macroscopic properties of the materials. Analysis of crystal images helps understand the defects and their effects on crystalline materials. While the defect analysis is traditionally done by visual inspection, the large amount of data made available due to advances in imaging and simulation

techniques creates a need of efficient computer-assisted or automated analysis.

Crystal deformation at the atomic scale is another important quantity that characterizes polycrystalline materials. When the deformation, denoted by ψ, is well-defined, the tensor field $F = \nabla\psi$ describes the local crystal strain; the polar decomposition of F at each point gives grain rotations; the curl of F provides information about defects and the well-known Burgers vector that represents the magnitude and direction of the lattice distortion resulting from a dislocation. Since it is almost impossible to estimate the deformation manually, the development of computer-aided analysis becomes important.

For crystalline materials, defects are physical domains of the materials such that it is not possible to identify a smooth crystal deformation $\phi = \psi^{-1}$ that maps the atomic configuration back to a perfect lattice. In other words, the deformation gradient $G = \nabla\phi = F^{-1}$ is irregular and has nonzero curl at the defect location. In the opposite case, when a smooth deformation map does exist, the affine transform given by the gradient of the map, $G = \nabla\phi$, transforms the image locally to an undistorted lattice of atoms. Therefore, for a defect-free region of the material, G is a gradient field and thus is curl-free: $\operatorname{curl} G = 0$. Crystal image analysis hence requires the detection of the defect regions and preferably also the estimation of the local elastic deformation G away from the defects.

In this work, we will limit our scope to $2D$ images of (slices of) polycrystalline materials and aim at extracting mesoscopic and microscopic information from the given images. This involves the identification of point defects, dislocations, deformations, grains and grain boundaries. Grains are material regions that are composed of a single crystal, possibly with different orientations (which will be referred as crystal rotations later, since we are working in $2D$); they are usually slightly deformed due to defects and interactions with neighboring grains at grain boundaries. Grains might contain point defects like vacancies and interstitials, for which we would like identify their positions. Crystal analysis should also be able to locate cores and Burgers vectors for dislocations, which play an important role in crystal plasticity. The estimated gradient field G should be curl-free in the interior of each grain. We refer the readers to [111] for more background details of polycrystals and crystal defects. Our proposed method provides a reliable and efficient way for extracting this information from crystal images.

Our contribution

Due to the lattice structure on the microscopic scale, crystal images are highly oscillatory (see Figure 6.1 (right) as an example). Inspired by this, we introduce a new characterization of grains by studying $2D$ general shape functions and $2D$ general intrinsic mode type functions, which are superpositions of nonlinear and non-stationary wave-like components (more precise definitions will be given in Section 6.1.2). Using these concepts, a crystal image can be considered as an assemblage of $2D$ general intrinsic mode type functions with non-overlapping supports, specified propagating directions and smoothly varying local wave vectors (see Figure 6.1 (left) as an example). In this model, crystal

defects and grain boundaries can be detected through the discontinuity and irregularity of these components; crystal rotations and crystal deformations are estimated from a linear system provided by local wave vectors of underlying wave-like components.

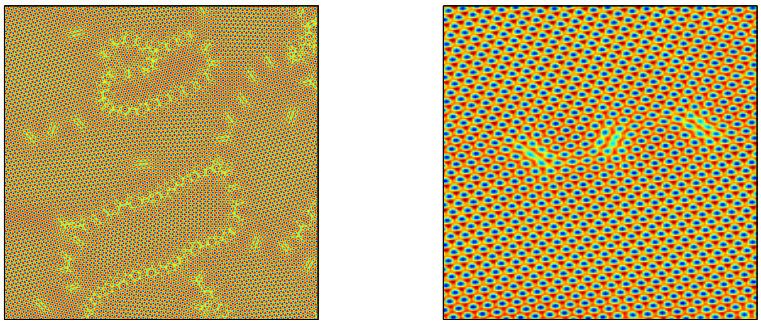


Figure 6.1: A phase field crystal (PFC) image and its zoomed-in image. Courtesy of Benedikt Wirth [64].

In this section, we propose a two-step method to analyze crystal images. In the first step, we adapt $2D$ synchrosqueezed transforms to the problems at hand. First, $2D$ synchrosqueezed transforms are applied to obtain the synchrosqueezed energy distributions of underlying wave-like components. Second, since the synchrosqueezed energy is concentrating on local wave vectors, these local wave vectors can be estimated by averaging the supports of the synchrosqueezed energy distribution. Third, the irregularity of each wave-like component can be measured by the irregularity of its corresponding synchrosqueezed energy distribution. Finally, with such information ready, the crystal image can be analyzed as discussed previously. In the second step, we propose a variational approach based on the information obtained in the first step. The optimization procedure improves the robustness of the analysis and, importantly, makes the results better agree with the physical nature of defects.

Previous works

One important class of methods for crystal image analysis is variation based. General variational methods for texture classification and segmentation have been extensively studied (see [131, 26, 165, 145, 166, 9], for example).

The method in [12] proposed to segment crystal images into disjoint regions with different constant crystal rotations using the Chan-Vese level-set approximation in [26] of the piecewise constant Mumford-Shah segmentation in [133]. The method involves search for a global deformation $\phi : \Omega \to \mathbb{R}^2$ acting on all grains. To speed up the expensive optimization in [12], the authors in [14] proposed a convex relaxation via functional lifting and the authors in [150] proposed a more efficient

version by penalizing the segmentation interfaces according to jumps in crystal rotations. Although a corresponding GPU implementation of these methods is very fast, a bottleneck still exists due to the large memory cost coming from the additional dimension for the functional lifting process.

More recently, Matt Elsey and Benedikt Wirth [62, 64] proposed a variational model based on finding a tensor map, the gradient of the inverse deformation $\nabla(\phi^{-1}) : \Omega \to \mathbb{R}^{2\times 2}$ and developed an efficient $L^1 - L^2$ regularization scheme. Crystal defects, rotations, grain boundaries and strain can be recovered by the information hidden in $\nabla(\phi^{-1})$. In addition, a corresponding GPU implementation is proposed and it shortens the runtime for a 1024^2 image to about 40 seconds. The variational method has been extended to 3D cases via sophisticated optimization algorithms in [65].

Another class of methods for texture classification and segmentation are based on a local, direction sensitive frequency analysis [159, 147]. [159] constructed an over-complete wavelet frame for texture feature extraction and segmented textures in a reduced feature space by clustering techniques. However, the frame is not sensitive to crystal rotations and local defects. Hence, it cannot distinguish two grains with a small angle boundary and cannot detect local defects. This method is neither capable of providing estimates of crystal deformations. The work [147] develops a heuristic method that uses a "wavelet like" patch according to a given reference crystal and quantify the similarity between local crystal patches with the reference. By looking up a prefabricated table of crystal rotation angles and their corresponding similarity, crystal rotations of each crystal patch can be estimated. However, in the case of deformed crystals, this method may be problematic.

The method to be presented in this section follows a different spirit from those methods. It is based on a novel model characterizing deformed periodic textures using $2D$ general intrinsic mode type functions and an efficient phase space representation method, $2D$ synchrosqueezed transforms proposed recently. An analytic characterization of periodic textures allows rigorous analysis and indeed it is proved that $2D$ synchrosqueezed transforms can estimate the local wave vectors of underlying wave-like components of textures (grains in this paper) precisely under certain conditions. Most of all, nonlinear deformations of crystals are available by solving a simple linear system provided by local wave vectors.

The rest of this section is organized as follows. In Section 6.1.2, we introduce a crystal image model on the microscopic length scale based on $2D$ general intrinsic mode type functions and prove that $2D$ synchrosqueezed transforms are able to estimate the local properties of the underlying wave-like components of general intrinsic mode type functions. In Section 6.1.3, two efficient algorithms based on $2D$ discrete band-limited synchrosqueezed transforms are proposed to detect crystal defects, estimate crystal rotations and elastic deformations. In Section 6.1.4, we present the variational optimization based on physical properties of atomic materials to refine the results provided by previous sections. In Section 6.1.5, several numerical examples of synthetic and real crystal images are provided to demonstrate the robustness and the reliability of our methods. Finally, we conclude with some discussion on future works in Section 6.1.6.

6.1.2 Crystal Image Models and Theory

In this section, at first, we describe a new model to characterize atomic crystal images. Inspired by the periodicity of crystal images, $2D$ general intrinsic mode type functions are defined as a key ingredient of the characterization of a perfect crystal image. Second, we prove that the $2D$ synchrosqueezed transforms accurately estimate the local wave vectors of these wave-like components, and hence provides a useful tool for crystal image analysis.

$2D$ general intrinsic mode type functions

A perfect crystal image, i.e., a single undeformed grain without defects, is characterized by a periodic function with two space variables. We will limit ourselves to simple crystals (Bravais lattices). In $2D$ space domain, there are five kinds of Bravais lattices: oblique, rectangular, centered rectangular (rhombic), hexagonal, and square [111]. In a discrete setting, we can denote the perfect reference lattice as

$$\mathcal{L} = \{c_1 a_1 + c_2 a_2 \,:\, c_1, c_2 \text{ integers}\},$$

where $a_1, a_2 \in \mathbb{R}^2$ represent two fixed lattice vectors. The lattices and corresponding unit cells are shown in Figure 6.2. For each lattice type, through an affine transform, we can transform the unit

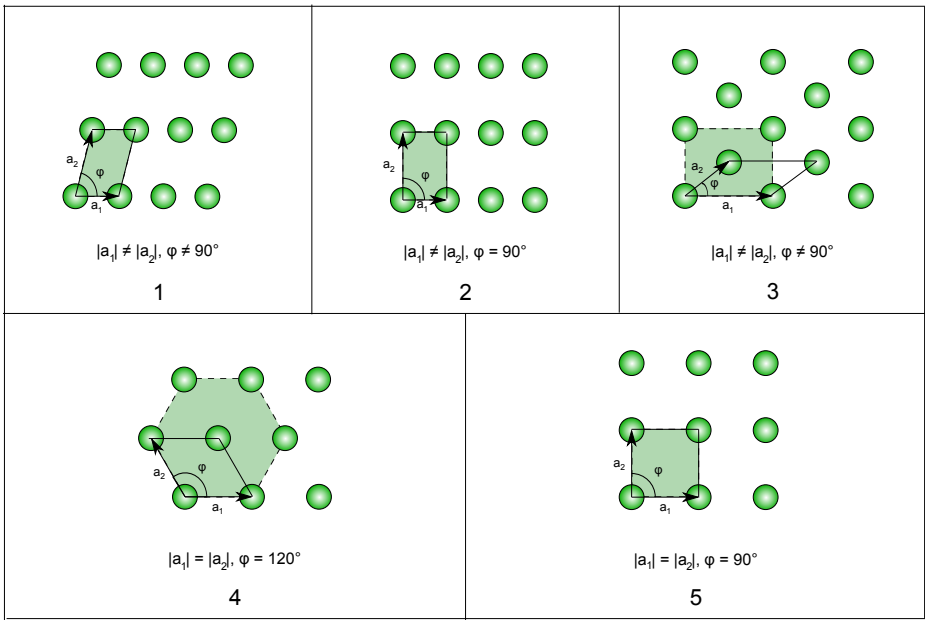


Figure 6.2: Five fundamental $2D$ Bravais lattices: 1 oblique, 2 rectangular, 3 centered rectangular (rhombic), 4 hexagonal, and 5 square. Courtesy of Wikipedia.

cell to a square. As an example, for the hexagonal lattice, the transform is given by

$$x \mapsto Fx; \quad F = \begin{pmatrix} 1 & -\frac{\sqrt{3}}{3} \\ 0 & \frac{2\sqrt{3}}{3} \end{pmatrix}.$$

Hence, by introducing the matrix F, we can set up a reference configuration function as

$$f(x) = \alpha S(2\pi Fx) + c,$$

for $x \in \mathbb{R}^2$. Here, $S(x)$ is a 2π periodic general shape function (the rigorous definition is given later), which has a unit $L^2([-\pi, \pi]^2)$-norm and a zero mean. α and c are two parameters.

Allowing a rotation and a translation, a crystal image function for an undeformed grain is then modeled by

$$f(x) = \alpha S(2\pi NF(R_\theta x + z)) + c,$$

where N is the reciprocal of the lattice parameter, R_θ is the rotation matrix

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

corresponding to a rotation angle θ, and $z \in \mathbb{R}^2$ gives the translation. In the case of multi-grains, it is expected that

$$f(x) = \sum_{k=1}^{M} \chi_{\Omega_k}(x) \left(\alpha_k S_k \left(2\pi N_k F_k (R_{\theta_k} x + z_k)\right) + c_k\right),$$

where $\chi_{\Omega_k}(x)$ is the indicator function defined as

$$\chi_{\Omega_k}(x) = \begin{cases} 1, & x \in \Omega_k \\ 0, & \text{otherwise}, \end{cases} \tag{6.1}$$

and Ω_k is the domain of the kth grain with $\Omega_k \cap \Omega_j = \emptyset$, if $k \neq j$. With these notations, grain boundaries are interpreted as $\cup\partial\Omega_k$ (in real crystal images, grain boundaries would be a thin transition region instead of a sharp boundary $\cup\partial\Omega_k$). In the presence of local defects, e.g., an isolated defect and a terminating line of defects, $\cup\partial\Omega_k$ may include irregular boundaries and may contain point boundaries inside $\cup\Omega_k$.

Considering an uneven distribution of atoms on the mesoscopic length scale and possible reflection of light when generating crystal images, the amplitudes α_k and the global trends c_k are assumed to be smooth functions $\alpha_k(x)$ and $c_k(x)$, respectively, in the domain Ω_k.

Notice that the rotation matrix R_{θ_k} and the translation position z_k act as a linear transformation ψ_k from x to $\psi_k(x) = R_{-\theta_k}(x - z_k)$. In a more complicated case, a smooth nonlinear deformation

$\psi_k : \mathbb{R}^2 \to \mathbb{R}^2$ transferring an atom from position x to $\psi_k(x)$ is introduced. Let $\phi_k(x) = \psi_k^{-1}(x)$ defined in Ω_k, then the crystal image function becomes

$$f(x) = \sum_{k=1}^{M} \chi_{\Omega_k}(x) \left(\alpha_k(x) S_k \left(2\pi N_k F_k \phi_k(x)\right) + c_k(x)\right). \tag{6.2}$$

This motivates the definition of $2D$ general shape functions and $2D$ general intrinsic mode type functions as follows.

Definition 6.1.1 ($2D$ general shape function)**.** *The $2D$ general shape function class S_M consists of periodic functions $S(x)$ with a periodicity $(2\pi, 2\pi)$, a unit $L^2([-\pi,\pi]^2)$-norm, and an L^∞-norm bounded by M satisfying the following conditions:*

1. *The $2D$ Fourier series of $S(x)$ is uniformly convergent;*
2. *$\sum_{n\in\mathbb{Z}^2} |\widehat{S}(n)| \le M$ and $\widehat{S}(0,0) = 0$;*
3. *Let Λ be the set of integers $\{|n_1| \in \mathbb{N} : \widehat{S}(n_1, n_2) \neq 0$ or $\widehat{S}(n_2, n_1) \neq 0$ for some $n_2 \in \mathbb{Z}\}$. The greatest common divisor of all the elements in Λ is 1.*

The requirement that $\widehat{S}(0,0) = 0$, which is equivalent to a zero mean over $[-\pi,\pi]^2$, guarantees a well-separation between the oscillatory part $\alpha_k(x) S_k\left(2\pi N_k F_k \phi_k(x)\right)$ and the smooth trend $c_k(x)$ in (6.2), when N_k is sufficiently large. The third condition implies the uniqueness of similar oscillatory patterns in S_M up to a scaling, i.e., if $S(x) \in S_M$, $S(Nx) \notin S_M$ for any positive integer $N > 1$.

Definition 6.1.2 ($2D$ general intrinsic mode type function (GIMT))**.** *$f(x) = \alpha(x) s(2\pi N F \phi(x))$ is a $2D$ GIMT of type (M, N, F), if $S(x) \in S_M$, $\alpha(x)$ and $\phi(x)$ satisfy the conditions below.*

$$\begin{aligned}
&\alpha(x) \in C^\infty, \quad |\nabla\alpha| \le M, \quad 1/M \le \alpha \le M,\\
&\phi(x) \in C^\infty, \quad 1/M \le \left|\nabla(n^\mathcal{T} F\phi)/|n^\mathcal{T} F|\right| \le M, \quad \textit{and}\\
&\left|\nabla^2(n^\mathcal{T} F\phi)/|n^\mathcal{T} F|\right| \le M, \quad \forall n \in \mathbb{Z}^2 \quad \textit{s.t.} \quad \widehat{s}(n) \neq 0.
\end{aligned}$$

Hence, in the domain Ω_k of each grain, the crystal image is a superposition of a $2D$ general intrinsic mode type function and a smooth trend. Applying the $2D$ Fourier series of each $2D$ general shape function $S_k(x)$, it holds that

$$\begin{aligned}
f(x) &= \sum_{k=1}^{M} \chi_{\Omega_k}(x) \left(\alpha_k(x) S_k\left(2\pi N_k F_k \phi_k(x)\right) + c_k(x)\right)\\
&= \sum_{k=1}^{M} \chi_{\Omega_k}(x) \left(\sum_{n\in\mathbb{Z}^2} \alpha_k(x) \widehat{S_k}(n) e^{2\pi i N_k n^\mathcal{T} F_k \phi_k(x)} + c_k(x)\right),
\end{aligned} \tag{6.3}$$

In the domain Ω_k, each underlying wave-like component

$$\alpha_k(x)\widehat{S_k}(n)e^{2\pi i N_k n^{\mathcal{T}} F_k \phi_k(x)}$$

is an intrinsic mode type function studied in [182, 184]. Hence, if they satisfy the well-separation condition defined in [182, 184] and each N_k is large enough, the $2D$ synchrosqueezed wave packet transforms and the synchrosqueezed curvelet transforms are expected to estimate the local wave vectors $N_k\nabla(n^{\mathcal{T}}F_k\phi_k(x))$ accurately for x away from $\partial\Omega_k$.

Based on the estimates of local wave vectors, it is possible to define and analyze crystal rotations as follows.

Definition 6.1.3. *Given a reference configuration $S(2\pi NFx)$, a deformation $\phi(x)$, an amplitude function $\alpha(x)$ and a trend function $c(x)$ such that $f(x) = \alpha(x)S(2\pi NF\phi(x)) + c(x)$ is a $2D$ GIMT of type (M, N, F), suppose $\widehat{S}(n) \neq 0$, then the reference configuration has a local wave vector $v(n) = Nn^{\mathcal{T}}F$ and $f(x)$ has a local wave vector $v_\phi = Nn^{\mathcal{T}}F\nabla\phi(x)$. The local rotation function of $f(x)$ with respect to $v(n)$ is defined as*

$$\beta(f)(x) = \arg(v_\phi(x)) - \arg(v(n)),$$

where $\arg(v)$ *means the argument of a vector v.*

In the case of an undeformed crystal image $f(x) = \alpha(x)S(2\pi NF(R_\theta x+z))+c(x)$, $\phi(x) = R_\theta x+z$ is a composition of a rotation and a translation. Then the local rotation function $\beta(f)(x) = \theta$ with respect to any local wave vector $v(n)$. This agrees with an intuition of a global crystal rotation. However, a global crystal rotation is not well defined in a real crystal image due to a nonlinear crystal deformation. This motivates the definition of a local crystal rotation in Definition 6.1.3. Because the nonlinear deformation $\phi(x)$ is a smooth function with $\det(\nabla\phi(x)) \approx 1$, local rotation functions $\beta(f)(x)$ vary smoothly and are approximately the same with respect to any local wave vector.

$2D$ synchrosqueezed transforms

Our goal is to apply and adapt the $2D$ synchrosqueezed transforms to analyze the $2D$ general intrinsic mode type functions in the image function (6.3). This problem is similar to but different from the $1D$ general mode decomposition problems studied in [170, 178], where $1D$ synchrosqueezed transforms are applied to estimate the instantaneous properties of $1D$ general intrinsic mode type functions. Generalizing the conclusions in [178, 184], the theorem below shows that the $2D$ synchrosqueezed transforms precisely estimate the local wave vectors of the wave-like components in (6.3) at the points away from boundaries.

Theorem 6.1.4. *For a $2D$ general intrinsic mode type function $f(x)$ of type (M, N, F) with $|F| \geq 1$,*

any $\epsilon > 0$ and any $r > 1$, we define

$$R_\epsilon = \left\{(a,\theta,b) : |W_f(a,\theta,b)| \geq a^{-\frac{s+t}{2}}\sqrt{\epsilon}, \quad a \leq 2MNr\right\}$$

and

$$Z_n = \left\{(a,\theta,b) : \left|A_a^{-1}R_\theta^{-1}\left(a\cdot u_\theta - N\nabla(n^{\mathcal{T}}F\phi(b))\right)\right| \leq 1, \quad a \leq 2MNr\right\}$$

For fixed M, r, s, t, and ϵ, there exists $N_0(M,r,s,t,\epsilon) > 0$ such that for any $N > N_0(M,r,s,t,\epsilon)$ and a 2D GIMT $f(x)$ of type (M,N,F), the following statements hold.

(i) $\left\{Z_n : \widehat{S}(n) \neq 0\right\}$ are disjoint and $R_\epsilon \subset \bigcup_{\widehat{S}(n)\neq 0} Z_n$;

(ii) For any $(a,\theta,b) \in R_\epsilon \cap Z_n$,

$$\frac{\left|v_f(a,\theta,b) - N\nabla(n^{\mathcal{T}}F\phi(b))\right|}{|N\nabla(n^{\mathcal{T}}F\phi(b))|} \lesssim \sqrt{\epsilon}.$$

For simplicity, the notations $\mathcal{O}(\cdot)$, $\lesssim$ and $\gtrsim$ are used when the implicit constants may only depend on M, s, t, and K. The proof of the theorem is similar to those theorems in Section 2.3 and [184, 178] and relies on two lemmas that have been proved in Section 2.3. Let us recall these lemmas again.

Lemma 6.1.5. *Suppose $f(x) = \sum_{k=1}^K f_k(x) = \sum_{k=1}^K e^{-(\phi_k(x)-c_k)^2/\sigma_k^2}\alpha_k(x)e^{2\pi i N\phi_k(x)}$ is a well-separated superposition of type (M,N,K). Set*

$$\Omega = \left\{(a,\theta) : a \in \left(\frac{N}{2M}, 2MN\right), \exists k \ s.t. \ \left|\theta_{\nabla\phi_k(b)} - \theta\right| < \theta_0\right\},$$

where $\theta_0 = \arcsin((\frac{M}{N})^{t-s})$. For any $\epsilon > 0$, there exists $N_0(M,s,t,\epsilon)$ such that the following estimation of $W_f(a,\theta,b)$ holds for any $N > N_0$.

(1) If $(a,\theta) \in \Omega$,

$$W_f(a,\theta,b) = a^{-\frac{s+t}{2}}\left(\sum_{k:\ |\theta_{\nabla\phi_k(b)}-\theta|<\theta_0} f_k(b)\widehat{w}\left(A_a^{-1}R_\theta^{-1}\left(a\cdot u_\theta - N\nabla\phi_k(b)\right)\right) + \mathcal{O}(\epsilon)\right);$$

(2) Otherwise,

$$W_f(a,\theta,b) = a^{-\frac{s+t}{2}}\mathcal{O}(\epsilon).$$

Lemma 6.1.6. *Suppose $f(x) = \sum_{k=1}^K f_k(x) = \sum_{k=1}^K e^{-(\phi_k(x)-c_k)^2/\sigma_k^2}\alpha_k(x)e^{2\pi i N\phi_k(x)}$ is a well-separated superposition of type (M,N,K). For any $\epsilon > 0$, there exists $N_0(M,s,t,\epsilon)$ such that the*

following estimation of $\nabla_b W_f(a,\theta,b)$ *holds for any* $N > N_0$.

$$\nabla_b W_f(a,\theta,b) = a^{-\frac{s+t}{2}} \left(2\pi i N \sum_{k:\ |\theta_{\nabla\phi_k(b)}-\theta|<\theta_0} \nabla\phi_k(b) f_k(b) \widehat{w}\left(A_a^{-1}R_\theta^{-1}(a\cdot u_\theta - N\nabla\phi(b))\right) + \mathcal{O}(\epsilon) \right),$$

when

$$(a,\theta)\in\Omega = \left\{ (a,\theta) : a \in \left(\frac{N}{2M}, 2MN\right), \exists k \ s.t.\ \left|\theta_{\nabla\phi_k(b)} - \theta\right| < \theta_0 \right\}.$$

In the scope of this section, we have $\sigma_k = \infty$ for all k. Now, we are ready to prove Theorem 6.1.4.

Proof of Theorem 6.1.4. By the uniform convergence of the $2D$ Fourier series of general shape functions, we have

$$W_f(a,\theta,b) = \sum_{n\in\mathbb{Z}^2} W_{f_n}(a,\theta,b),$$

where $f_n(x) = \widehat{S}(n)\alpha(x)e^{2\pi i N n^{\mathcal{T}} F\phi(x)}$. Introduce the short hand notation, $\widetilde{\phi}_n(x) = n^{\mathcal{T}}F\phi(x)/|n^{\mathcal{T}}F|$, then

$$f_n(x) = \widehat{S}(n)\alpha(x)e^{2\pi i N|n^{\mathcal{T}}F|\widetilde{\phi}_n(x)}.$$

By the property of $2D$ general intrinsic mode functions, $f_n(x)$ is a well-separated superposition of type $(M, N|n^{\mathcal{T}}F|, 1)$ defined in Section 2.3.

For each n, we estimate $W_{f_n}(a,\theta,b)$. By Lemma 6.1.5, there exists a uniform $N_1(M,s,t,d,\epsilon)$ independent of n such that, if $N|n^{\mathcal{T}}F| > N_1$,

$$W_{f_n}(a,\theta,b) = a^{-\frac{s+t}{2}} \left(f_n(b)\widehat{w}\left(A_a^{-1}R_\theta^{-1}\left(a\cdot u_\theta - N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b)\right)\right) + \left|\widehat{S}(n)\right|\mathcal{O}(\epsilon)\right)$$

for $(a,\theta)\in\Omega_n$, and

$$W_{f_n}(a,\theta,b) = a^{-\frac{s+t}{2}}\left|\widehat{S}(n)\right|\mathcal{O}(\epsilon)$$

for $(a,\theta)\notin\Omega_n$. Here

$$\Omega_n = \left\{ (a,\theta) : a\in\left(\frac{N|n^{\mathcal{T}}F|}{2M}, 2MN|n^{\mathcal{T}}F|\right), \left|\theta_{\nabla\widetilde{\phi}_n(b)} - \theta\right| < \theta_0 \right\},$$

and $\theta_0 = \arcsin((\frac{M}{N|n^{\mathcal{T}}F|})^{t-s})$. Let $\Gamma_{a\theta} = \{n\in\mathbb{Z}^2 : (a,\theta)\in\Omega_n\}$, then

$$\begin{aligned} W_f(a,\theta,b) &= a^{-\frac{s+t}{2}}\left(\sum_{n\in\Gamma_{a\theta}} f_n(b)\widehat{w}\left(A_a^{-1}R_\theta^{-1}\left(a\cdot u_\theta - N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b)\right)\right) + \sum_n \left|\widehat{S}(n)\right|\mathcal{O}(\epsilon)\right) \\ &= a^{-\frac{s+t}{2}}\left(\sum_{n\in\Gamma_{a\theta}} f_n(b)\widehat{w}\left(A_a^{-1}R_\theta^{-1}\left(a\cdot u_\theta - N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b)\right)\right) + \mathcal{O}(\epsilon)\right). \end{aligned}$$

Notice that for any $n \neq \widetilde{n}$, the distance between the local wave vectors $N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b)$ and $N|\widetilde{n}^{\mathcal{T}}F|\nabla\widetilde{\phi}_{\widetilde{n}}(b)$ are bounded below. In fact,

$$\begin{aligned}\left|N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b) - N|\widetilde{n}^{\mathcal{T}}F|\nabla\widetilde{\phi}_{\widetilde{n}}(b)\right| &= N\left|(n-\widetilde{n})^{\mathcal{T}}F\nabla\phi(b)\right| \\ &\geq \frac{N}{M}|(n-\widetilde{n})^{\mathcal{T}}F| \geq \frac{N}{M}.\end{aligned}$$

The first inequality above is due to the definition of $2D$ general intrinsic mode type function. Observe that the support of a general wave packet centered at $(a\cos(\theta), a\sin(\theta))$ is within a disk with a radius of length a^t. Because the range of a of interest is $a \leq 2MNr$, the general wave packets of interest have supports of size at most $2(2MNr)^t$. Hence, if $\frac{N}{M} \geq 2(2MNr)^t$, which is equivalent to $N \geq (2^{1+t}M^{1+t}r^t)^{\frac{1}{1-t}}$, then for each (a,θ,b) of interest, there is at most one $n \in \mathbb{Z}^2$ such that

$$\left|A_a^{-1}R_\theta^{-1}\left(a\cdot u_\theta - N\nabla(n^{\mathcal{T}}F\phi(b))\right)\right| \leq 1.$$

This implies that $\{Z_n\}$ are disjoint sets. Notice that $\widehat{w}(x)$ decays when $|x| \geq 1$. The above statement also indicates that there is at most one $n \in \Gamma_{a\theta}$ such that

$$f_n(b)\widehat{w}\left(A_a^{-1}R_\theta^{-1}\left(a\cdot u_\theta - N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b)\right)\right) \neq 0.$$

Hence, if $(a,\theta,b) \in R_\epsilon$, there must be some n such that $\widehat{S}(n) \neq 0$ and

$$W_f(a,\theta,b) = a^{-\frac{s+t}{2}}\left(f_n(b)\widehat{w}\left(A_a^{-1}R_\theta^{-1}\left(a\cdot u_\theta - N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b)\right)\right) + \mathcal{O}(\epsilon)\right). \tag{6.4}$$

By the definition of Z_n, we see $(a,\theta,b) \in Z_n$. So, $R_\epsilon \subset \bigcup_{\widehat{S}(n)\neq 0} Z_n$ and (1) is proved.

Now, we estimate $\nabla_b W_f(a,\theta,b)$. Suppose $(a,\theta,b) \in R_\epsilon \cap Z_n$. Similar to the estimate of $W_f(a,\theta,b)$, by Lemma 6.1.6, there exits $N_2(M,s,t,\epsilon)$ such that if $N > N_2$ then

$$\begin{aligned}&\nabla_b W_f(a,\theta,b) \\ &= a^{-\frac{s+t}{2}}\left(2\pi i N\sum_{n\in\Gamma_{a\theta}}|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b)f_n(b)\widehat{w}\left(A_a^{-1}R_\theta^{-1}(a\cdot u_\theta - N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b))\right) + \mathcal{O}(\epsilon)\right) \\ &= a^{-\frac{s+t}{2}}\left(2\pi i N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b)f_n(b)\widehat{w}\left(A_a^{-1}R_\theta^{-1}(a\cdot u_\theta - N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b))\right) + \mathcal{O}(\epsilon)\right)\end{aligned}$$

for the same n in (6.4).

Let $g = f_n(b)\widehat{w}\left(A_a^{-1}R_\theta^{-1}(a\cdot u_\theta - N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b))\right)$, then

$$v_f(a,\theta,b) = \frac{N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b)g + \mathcal{O}(\epsilon)}{g + \mathcal{O}(\epsilon)}.$$

Since $|W_f(a,\theta,b)| \geq a^{-\frac{s+t}{2}}\sqrt{\epsilon}$ for $(a,\theta,b) \in R_\epsilon$, then $|g| \gtrsim \sqrt{\epsilon}$. So

$$\begin{aligned}\frac{\left|v_f(a,\theta,b) - N\nabla(n^{\mathcal{T}}F\phi(b))\right|}{|N\nabla(n^{\mathcal{T}}F\phi(b))|} &= \frac{\left|v_f(a,\theta,b) - N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b)\right|}{\left|N|n^{\mathcal{T}}F|\nabla\widetilde{\phi}_n(b)\right|} \\ &\lesssim \left|\frac{\mathcal{O}(\epsilon)}{g+\mathcal{O}(\epsilon)}\right| \\ &\lesssim \sqrt{\epsilon}.\end{aligned}$$

The above estimate holds for $N > N_0 = \max\{N_1, N_2, (2^{1+t}M^{1+t}r^t)^{\frac{1}{1-t}}\}$. The proof is complete. □

In the crystal image analysis, grains would have local defects, irregular boundaries and smooth trends caused by reflection etc. First, it is important to know where we can keep the local wave vector estimates accurate. Suppose a $2D$ general intrinsic mode type function $\widehat{S}(n)\alpha(x)e^{2\pi iNn^{\mathcal{T}}F\phi(x)}$ is defined in a domain Ω with a boundary $\partial\Omega$. The smallest scale used to estimate local wave vectors is of order $\frac{N}{M}$. Hence, the general wave packets used have supports of size at most $\frac{M^t}{N^t}$ by $\frac{M^t}{N^t}$. Let us denote a perfect interior of Ω as

$$\widetilde{\Omega} = \left\{x \in \Omega : |x-y| > \frac{M^t}{N^t}, \forall y \in \partial\Omega\right\},$$

then the estimates of local wave vectors remain accurate in $\widetilde{\Omega}$. As the numerical results in [43, 182, 184, 178] show, synchrosqueezed transforms can still approximately recover instantaneous or local properties near $\partial\Omega$.

The second concern is the influence of the smooth trends $c_k(x)$ in (6.2). By the method of stationary phase, the Fourier transform $\widehat{c_k}(\xi)$ would decay quickly as $|\xi|$ increases. Hence, the influence of smooth trends is essentially negligible when the synchrosqueezed transforms are applied to estimate local wave vectors $Nn^{\mathcal{T}}F\nabla\phi(x))$ with a sufficiently large wave number.

6.1.3 Crystal Defect Analysis Algorithms and Implementations

This section introduces several algorithms based on the features of crystal images and $2D$ synchrosqueezed transforms for a fast analysis of local defects, crystal rotations and deformations. As discussed in the introduction, we will focus on the analysis of an image with only one type of crystals, i.e., suppose

$$\begin{aligned}f(x) &= \sum_{k=1}^{M}\chi_{\Omega_k}(x)\left(\alpha_k(x)S\left(2\pi NF\phi_k(x)\right) + c_k(x)\right) \\ &= \sum_{k=1}^{M}\chi_{\Omega_k}(x)\left(\sum_{n\in\mathbb{Z}^2}\widehat{S}(n)\alpha_k(x)e^{2\pi iNn^{\mathcal{T}}F\phi(x)} + c_k(x)\right).\end{aligned}$$

In the $2D$ space domain, there are five kinds of Bravais lattices, which are oblique, rectangular, centered rectangular (rhombic), hexagonal, and square [111] as shown in Figure 6.2. Accordingly, there are five kinds of $2D$ Fourier power spectra $|\widehat{S}(n)|$. These spectra have a few dominant wave vectors in terms of energy, i.e., a few $|\widehat{S}(n)|$ with large values. At each interior point of a grain, a sufficiently localized Fourier transform is able to recover an approximate distribution of the $2D$ Fourier power spectrum, based on which the corresponding reference configuration of this grain can be specified. Hence, for the simplicity of a presentation, we will restrict ourselves to the analysis of images of hexagonal crystals (e.g. Figure 6.3 (left)). A generalization of our algorithms to other kinds of simple crystal images is straightforward. In the case of a complex lattice, there might be numerous local wave vectors with relatively large energy. Feature extraction and dimension reduction techniques should be applied to provide a few crucial local wave vectors. This would be an interesting future work.

We will introduce two fast algorithms for crystal image analysis. Algorithm 6.1.7 provides estimates of grain boundaries and crystal rotations; while Algorithm 6.1.8 further identifies point defects, dislocations, and deformations. To make our presentation more transparent, the algorithms and implementations are introduced with toy examples. For Algorithm 6.1.7, we will use the example in Figure 6.3 (left) which contains two undeformed grains with a straight grain boundary. While Figure 6.3 is a synthetic example, it illustrates nicely the key feature of atomic crystal images and the idea of local phase plane spectrum. In this example, the reciprocal lattice parameter $N = 120$ and the crystal rotations are given by 15 and 52.5 degrees on the left and right respectively. We introduce Algorithm 6.1.8 using strained examples of a small angle boundary (see Figure 6.6 (left)) and with some isolated dislocations (see Figure 6.6 (right)). These are examples from phase field crystal simulations [63].

Band-limited $2D$ fast SST

Typically, each grain

$$\chi_{\Omega_k}(x)\left(\alpha_k(x)S\left(2\pi NF\phi_k(x)\right)+c_k(x)\right)$$

in a polycrystalline crystal image can be identified as a $2D$ general intrinsic mode type function of type (M, N, F) with a small M near 1, unless the strain is too large. Hence, the $2D$ Fourier power spectrum of a multi-grain image would have several well-separated nonzero energy annuli centered at the origin due to crystal rotations (see an example shown in Figure 6.3 (middle)). Suppose the radially average Fourier power spectrum is defined as

$$E(r)=\frac{1}{r}\int_0^{2\pi}\left|\widehat{f}(r,\theta)\right|d\theta,$$

where $\widehat{f}(r,\theta) = \widehat{f}(\xi)$ for $\xi = (r\cos(\theta), r\sin(\theta))^T \in \mathbb{R}^2$. Then $E(r)$ would have several well-separated energy bumps (see Figure 6.3 (right)) for the same reason.

As we can see in Figure 6.3 (middle), a hexagonal crystal image with a single grain

$$f(x) = \alpha(x) S\left(2\pi N F \phi(x)\right) + c(x)$$

has six dominant local wave vectors close to $v_j(\theta(x))$, $j = 0, 1, \ldots, 5$, which are the vertices of a hexagon centered at the origin in the Fourier domain, i.e.,

$$v_j(\theta(x)) = \left(N\cos(\theta(x) + \frac{j\pi}{3}), N\sin(\theta(x) + \frac{j\pi}{3})\right)^{\mathcal{T}}, \quad j = 0, 1, \ldots, 5,$$

for $\theta(x) \in [0, \frac{\pi}{3})$. Suppose $S\left(2\pi N F(x)\right)$ is a reference configuration, then the local rotation function with respect to each vertex $v_j(0)$ is

$$\beta(f)(x) \approx \arg(v_j(\theta(x))) - \arg(v_j(0)) = \theta(x). \tag{6.5}$$

Actually, $f(x)$ can approximately be considered as a rotated version of $\alpha S\left(2\pi N F(x+z)\right) + c$ by an angle $\theta(x) + \frac{k\pi}{3}$ for any $k \in \mathbb{Z}$ due to the crystal symmetry. The restriction $\theta(x) \in [0, \frac{\pi}{3})$ guarantees a unique local rotation function $\beta(f)(x)$.

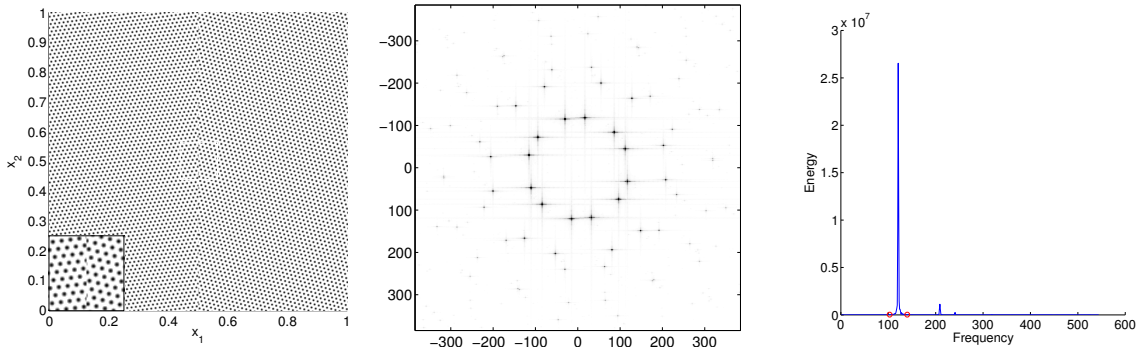


Figure 6.3: Left: An undeformed example of two hexagonal grains with a vertical line boundary. Middle: Its Fourier power spectrum. The number of dominant local wave vectors is 12 due to the superposition of spectrum of the two grains. Right: Its radially average Fourier power spectrum with the identified most dominant energy bump indicated by two red circles.

The discussion above shows that it is sufficient to compute the local rotation function using the dominant local wave vectors. To reduce the computational cost of $2D$ SST, the support of the most dominant energy bump in the radially average Fourier power spectrum should be identified (see Figure 6.3 (right)). Suppose the support (i.e. frequency band) is $[r_1, r_2]$. Then a band-limited $2D$ fast SST can be introduced following a similar methodology in [184] to estimate local wave vectors

with wave numbers in $[r_1, r_2]$ provided by the frequency band detection.

As an example, Figure 6.4 shows the synchrosqueezed energy distribution $T_f(a, \theta, b)$ in a polar coordinate at three different positions. Because the crystal image is real, it is enough to compute the synchrosqueezed energy distribution for $\theta \in [0, \pi)$. The results show that the essential support of $T_f(a, \theta, b)$ can accurately estimate local wave vectors when location b is not at the boundary. When b is at the boundary, the essential support of $T_f(a, \theta, b)$ can still provide some information, e.g. crystal rotations.

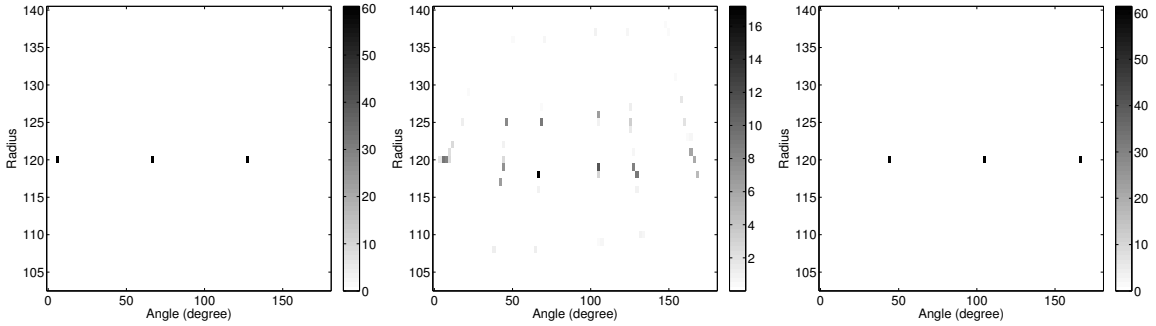


Figure 6.4: The synchrosqueezed energy distribution $T_f(a, \theta, b)$ of Figure 6.3 (left) at three different points b_i, $i = 1, 2, 3$. Left: $b_1 = (0.25, 0.5)$ is in the middle of the left grain. Middle: $b_2 = (0.5, 0.5)$ is at the boundary. Right: $b_3 = (0.75, 0.5)$ is in the middle of the right grain. Note that the scale of colorbar is different in the middle panel.

Defect detection algorithms

As we can see in Figure 6.4, the synchrosqueezed energy around each local wave vector $\nabla\left(Nn^{\mathcal{T}}F\phi(x)\right)$ is stable and of order $|\widehat{S}(n)|\alpha_k(x)$ when x is in the perfect interior $\widetilde{\Omega}_k$. Moreover, the energy would decrease fast near the boundary $\partial\Omega_k$ and becomes zero soon outside Ω_k. This motivates the application of synchrosqueezed energy distribution to identify grain boundaries by detecting the irregularity of energy distribution as follows.

Algorithm 6.1.7 (Fast defect detection algorithm based on stacked synchrosqueezed energy)**.**

- ***Step 1:*** *Stack the synchrosqueezed energy distribution and define a stacked synchrosqueezed energy distribution as*

$$\widetilde{T}_f(a, \theta, b) = T_f(a, \theta, b) + T_f(a, \theta + \frac{\pi}{3}, b) + T_f(a, \theta + \frac{2\pi}{3}, b)$$

for $\theta \in [0, \frac{\pi}{3})$.

- ***Step 2:*** *Compute an angular total energy distribution*

$$E_a(\theta, b) = \int_{r_1}^{r_2} \widetilde{T}_f(a, \theta, b) da.$$

- ***Step 3:*** *For each b, identify the most dominant energy bump in $E_a(\theta, b)$. Denote the range of this bump as $[\theta_{11}(b), \theta_{12}(b)]$.*

- ***Step 4:*** *Compute the total energy of the first identified bump as*

$$TE_1(b) = \int_{\theta_{11}(b)}^{\theta_{12}(b)} E_a(\theta, b) d\theta.$$

- ***Step 5:*** *Compute a weighted average angle of the first bump as*

$$Angle(b) = \frac{1}{TE_1(b)} \int_{\theta_{11}(b)}^{\theta_{12}(b)} E_a(\theta, b) \theta d\theta.$$

- ***Step 6:*** *For each b, update E_a such that $E_a(\theta, b) = 0$, if $\theta \in [\theta_{11}(b), \theta_{12}(b)]$. Update the most dominant energy bump in E_a. Denote the range of this bump as $[\theta_{21}(b), \theta_{22}(b)]$.*

- ***Step 7:*** *Compute the total energy of the second identified bump by*

$$TE_2(b) = \int_{\theta_{21}(b)}^{\theta_{22}(b)} E_a(\theta, b) d\theta.$$

- ***Step 8:*** *Compute the boundary indicator function*

$$BD(b) = \frac{1}{\sqrt{TE_1(b) - TE_2(b) + 1}}.$$

The synchrosqueezed energy distribution of each grain behaves like an energy bump essentially supported in Ω_k and quickly decays outside Ω_k. Hence, two energy bumps have close energy in the transition area containing the grain boundary. Since $\mathrm{TE}_1(b)$ acts as the maximum of two energy distributions and $\mathrm{TE}_2(b)$ acts as the minimum (see Figure 6.5 (left)), $\mathrm{TE}_1(b) - \mathrm{TE}_2(b)$ decays near the grain boundary and the boundary indicator function $\mathrm{BD}(b)$ is relatively large at the grain boundary. As Figure 6.5 (middle) shows, the grain boundary can be identified from a gray-scale image of $\mathrm{BD}(b)$.

The local rotation function with respect to each local wave vector of the reference hexagonal configuration are approximately the same. Since, the synchrosqueezed energy distribution $T_f(a, \theta, b)$ is stacked together per $\frac{\pi}{3}$ in θ, the weighted average $\mathrm{Angle}(b)$ approximates the local rotation functions in a weighted average sense. As shown in Figure 6.5 (right), $\mathrm{Angle}(b)$ accurately reflects the crystal rotations in this example.

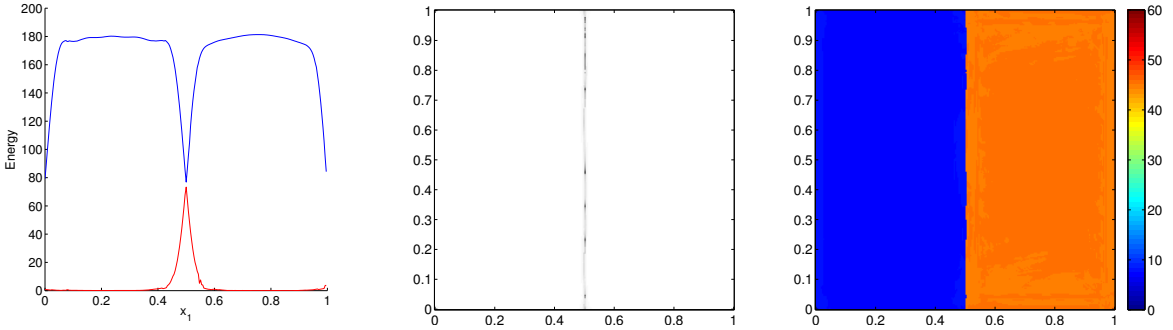


Figure 6.5: Left: The total energy function $TE_1(b)$ in blue and $TE_2(b)$ in red for fixed $b_2 = 0.5$. Middle: The boundary indicator function $\mathrm{BD}(b)$. Right: The weighted average angle $\mathrm{Angle}(b)$ as an approximation of crystal rotations. The real crystal rotations are 15 degrees on the left and 57.5 degrees on the right.

The stacking step in Algorithm 6.1.7 is averaging the influence of each local wave vector. This gives a stable result of grain boundary and crystal rotation estimates even with severe noise as will be illustrated in Section 6.1.5. However, Algorithm 6.1.7 might miss some local defects that would not influence all local wave vectors simultaneously. For example, in Figure 6.6 (left), two of the underlying wave-like components have smoothly changing directions, while the third one changes its direction suddenly at a line segment, resulting in a small angle boundary. At some local dislocations, as Figure 6.6 (right) shows, the dislocation might not cause irregularity to all wave-like components. Hence, to be more sensitive to local irregularity, it is reasonable to remove the stacking step in Algorithm 6.1.7, if noise is relatively small. This motivates the following algorithm.

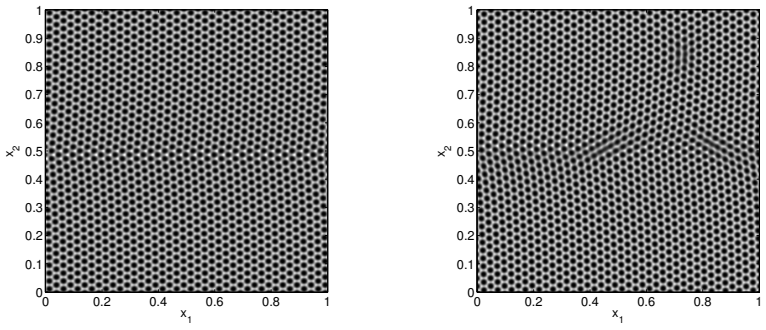


Figure 6.6: Left: An example of a small angle boundary. Right: An example of some isolated dislocations. Courtesy of Benedikt Wirth. The sizes of these images are 512×512 pixels.

Algorithm 6.1.8 (Fast defect detection algorithm with enhanced sensitivity)**.**

- ***Step 1:*** *Define* $\widetilde{T}_j(a,\theta,b) = T_f(a,\theta + j\pi/3, b)$ *for* $\theta \in [0, \frac{\pi}{3})$ *and* $j = 0, 1, 2$. *Apply Steps* $2-8$ *in Algorithm 6.1.7 to* $\widetilde{T}_j(a,\theta,b)$ *to compute* $TE_{1j}(b)$, $TE_{2j}(b)$, $Angle_j(b)$ *and* $BD_j(b)$ *for each* j.

- ***Step 2:*** *For each* $j = 0, 1, 2$, *compute the weight function*

$$W_j(b) = \frac{TE_{1j}(b) + TE_{2j}(b)}{\sum_j \left(TE_{1j}(b) + TE_{2j}(b)\right)}.$$

- ***Step 3:*** *Compute the weighted average angle*

$$Angle(b) = \sum_j W_j(b) Angle_j(b).$$

- ***Step 4:*** *Compute the weighted boundary indicator function*

$$BD(b) = \sum_j W_j(b) BD_j(b).$$

In the example in Figure 6.6 (left), only the second local wave vector exhibits irregularity at the small angle boundary, resulting in a sharp decrease of $\mathrm{TE}_{11}(b)$ and a sharp increase of $\mathrm{TE}_{21}(b)$ at the boundary. Hence, as shown in Figure 6.7, the weighted boundary indicator function $\mathrm{BD}(b)$ with $\mathrm{BD}_1(b)$ as a key integrand can clearly indicate the small angle boundary. As Figure 6.8 shows, the top point dislocation in the example in Figure 6.6 (right) interrupts the first and third underlying wave-like component, resulting in a sharp decrease in $\mathrm{TE}_{10}(b)$ and $\mathrm{TE}_{12}(b)$ and a sharp increase in $\mathrm{TE}_{20}(b)$ and $\mathrm{TE}_{22}(b)$. Therefore, the weighted boundary indicator function can reveal this point dislocation. The results in Figure 6.7 and 6.8 indicates that the information of some local defects is hidden behind some particular local wave vectors. By using the synchrosqueezed energy distribution of each local wave vector individually, more information of local defects can be discovered.

Recovery of inverse deformation gradient

In addition to grain boundaries and crystal rotations, a reliable extraction of the elastic deformation of a crystal image is also essential for an efficient material characterization. Instead of estimating the elastic inverse deformation $\phi(x)$ directly, we would emphasize how to recover the inverse deformation gradient

$$\nabla\phi(x) = \begin{pmatrix} \partial_{x_1}\phi_1(x) & \partial_{x_2}\phi_1(x) \\ \partial_{x_1}\phi_2(x) & \partial_{x_2}\phi_2(x) \end{pmatrix},$$

and how to read more information from $\nabla\phi(x)$, e.g., directions of Burgers vectors.

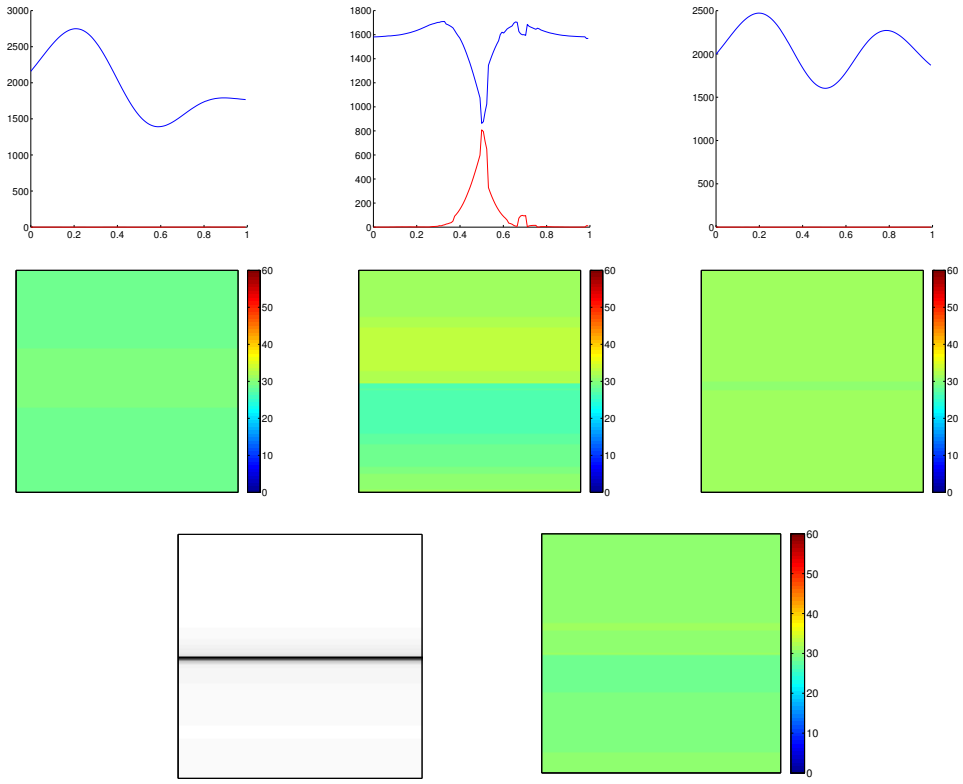


Figure 6.7: Top: $\mathrm{TE}_{1j}(b)$ (in blue) and $\mathrm{TE}_{2j}(b)$ (in red) of Figure 6.6 (left) for fixed $b_1 = 0.5$ and $j = 0, 1, 2$ from left to right, respectively. Center: The weighted average angle functions $\mathrm{Angle}_j(b)$ provided by each local wave vector for $j = 0, 1, 2$, respectively. Bottom left: the weighted boundary indicator function $\mathrm{BD}(b)$. Bottom right: the weighted average angle function $\mathrm{Angle}(b)$.

The estimation of the inverse deformation gradient $\nabla\phi(x)$ relies on the complete estimate of at least two local wave vectors $\nabla\left(Nn^{\mathcal{T}}F\phi(x)\right)$. Let us continue with hexagonal crystal images as an example. In this case, there are six local wave vectors of interest as discussed in Section 6.1.3. By symmetry, it is enough to consider those in the upper half plane of the Fourier domain. They are

$$v_j(x) = (\nabla\phi(x))^{\mathcal{T}} \left(N\cos(\frac{j\pi}{3}), N\sin(\frac{j\pi}{3})\right)^{\mathcal{T}}, \quad j = 0, 1, 2.$$

At each location x, we can estimate $v_j(x)$ efficiently by identifying peaks in the synchrosqueezed energy distribution $T_f(a, \theta, x)$. Denote these estimates as $v_j^{\mathrm{est}}(x)$, then we have an over-determined linear system at each x

$$v_j^{\mathrm{est}}(x) \approx (\nabla\phi(x))^{\mathcal{T}} \left(N\cos(\frac{j\pi}{3}), N\sin(\frac{j\pi}{3})\right)^{\mathcal{T}}, \quad j = 0, 1, 2.$$

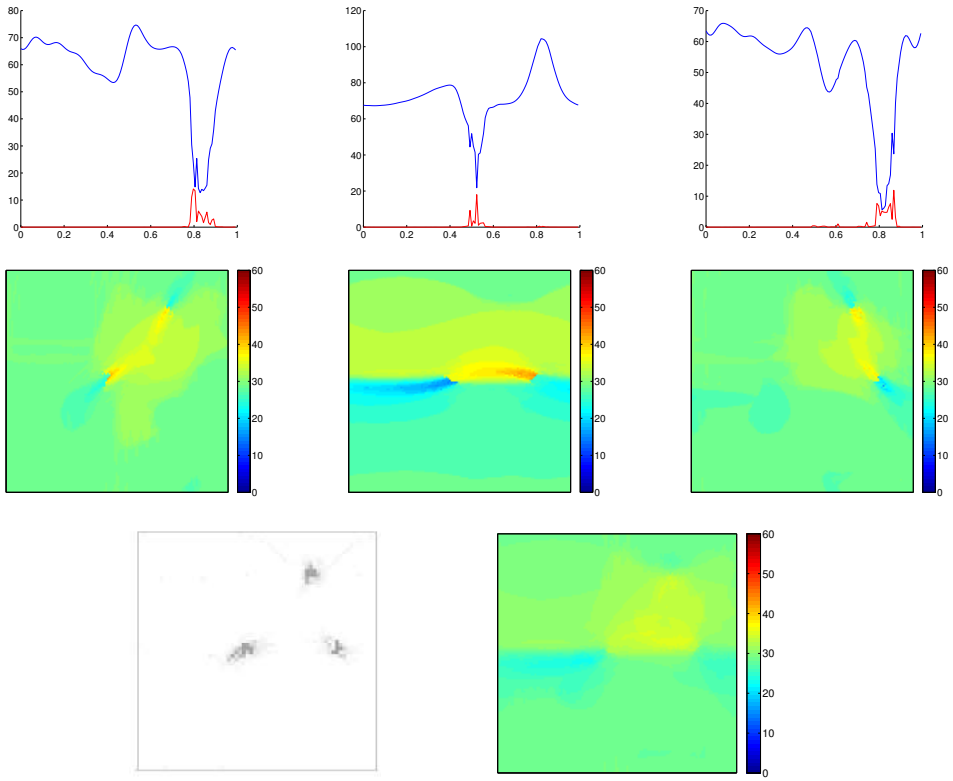


Figure 6.8: Top: $\mathrm{TE}_{1j}(b)$ (in blue) and $\mathrm{TE}_{2j}(b)$ (in red) of Figure 6.6 (right) for fixed $b_1 = 0.8$ and $j = 0, 1, 2$, respectively. Center: The weighted average angle functions $\mathrm{Angle}_j(b)$ provided by each local wave vector for $j = 0, 1, 2$, respectively. Bottom left: the weighted boundary indicator function $\mathrm{BD}(b)$. Bottom right: the weighted average angle function $\mathrm{Angle}(b)$.

Notice that the reciprocal number N can be estimated by $\arg\max E(r)$, where $E(r)$ is the radially average Fourier power spectrum of the given crystal image. Hence, a least square method is sufficient to provide a good estimate $G_0(x) \in \mathbb{R}^{2\times 2}$ of the inverse deformation gradient $\nabla\phi(x)$:

$$G_0(x) = \underset{G}{\operatorname{argmin}} \sum_{j=1}^{3} \left\| v_j^{\mathrm{est}}(x) - G\left(N\cos(\frac{j\pi}{3}), N\sin(\frac{j\pi}{3})\right)\right\|_2^2.$$

As the synchrosqueezed energy distribution $T_f(a, \theta, x)$ is no longer valid around the crystal defects, we may characterize the defect region by using an indicator function generated by thresholding a smoothed version of the weighted boundary indicator function $\mathrm{BD}(x)$.

To better interpret the inverse deformation gradient G_0, we compute its polar decomposition $G_0(x) = U_0(x)P_0(x)$ for each point x, where $U_0(x)$ is a rotation matrix and $P_0(x)$ is a positive-semidefinite symmetric matrix. The rotation angle of $U_0(x)$ describes the crystal orientation at x;

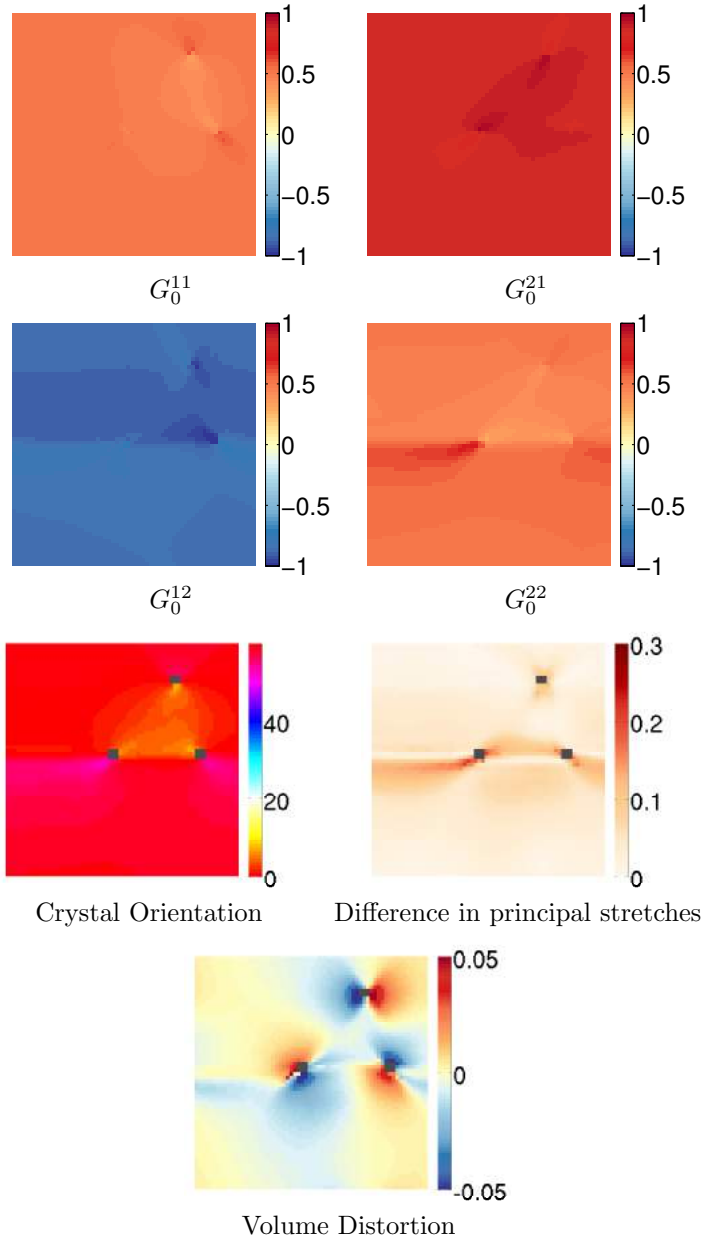


Figure 6.9: Top panel: Estimated inverse deformation gradient $G_0 \in \mathbb{R}^{2\times 2}$ of the atomic crystal image in Figure 6.6 (right). Bottom panel: The crystal orientation, the difference in principal stretches, and the volume distortion of G_0. The grey mask in these figures is the defect region identified by thresholding the smoothed boundary indication function.

$\mathrm{Vol}(x) = \det(G_0(x)) - 1$ indicates the volume distortion of $G_0(x)$; the quantity $|\lambda_1(x) - \lambda_2(x)|$, where $\lambda_1(x)$ and $\lambda_2(x)$ are the eigenvalues of $P_0(x)$, characterizes the difference in the principal stretches

of $G_0(x)$ as a measure of shear strength. The bottom panel of Figure 6.9 shows these quantities corresponding to the estimate of G_0 in the top panel. In the later numerical examples, instead of G itself we will always present the crystal orientation, the volume distortion, and the difference in principal stretches.

In the presence of a lattice distortion of a dislocation, a Burgers vector is introduce to represent this distortion. On one side of a Burgers vector, the space between atoms is slightly compressed, while the space on the other side is expanded. Hence, the local volume distortion would be positive on one side of the Burgers vector and negative on the other side. Figure 6.9 (bottom) shows the local distortion volume estimate $\mathrm{Vol}(x)$ of the example in Figure 6.6 (right). $\mathrm{Vol}(x)$ reaches its local maximum and local minimum near each point dislocation and the direction of the corresponding Burgers vector can be directly read off from the color coding of $\mathrm{Vol}(x)$.

6.1.4 Variational Model to Retrieve Deformation Gradient

We have shown that the synchrosqueezed transform is able to provide good mesoscopic information of atomic crystal images. This information is obtained point-wise directly from the synchrosqueezed energy distribution. In what follows, we will discuss a variational model based on continuous physical properties of crystal materials to improve the results provided by the synchrosqueezed transform. In particular, we retrieve the inverse deformation gradient G_0 to obtain a new deformation gradient G that better agrees with physical system. Before introducing the optimization model, we are going to summarize a few properties of atomic crystal.

Elastic energy in grain interior

Assume the image domain is Ω and the defect region is given by $\Omega_d \subset \Omega$. We expect the displacement field G to minimize the elastic energy of the system outside the defect region, since the system under imaging is in a quasistatic state. Given G_0 a rough guess of the deformation gradient, this motivates the energy minimization

$$\min_G \int_{\Omega\setminus\Omega_d} |G - G_0|^2 + W(G)\,\mathrm{d}x \tag{6.6}$$

where $|\cdot|$ denotes the Frobenius norm of a matrix, $|A| = (\mathrm{tr}(A^{\mathcal{T}}A))^{1/2}$, and W is the elastic stored energy density.

Since our reference lattice represents the undeformed equilibrium state of the crystal and the atom configuration in the image is produced by the (local) deformation ϕ^{-1}, the stored elastic energy can be expressed in the standard Lagrangian form as the integral over the reference domain $\phi(\Omega \setminus \Omega_d)$ of an elastic energy density w that depends on $\nabla(\phi^{-1}) = G^{-1} \circ \phi^{-1}$,

$$\int_{\phi(\Omega\setminus\Omega_d)} w(G^{-1} \circ \phi^{-1}(y))\ \mathrm{d}y\,.$$

Here, w satisfies the standard conditions coming from first principles, i. e. w is frame indifferent, $w(A) = 0$ for $A \in SO(2)$, $w(A) > 0$ else, and $w(A) = \infty$ if $\det A \leq 0$. After a change of variables the elastic energy turns into

$$\int_{\Omega \setminus \Omega_d} W(G) \, \mathrm{d}x$$

for $W(G) = w(G^{-1}) \det G$, where it is easy to see that W has the same above properties as w. For w (or equivalently W) one can use a material-specific, possibly anisotropic energy density. To be specific, since our numerical examples are all concerned with a triangular lattice exhibiting isotropic elastic behavior, we here simply restrict ourselves to the following neo-Hookean-type elastic energy density

$$\begin{aligned} W(G) = \frac{\mu}{2}(|G|^2 - 2) + \big(\frac{\mu}{2} + \frac{\lambda}{2}\big)(\det G - 1)^2 \\ - \mu(\det G - 1). \end{aligned} \tag{6.7}$$

Note that in (6.6), the fidelity and elastic energy terms are both evaluated outside the defect region. Within the defect region, since it is not possible to map the local configuration of atoms back to the reference state, the estimate G_0 is not trustworthy. It is also well known that the elastic energy blows up logarithmically approaching the dislocation core, and hence it only makes sense to penalize the elastic energy away from the defects.

Burgers vectors and $\operatorname{curl} G$

As explained previously, away from defects, G can be interpreted as the gradient $\nabla\phi$ of an inverse deformation ϕ deforming the configuration of the given image into a perfect reference crystal of a fixed orientation. Now the fact that gradient fields are always curl-free can be exploited as a constraint

$$\operatorname{curl} G = \begin{pmatrix} \partial_{x_1} G_{12} - \partial_{x_2} G_{11} \\ \partial_{x_1} G_{22} - \partial_{x_2} G_{21} \end{pmatrix} = 0 \quad \text{on } \Omega \setminus \Omega_d, \tag{6.8}$$

where Ω and Ω_d denote the image domain and the defect region, respectively. In the defect region Ω_d, however, the interpretation of G as a deformation gradient breaks down since there is no smooth deformation of the crystal that can undo the lattice defect. In fact, denoting the connected components of Ω_d by $\Omega_d^1, \ldots, \Omega_d^l$, it is shown in [64] that the integral

$$B_i = \int_{\Omega_d^i} \operatorname{curl} G \, \mathrm{d}x \tag{6.9}$$

is related to the Burgers vectors associated with the defects in Ω_d^i. If Ω_d^i contains an isolated dislocation surrounded by a regular lattice, B_i just represents the Burgers vector of that dislocation. If Ω_d^i contains multiple dislocations or even a section of a high angle grain boundary, B_i represents the accumulated Burgers vector of all defects in Ω_d^i (note that a high angle grain boundary may

be thought of as a string of dislocations with distance smaller than the lattice spacing so that the single dislocations are not clearly spatially separated). As in the case of $\Omega \setminus \Omega_d$, where we know $\operatorname{curl} G = 0$, we also have a priori information on $\operatorname{curl} G$ in Ω_d^i. In particular we know that B_i is a Burgers vector and thus must lie in the discrete set of Bravais lattice vectors of the perfect reference lattice $\mathcal{L}$. We thus identify B_i by projecting the (potentially noisy) estimate $\int_{\Omega_d^i} \operatorname{curl} G_0 \, \mathrm{d}x$ onto $\mathcal{L}$ and then impose (6.9) as a constraint on G. In fact, instead of prescribing the accumulated curl in Ω_d^i via (6.9) we may just as well prescribe

$$\operatorname{curl} G = b_i \quad \text{on } \Omega_d^i \tag{6.10}$$

for a function $b_i : \Omega_d^i \to \mathbb{R}^2$ with $\int_{\Omega_d^i} b_i \, \mathrm{d}x = B_i$ (in mathematically more precise terms, b_i may be a distribution). This is possible since we are only interested in the field G outside of Ω_d and since any field $G : \Omega \to \mathbb{R}^{2\times 2}$ satisfying (6.8) and (6.9) can be modified on Ω_d to a field satisfying (6.10). The function b_i is here simply chosen as $b_i = \operatorname{diag}(\alpha, \beta) \operatorname{curl} G_0$ with $\alpha, \beta \in \mathbb{R}$ such that $\int_{\Omega_d^i} b_i \, \mathrm{d}x = B_i$ (*i.e.*, an overall scaling of $\operatorname{curl} G_0$). Summarizing, a potential constrain in our variational method to extract G from the initial guess G_0 we will prescribe the constraint

$$\operatorname{curl} G = b \qquad \text{for} \quad b = \begin{cases} 0 & \text{on } \Omega \setminus \Omega_d; \\ b_i & \text{on } \Omega_d^i \,. \end{cases} \tag{6.11}$$

Refined defect regions

On the one hand, the threshold to identify Ω_d should be chosen very low to yield thin and localized defect regions (e. g. such that defect regions Ω_d^i around single dislocations stay separated from each other), on the other hand, the thinner the identified defect region Ω_d^i the worse will the estimate of the Burgers vectors B_i be. A compromise is to first use thick defect regions $\tilde{\Omega}_d^i$ in order to estimate the Burgers vectors B_i and then to impose the constraint (6.11) with a much finer estimate of the Ω_d^i. However, it may happen that a thick patch $\tilde{\Omega}_d^i$ contains multiple thin connected components $\Omega_d^{i_1}, \ldots, \Omega_d^{i_k}$. In that case the B_{i_j} are defined as the closest projections of $\int_{\Omega_d^{i_j}} \operatorname{curl} G_0 \, \mathrm{d}x$ onto $\mathcal{L}$ under the constraint $B_{i_1} + \ldots + B_{i_k} = \tilde{B}_i$, where $\tilde{B}_i$ is the accumulated Burgers vector of the patch $\tilde{\Omega}_d^i$. In order to obtain a very thin and localized Ω_d we simply identify the ridge of $3 - \operatorname{mass}(b)$ inside the thick $\tilde{\Omega}_d$ and then dilate this ridge by a few pixels.

Introduction of jump sets for topological consistency

A Bravais lattice does typically not only exhibit translational, but also rotational symmetry. The so-called point group $P \subset SO(2)$ comprises all those rotations which leave the reference lattice invariant. This leads to an ambiguity in the deformation gradient G: if G correctly describes the local configuration of the crystal, then RG for any $R \in P$ does so as well. Even though the constraint

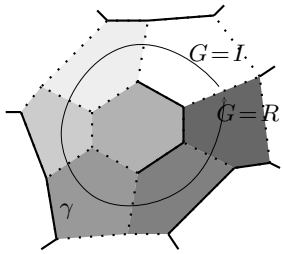


Figure 6.10: Along a closed path γ traversing a sequence of crystal grains, the deformation gradient G changes continuously from I to $R \neq I$. The gray shade indicates the local crystal orientation from the identity I (white) to R (dark gray). Dots represent point dislocations; lines indicate high angle grain boundaries. Along the path γ all grains are connected by low angle grain boundaries.

(6.11) has the effect that the matrix field G will be locally consistent (in the sense that for any y in a neighborhood of $x \in \Omega$ we have $|G(y) - G(x)| = \min_{R \in P} |RG(y) - G(x)|$), global consistency is often not guaranteed. Indeed, Figure 6.10 shows a situation in which along a closed path $\gamma \subset \Omega$, G changes continuously from the identity I to an element $R \neq I$ of the point group. Since R describes the same local crystal configuration as I, the curl where G jumps from R to I is spurious. As in [64], this inconsistency can be remedied by introducing a cut set S across which G is allowed to jump by a point group element,

$$G^- = RG^+ \text{ for some } R \in P\,,$$

where G^- and G^+ denote the value of G on either side of S. Henceforward, we consider the constrain

$$\operatorname{curl} G = b \text{ on } \Omega \setminus S\,, \quad G^-(G^+)^{-1} \in P \text{ on } S \tag{6.12}$$

instead of (6.8)

To find a good cut set S, let $Q \neq I$ denote the element of P closest to I. For each point $x \in \Omega$ we compute $G_i(x)$, $i = 1, 2, 3$, according to $G_i(x) = R_i(x)G_0(x)$ with $R_i(x) = \operatorname{argmin}_{R \in P} |RG_0(x) - Q^{\frac{i}{3}}|$. Here, $Q^{\frac{i}{3}}$ denotes rotation by $\frac{i}{3}$ of the rotation angle of Q. Of all three matrix fields we compute the curl and identify the regions where each of the G_i has the least curl. The boundaries between those regions are chosen as S, and in each region we reinitialize G_0 as that G_i with the least curl. Since the spurious curl at each point can occur at most for one of the G_i, the new initialization does not exhibit any spurious curl. Note that we also take care that $\Omega_d \cap S = \emptyset$ so that the estimation of Burgers vectors is not impaired.

Constrained minimization model

Combining the discussion for 6.6 and (6.12), now we are ready to describe the variational model to retrieve the initial guess G_0 for a new G that better agrees with physical meaning:

$$\min_{G:\Omega\to\mathbb{R}^{2\times 2}} \int_{\Omega\setminus\Omega_d} |G - G_0|^2 + W(G)\,\mathrm{d}x \qquad (6.13)$$
$$\text{s. t. } \operatorname{curl} G = b \text{ on } \Omega \setminus S\,, \quad G^-(G^+)^{-1} \in P \text{ on } S\,,$$

where G^- and G^+ denote the value of G on either side of S. It is convenient to describe the numerical algorithm to solve this optimization in discretized form, even though all objects have a continuous equivalent. Throughout, a superscript d denotes discrete differential operators.

In the discrete setting, the image domain Ω is discretized via $M \cdot N$ Cartesian pixels, indexed by x-y-position $(m,n) \in \Omega = \{1,\dots,M\} \times \{1,\dots,N\}$ (the pixel spacing is assumed to be one). For an index m we denote by $m^+ = m+1$ and $m^- = m-1$ the next larger or next smaller index, where for simplicity we assume periodic boundary conditions and use cyclic indexing, i. e. $(M^+,n) = (1,n)$, $(m,N^+) = (m,1)$, $(1^-,n) = (M,n)$, $(m,1^-) = (m,N)$ for all $(m,n) \in \Omega$. The matrix fields are discretized accordingly, $G, G_0 \in (\mathbb{R}^{2\times 2})^{M\times N}$. The jump set S follows the edges between the pixels and is represented as a collection of horizontal or vertical pixel pairs, $S \subset \{((m,n),(k,l)) \in \Omega\times\Omega \;:\; (k,l) = (m^+,n) \text{ or } (k,l) = (m,n^+)\}$. Furthermore, we define the function $R : S \to P$ such that $R_{(m,n),(k,l)}$ is the point group element with smallest distance to $(G_0)_{m,n}(G_0)^{-1}_{k,l}$. R is extended to $\Omega\times\Omega$ by the identity in P. Derivatives in x- and y-direction are replaced by finite differences that respect the point group equivalence across S,

$$(\partial_x^d G)^{ij}_{m,n} = (R_{(m,n),(m^+,n)} G_{m^+,n})^{ij} - G^{ij}_{m,n}\,,$$
$$(\partial_y^d G)^{ij}_{m,n} = (R_{(m,n),(m,n^+)} G_{m,n^+})^{ij} - G^{ij}_{m,n}\,,$$

where superscript ij denotes the (i,j)-matrix entry. In particular, the discrete curl and Laplacian are defined as

$$\operatorname{curl}^d : (\mathbb{R}^{2\times 2})^{M\times N} \to (\mathbb{R}^2)^{M\times N}\,,$$
$$(\operatorname{curl}^d G)_{m,n} = \partial_x^d G^{:2}_{m,n} - \partial_y^d G^{:1}_{m,n}\,,$$
$$\Delta^d : (\mathbb{R}^2)^{M\times N} \to (\mathbb{R}^2)^{M\times N}\,,$$
$$\Delta^d = -\operatorname{curl}^d (\operatorname{curl}^d)^*\,,$$

where $G^{:i}_{m,n}$ denotes the i^{th} column of the matrix $G_{m,n}$ and the superscript $*$ denotes the adjoint

operator, which in this particular case is given by

$$
\begin{aligned}
&(\operatorname{curl}^d)^* : (\mathbb{R}^2)^{M\times N} \to (\mathbb{R}^{2\times 2})^{M\times N}\,,\\
&((\operatorname{curl}^d)^* V)_{m,n} =\\
&\left[V_{m,n} - R^T_{(m,n^-),(m,n)} V_{m,n^-} \,\middle|\, R^T_{(m^-,n),(m,n)} V_{m^-,n} - V_{m,n} \right] .
\end{aligned}
$$

In sum, after discretization, our optimization problem reads

$$
\begin{gathered}
\min_{G\in C_b} E[G]\\
\text{for } E[G] = \textstyle\sum_{(m,n)\in\Omega} \left(|G_{m,n} - (G_0)_{m,n}|^2 + W(G_{m,n}) \right)\\
\text{and } C_b = \{ G \in (\mathbb{R}^{2\times 2})^{M\times N} \;:\; \operatorname{curl}^d G = b \}\,.
\end{gathered}
$$

Fast algorithm for the constrained minimization

In the constrained minimization, the constraint space C_b is an affine space and can be expressed as $\hat{G} + C_0$ for a $\hat{G}$ with $\operatorname{curl}^d \hat{G} = b$. Hence, the energy can be minimized using a standard projected nonlinear conjugate gradient (NCG) descent in C_b. In more detail, we employ a Fletcher–Reeves NCG method in which the derivative $\partial^d_G E$ of E with respect to G is always orthogonally projected onto C_0 (i. e. onto its component parallel to C_b) so that the algorithm is performed within the subspace C_b. Due to accumulating numerical errors we also have to project the current estimate G back onto C_b from time to time. Denoting the projection onto C_b by $\operatorname{proj}_{C_b}$, the NCG algorithm is initialized with $\hat{G} = \operatorname{proj}_{C_b} G_0$.

The projection $\operatorname{proj}_{C_b} F$ is the solution to the constraint minimization $\min_{\operatorname{curl}^d G = b} \sum_{m,n} |G_{m,n} - F_{m,n}|^2$, which satisfies the optimality conditions

$$
b = \operatorname{curl}^d F\,, \qquad 0 = F - G + (\operatorname{curl}^d)^* \Lambda
$$

for a Lagrange multiplier $\Lambda \in (\mathbb{R}^2)^{M\times N}$. Applying curl^d to the second equation we obtain

$$
\operatorname{curl}^d G - b = -\Delta^d \Lambda\,.
$$

Note that $\ker(\Delta^d) \perp \operatorname{range}(\operatorname{curl}^d)$. Denoting by $(-\Delta^d)^{-1} : \operatorname{range}(\Delta^d) \to \ker(\Delta^d)^\perp$ the inverse of $-\Delta^d$, we obtain $\Lambda = (-\Delta^d)^{-1}(\operatorname{curl}^d G - b)$ and thus

$$
\operatorname{proj}_{C_b} G = F = G - (\operatorname{curl}^d)^* (-\Delta^d)^{-1} (\operatorname{curl}^d G - b)\,.
$$

Once $\hat{G}$ is computed, the projection onto C_b can also be obtained as

$$
\operatorname{proj}_{C_b} F = \hat{G} + \operatorname{proj}_{C_0}(F - \hat{G})\,.
$$

For the above projection we need to invert the discrete Laplacian operator. Using periodic boundary conditions this would be very fast using FFT if there was no jump set S. However, with nonempty jump set S, our finite difference operators do not turn into pointwise multiplications in Fourier space. In order to obtain a fast inversion we decompose $\Delta^d = \Delta_0^d + J$, where Δ_0^d is the standard discrete Laplacian and J the linear operator accounting for the point group,

$$\begin{aligned}\Delta_0^d &: (\mathbb{R}^2)^{M\times N} \to (\mathbb{R}^2)^{M\times N}\,,\\ (\Delta_0^d V)_{m,n} &= V_{m^-,n} + V_{m^+,n} + V_{m,n^-} + V_{m,n^+} - 4V_{m,n}\,,\end{aligned}$$

$$\begin{aligned}J &: (\mathbb{R}^2)^{M\times N} \to (\mathbb{R}^2)^{M\times N}\,,\\ (JV)_{m,n} &= (R^T_{(m^-,n),(m,n)} - I)V_{m^-,n}\\ &\quad + (R_{(m,n),(m^+,n)} - I)V_{m^+,n}\\ &\quad + (R^T_{(m,n^-),(m,n)} - I)V_{m,n^-}\\ &\quad + (R_{(m,n),(m,n^+)} - I)V_{m,n^+}\,.\end{aligned}$$

Note that J is symmetric, and it is highly sparse and thus has a very small range $L = \text{range} J$ and a large kernel $\ker J = \ker J^{\mathcal{T}} = L^\perp$. If we decompose $V \in (\mathbb{R}^2)^{M\times N}$ into

$$V = V_L + V_{L^\perp} \in L \oplus L^\perp\,,$$

then we obtain

$$-\Delta^d V = B \qquad \Leftrightarrow \qquad -\Delta_0^d V_{L^\perp} = (\Delta_0^d + J)V_L + B\,. \tag{6.14}$$

The solvability condition tells us that $\ker \Delta_0^d$ is orthogonal to the right-hand side, so we can just as well project the right-hand side onto $(\ker \Delta_0^d)^\perp$ by subtracting the mean,

$$-\Delta_0^d V_{L^\perp} = \text{proj}_{(\ker \Delta_0^d)^\perp}\left((\Delta_0^d + J)V_L + B\right)\,.$$

Now choosing a basis $L = \langle v_1, \ldots, v_K\rangle$ and $\ker \Delta_0^d = \langle s_1, s_2\rangle$, we write

$$V_L = \sum_{i=1}^K \lambda_i v_i\,,$$

$$\begin{aligned} V_{L^\perp} &= a_1 s_1 + a_2 s_2 + (-\Delta_0^d)^{-1}\mathrm{proj}_{(\ker \Delta_0^d)^\perp}\left((\Delta_0^d + J)V_L + B\right) \\ &= a_1 s_1 + a_2 s_2 + (-\Delta_0^d)^{-1}\mathrm{proj}_{(\ker \Delta_0^d)^\perp} B \\ &\quad + \sum_{i=1}^{K} \lambda_i \left((-\Delta_0^d)^{-1}\mathrm{proj}_{(\ker \Delta_0^d)^\perp} J v_i - v_i\right) , \end{aligned}$$

where $(-\Delta_0^d)^{-1} : (\ker \Delta_0^d)^\perp \to (\ker \Delta_0^d)^\perp$ denotes the inverse of $-\Delta_0^d$. The degrees of freedom $\lambda_1, \ldots, \lambda_K, a_1, a_2$ now have to satisfy the $K+2$ equations $v_j \cdot V_{L^\perp} = 0$, $j = 1, \ldots, K$, (due to $v_j \in L$) and $s_j \cdot \left((\Delta_0^d + J)V_L + B\right) = 0$, $j = 1, 2$, (the solvability condition for (6.14)) where the dot denotes the dot product. In detail, the equations are given by

$$\begin{aligned} 0 &= a_1 v_j \cdot s_1 + a_2 v_j \cdot s_2 + v_j \cdot (-\Delta_0^d)^{-1}\mathrm{proj}_{(\ker \Delta_0^d)^\perp} B \\ &\quad + \sum_{i=1}^{K} \lambda_i \left(\left((-\Delta_0^d)^{-1}\mathrm{proj}_{(\ker \Delta_0^d)^\perp} v_j\right) \cdot J v_i - v_j \cdot v_i\right) , \\ 0 &= s_j \cdot B + \sum_{i=1}^{K} \lambda_i s_j \cdot J v_i \, . \end{aligned}$$

To solve for $\lambda := (\lambda_1, \ldots, \lambda_K, a_1, a_2)^T$, we write the linear system as a matrix vector equation. Since we need to solve systems of the form $-\Delta^d V = B$ several times for different values of B, we perform a QR decomposition,

$$QRP\lambda = \mathrm{rhs} \, ,$$

where P is a permutation matrix such that the lower rows of R are zero (this is possible since the solution to $-\Delta^d V = B$ is only uniquely specified up to a two-dimensional subspace due to $\dim \ker \Delta^d = 2$) and all other rows have nonzero diagonal elements. Note that this decomposition is the bottleneck of the algorithm with a complexity of $O(K^3)$. The degrees of freedom corresponding to the zero-rows can now be chosen freely (say, as zero), and the resulting

$$\begin{aligned} V &= a_1 s_1 + a_2 s_2 + (-\Delta_0^d)^{-1}\mathrm{proj}_{(\ker \Delta_0^d)^\perp} B \\ &\quad + \sum_{i=1}^{K} \lambda_i (-\Delta_0^d)^{-1}\mathrm{proj}_{(\ker \Delta_0^d)^\perp} J v_i \\ &= a_1 s_1 + a_2 s_2 + (-\Delta_0^d)^{-1}\mathrm{proj}_{(\ker \Delta_0^d)^\perp} \left[\sum_{i=1}^{K} \lambda_i J v_i + B\right] \end{aligned}$$

solves $B = -\Delta^d V$. Note that $(-\Delta_0^d)^{-1}$ can readily be computed via FFT and that the v_i can be chosen as shifted versions of $\hat{v}^j \in (\mathbb{R}^2)^{M \times N}$, $j = 1, 2$, with $\hat{v}^j = 0$ except for the j^{th} entry of $\hat{v}^j_{1,1}$ being one. Thus, $(-\Delta_0^d)^{-1}\mathrm{proj}_{(\ker \Delta_0^d)^\perp} v_i$ can be evaluated efficiently as just a shift of $(-\Delta_0^d)^{-1}\mathrm{proj}_{(\ker \Delta_0^d)^\perp} \hat{v}^j$.

6.1.5 Examples and Discussions

In this section, there will be a series of experiments of synthetic and real images to illustrate the performance of Algorithm 6.1.7, 6.1.8 and the variational model. In the first part of this section, we focus on the application of Algorithm 6.1.7 to detect grain boundaries and to estimate crystal rotations. The robustness of this method will be emphasized and supported by some noisy examples and examples with blurry grain boundaries. In the second part, Algorithm 6.1.8 is applied to obtain initial results of crystal analysis followed by the variational model to retrieve the initial results. Algorithm 6.1.8 is more sensitive to local defects and is able to discover Burgers vectors of these dislocations by estimating the local volume distortion Vol(b).

The main parameters for the SST-based analysis are two geometric scaling parameters s and t in the 2D SST (for details see [183]). Smaller scaling parameters result in better robustness of SSTs while larger scaling parameters give more accurate estimates in noiseless cases [183]. Hence, we adopt $t = s \approx 1$ in the examples with less noise and use $t = s \approx 0.8$ when the crystal image is noisy. As discussed in [183], for images with heavy noise, the synchrosqueezed transform can still provide reasonable initial guess via a highly redundant transform with more computational cost. The variational model parameters λ and μ in (6.7) are simply set to 1 in all of the following examples.

The synchrosqueezed transform is very efficient. In the first part, all numerical results by Algorithm 6.1.7 are obtained within 5 seconds. In the second part, the SST-based method can downsample the original image in the frequency domain and reduce the mesh size for the variational optimization. Hence, the computational mesh of the optimization model is independent of the size of crystal images. The main computational cost comes from the inverse of the Laplacian operator. This cost is depending on the number of points of the singularity set S. By reducing the mesh size, the number of points in S can be reduced and the computational cost is significantly reduced, e.g., from 13 minutes to 20 seconds in the example of PFC image of size 1024×1024. The runtime for the first two real examples is less than 10 seconds. The runtime for the last example with heavy noise is about 1 minute due to extra effort to obtain robustness of the synchrosqueezed transform.

Examples for Algorithm 6.1.7

Our first example is a phase field crystal (PFC) image in Figure 6.11 (left). It contains several grains with low and high angle grain boundaries and some point dislocations. As shown in Figure 6.11 (middle), the weighted average angle Angle(b) is changing gradually in the interior of a grain and sharp at a large angle boundary. Figure 6.11 (right) shows the boundary indicator function BD(b). Large angle grain boundaries appear in a form of line segments. Adjacent point dislocations are connected and identified as grain boundaries, while light grey regions identify well-separated point dislocations. As pointed out later, some isolated point dislocation may be missed due to the stacking step, while Algorithm 6.1.8 is better suited for detecting local defects.

We next consider a real data example of a twin boundary in a TEM-image in GaN (see Figure

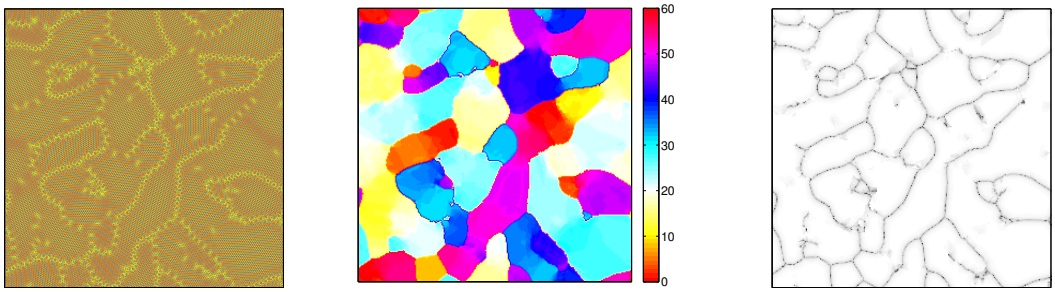


Figure 6.11: Analysis results of a large phase field crystal image (of size 1024×1024 pixels) provided by Algorithm 6.1.7. Left: A phase field crystal (PFC) image and its zoomed-in image. Courtesy of Benedikt Wirth [64]. Middle: The weighted average angle Angle(b) and its zoomed-in result. Right: The boundary indicator function BD(b) and its zoomed-in result.

6.12 (left)). Algorithm 6.1.7 identifies two grains as shown in Figure 6.12 (middle) and estimates their rotation angles. The grain boundary is approximated by a smooth curve in the image of the boundary indicator function BD(b) in Figure 6.12 (right).

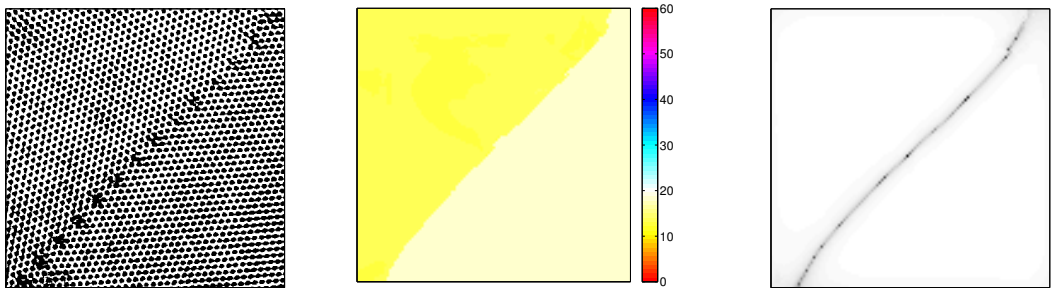


Figure 6.12: Left: A TEM-image of size 420×444 pixels in GaN. Courtesy of David M. Tricker (Department of material science and metallurgy, University of Cambridge). Middle: The weighted average angle Angle(b) provided by Algorithm 6.1.7. Right: The boundary indicator function BD(b) provided by Algorithm 6.1.7.

Figure 6.13 shows an example of wide boundaries in a photograph of a bubble raft. In this case, the transition between two grains is not sharp due to large distortion and the local crystal structure is heavily disturbed near the boundary. Nevertheless, Algorithm 6.1.7 is capable of identifying three grains with sharp grain boundaries matching the distortion area as shown in Figure 6.13 (middle) and 6.13 (right).

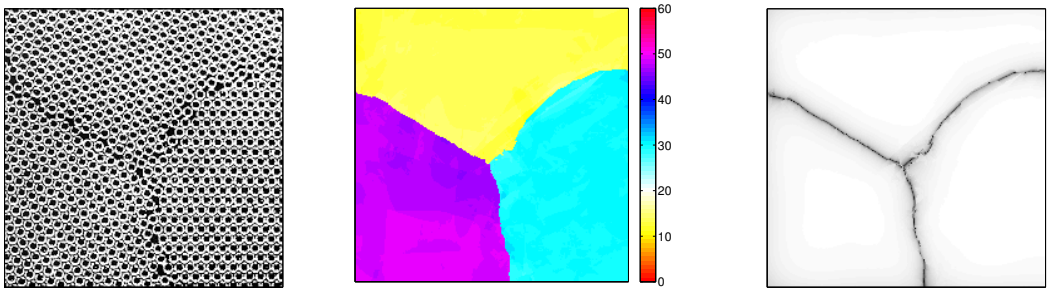


Figure 6.13: Left: A photograph of a bubble raft with large disorders and wide boundaries. Courtesy to Barrie S. H. Royce in Princeton University. The image size is 223×415. Middle and right: The weighted average angle Angle(b) and the boundary indicator function provided by Algorithm 6.1.7.

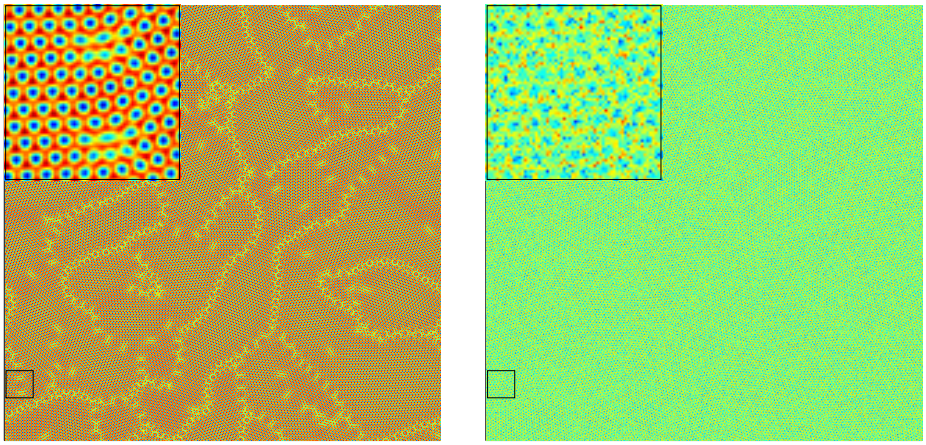


Figure 6.14: A noise-free PFC image (left) and its noisy version (right) with a zoom-in detailing the marked part.

Examples for Algorithm 6.1.8 **and the variational model**

In the examples in this part, we compare the initial estimated strain G_0 provided by Algorithm 6.1.8 and the improved results G from the variational method (6.13), where each time we display crystal orientations, difference in principal stretches, and volume distortion. For better visualization we mask out the identified defect regions, which is generated by thresholding a smoothed version of boundary indicator function BD(x) from Algorithm 6.1.8, since there is no meaningful notion of strain in these regions. The curl of G_0 (which in general violates the physical constraint of being zero outside Ω_d and of being compatible with the defects' Burgers vectors) as well as the average $\operatorname{curl} G$ per connected defect region (which is compatible with the defects' Burgers vectors) are also shown.

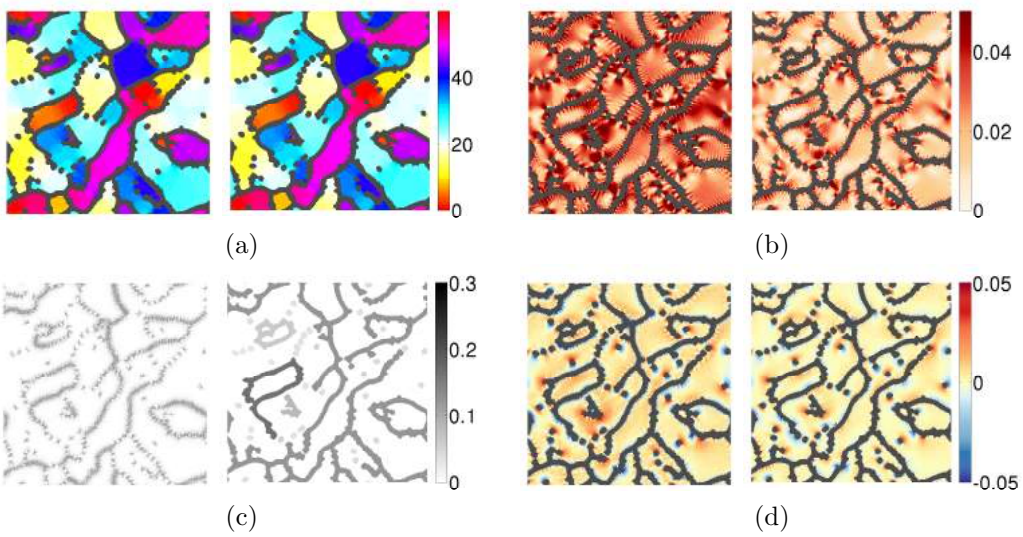


Figure 6.15: Results of the image in Figure 6.14 (left). From panel (a) to (d): crystal orientation, difference in principal stretches, curl of the inverse deformation gradient, and volume distortion. In each panel, the left figure shows the initial results from SST and the right one shows the optimized results from the variational method. Particularly in (c)-right, the average curl on each connected defect region is shown.

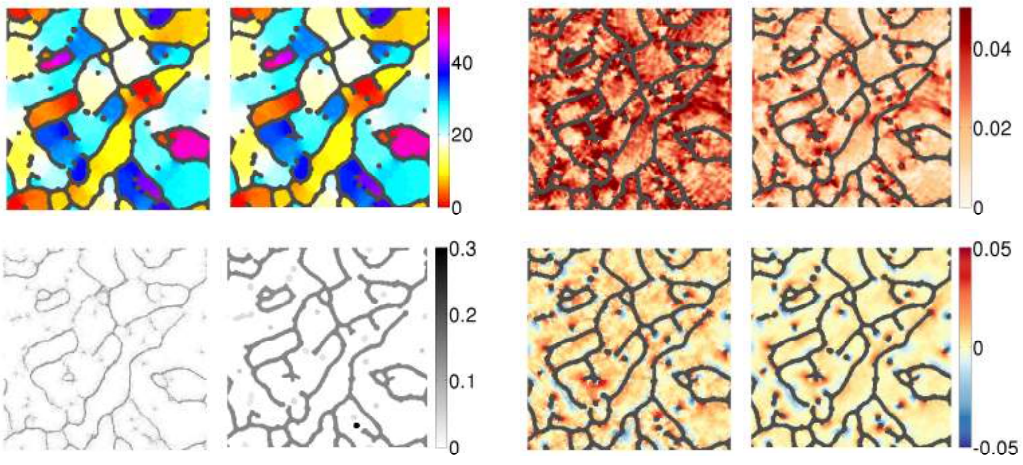


Figure 6.16: Results of Figure 6.14 (right) using the same visualization as in Figure 6.15.

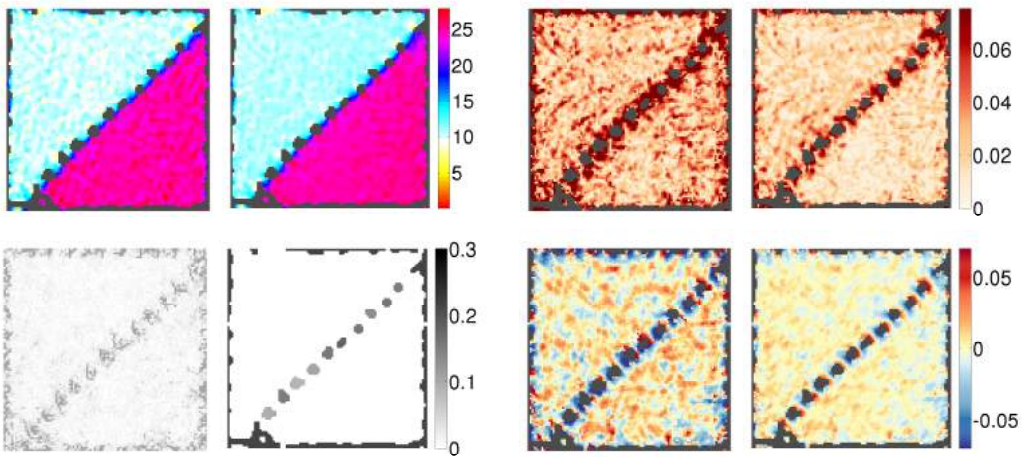


Figure 6.17: Results of crystal analysis for Figure 6.12 (left) and, using the same visualization as in Figure 6.15.

Since the PFC model is a well-established method to simulate elastic and plastic deformations, free surfaces, and multiple crystal orientations in nonequilibrium processes. We revisit it again to test Algorithm 6.1.8 and the variational model. A noiseless PFC image is given in Figure 6.14 (left) and its results are presented in Figure 6.15. Compared to the initial strain estimate G_0, the optimized G is smoother, exhibits a much smaller overall volume distortion and shear (as visualized by the difference in principal stretches), and sharpens the compression-dilation dipoles around each single dislocation.

The noisy PFC image of Figure 6.14 (right) is generated by adding 50% Gaussian random noise. Obviously, this leads to strong artifacts in the estimated deformation G_0. Remarkably, after the optimization we retrieve an estimate G almost as good as in the noiseless case as shown in Figure 6.16, which demonstrates the robustness of our method.

We also revisit other two real examples in the first part of this section. The TEM-image of GaN (Figure 6.12 left), contains a string of dislocations forming a large angle grain boundary. The artificial strong spatial variation of crystal orientation, shear and volume distortion in the SST result is greatly reduced after applying the optimization as shown in Figure 6.17. Even more importantly, the unphysical curl away from the defects is completely removed such that the curl of G is fully concentrated in the single defect regions around each dislocation (the total curl of each region equaling the dislocation's Burgers vector).

The photograph of a bubble raft with strongly disordered and blurry grain boundaries has been shown in Figure 6.13 (right). The result of Algorithm 6.1.8 (see Figure 6.18) shows a spurious strong shear of the local crystal structure close to the grain boundaries, especially near the triple junction. One of the reasons for this behavior is that the SST, like any wavelet type transform, extracts

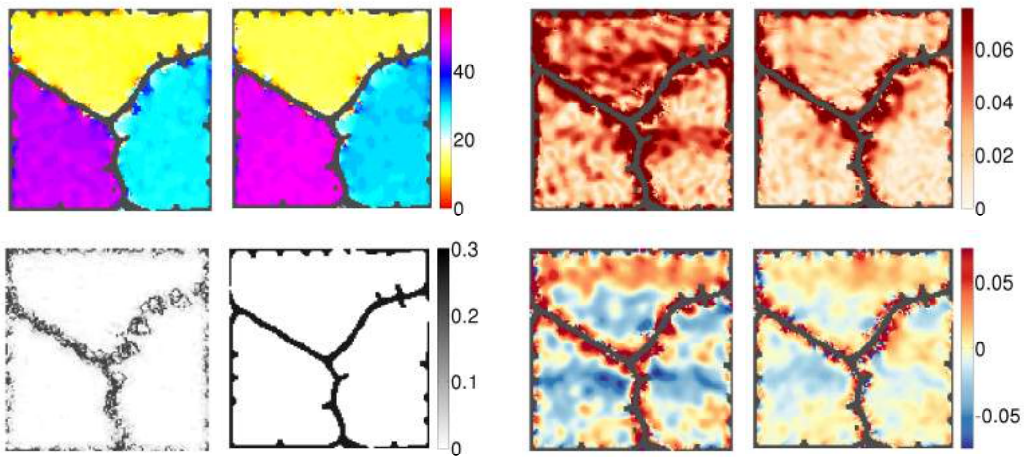


Figure 6.18: Results of crystal analysis for Figure 6.13 (left), using the same visualization as in Figure 6.15.

directional and strain information from image patches, which here have to be larger than the unit cells. Thus, the grain boundaries are diffused, and information near the grain boundary is not trustworthy. The optimization can mostly remedies this effect.

The last experimental example is a TEM-image of a twin and a high angle grain boundary in Al (Figure 6.19). The crystal structures in each grain are also slightly stretched. Although this example is very challenging, the Algorithm 6.1.8 can still provide an accurate defect region estimate and a reasonable initial guess G_0 (see Figure 6.20). After optimization, we obtain a curl-free inverse deformation gradient G in the grain interior. The difference of principal stretches becomes smaller and the volume distortion gets closer to zero outside the defect region.

6.1.6 Conclusion

This section has introduced a new model for atomic crystal images and several tools for their multiscale analysis. Through various synthetic and real data, it has been shown that the proposed methods are able to provide robust and reliable estimates of mesoscopic properties, e.g., crystal defects, rotations, elastic deformations and grain boundaries. Since these methods are well suitable for parallelization, parallel computing will considerably reduce the runtime. This would be appealing in the analysis of a series of large crystal images to study the time evolution of crystals on a microscopic length scale.

We focus on the analysis of images with the presence of only one type of crystal and without solid and liquid interfaces in this paper. The extension is not difficult. In fact, in the presence of liquid, the solid-liquid interface can be identified as “boundary between grains” by our method. One

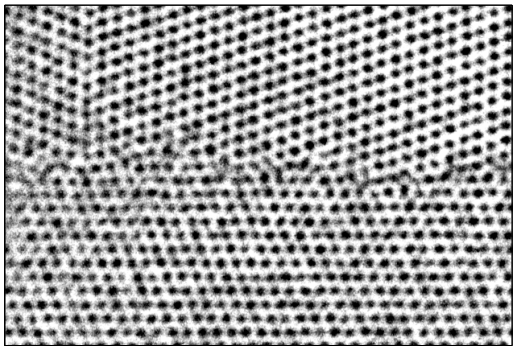


Figure 6.19: A TEM-image of Al (courtesy of the National Center for Electron Microscopy at the Lawrence Berkeley National Laboratory).

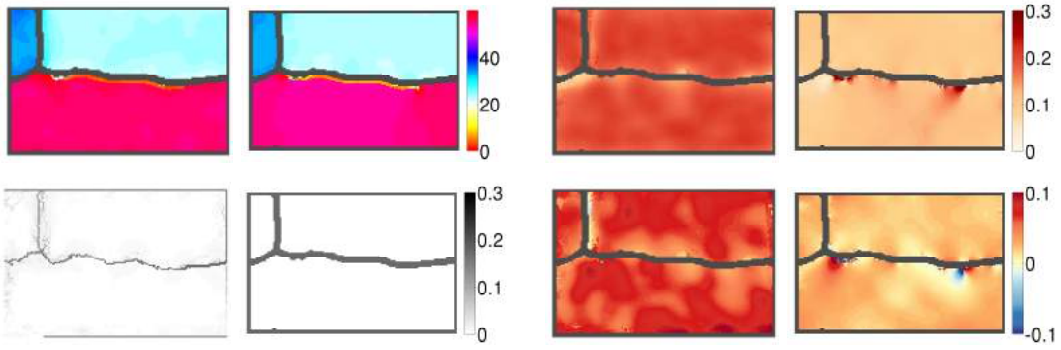


Figure 6.20: Results of crystal analysis for Figure 6.19, using the same visualization as in Figure 6.15.

could use imaging methods for detecting geometric objects in the cartoon part of images [18, 108] to identify the liquid part immediately after grains are identified by our method. Moreover, when the given image consists of multiple types of crystals, local Fourier transforms taken at enough sampling positions can identify reference crystals. Fixing one type of reference crystals, we apply our method to extract the boundaries, the rotations, the defects and the deformations of grains of this type. A complete analysis can be obtained by combining the results of each type of reference crystals.

Another interesting and challenging future direction is to analyze crystal images corresponding to complex lattices. First, it might be difficult to automatically identify reference crystals directly from a given image. The information hidden in the image is very redundant and hence feature extraction and dimension reduction techniques are necessary. Second, the well-separation condition for synchrosqueezed transforms may not hold due to a large number of underlying wave-like components of each grain. A $2D$ generalization of the $1D$ diffeomorphism based spectral analysis method in [178] may provide a solution to this problem.

The current methods can be easily extended to $3D$ crystal analysis by designing a $3D$ synchrosqueezed transform. This should be relevant for applications.

6.2 Canvas Weave Analysis in Art Forensics

6.2.1 Introduction

Quantitative canvas weave analysis has many applications in art investigations of paintings, including dating, forensics, canvas rollmate identification [110, 125, 160]. Traditionally, canvas analysis is based on X-radiograph. Prior to serving as a painting canvas, a piece of fabric is coated with a priming agent; smoothing its surface makes this layer thicker between and thinner right on top of weave threads. These variations affect the X-ray absorption, making the weave pattern stand out in X-ray images of the finished painting. To characterize this pattern, it is customary to visually inspect small areas within the X-radiograph and count the number of horizontal and vertical weave threads; averages of these then estimate the overall canvas weave density. The tedium of this process typically limits its practice to just a few sample regions of the canvas. In addition, it does not capture subtler information beyond weave density, such as thread angles or variations in the weave pattern. Application of signal processing techniques to art investigation are now increasingly used to develop computer-assisted canvas weave analysis tools.

In their pioneering work [100], Johnson *et al* developed an algorithm for canvas thread-counting based on windowed Fourier transforms (wFT); further developments in [101, 104] extract more information, such as thread angles and weave patterns. Successful applications to paintings of art historical interest include works by van Gogh [85, 164], Diego Velázquez [57], Johannes Vermeer [121], among others [102, 103, 105, 106, 107].

A more robust and automated analysis technique was later developed by Erdmann *et al* [68],

based on autocorrelation and pattern recognition algorithms, requiring less human intervention (e.g., choosing proper frequency range and window size of windowed Fourier transforms). Unlike the Fourier-space based approach of [100], [68] uses only the real-space representation of the canvas. Likewise, [41] also uses real-space based features for canvas texture characterization.

We consider here a new automated analysis technique for quantitative canvas analysis, based on the 2D synchrosqueezed transforms. Synchrosqueezing has shown to be a useful tool in problems presented previously. Observing that the canvas weave texture of interest consists of a sparse superposition of close to but not quite periodic template functions, it is natural to consider 2D SST as an alternative for canvas analysis; as illustrated by the results we obtained, reported here, this intuition proved to be correct. The method, as shown below, is very robust and offers fine scale weave density and thread angle information for the canvas. We compare our results with those in [68, 100, 101, 104].

We explain our model for X-radiography images for canvas analysis in Section 6.1.2; the use and limitations of windowed Fourier transforms are discussed in Section 6.2.3. Section 6.2.3 introduces the synchrosqueezed transform, with applications to quantitative canvas analysis; section 6.2.4 presents various examples, applying our technique in art investigation.

6.2.2 Model of the Canvas Weave Pattern in X-Radiography

We denote by f the intensity of an X-radiograph of a painting; see Figure 6.21a for a (zoomed-in) example. Because X-rays penetrate deeply, the image consists of several components: the paint layer itself, primer, canvas (if the painting is on canvas or on wood panel overlaid with canvas), possibly a wood panel (if the painting is on wood), and sometimes extra slats (stretchers for a painting on canvas, or a cradle for a painting on wood, thinned and cradled according to earlier conservation practice.) This X-ray image may be affected by noise or artifacts of the acquisition process. We model the intensity function f as an additive superposition of the canvas contribution, denoted by $c(x)$, and a remainder, denoted by $p(x)$, that incorporates all the other components. Our approach to quantitative canvas analysis relies on a simple model for the X-ray image of the weave pattern in the "ideal" situation. Since it is produced by the interleaving of horizontal and vertical threads in a periodic fashion, a natural general model is

$$f(x) = c(x) + p(x) := a(x)S(2\pi N\phi(x)) + p(x). \tag{6.15}$$

In this expression, S is a periodic function on the square $[0, 2\pi)^2$, the details of which reflect the basic weave pattern of the canvas, e.g., whether it is a plain weave or perhaps a twill weave. This is a generalization of more specific assumptions used in the literature – for instance, in [100] plain weave canvas is modeled by taking for S a sum of sinusoidal functions in the x and y directions; in [104], more general weave patterns (in particular twill) are considered. The parameter N in (6.15)

gives the averaged overall weave density of the canvas (in both directions). The function ϕ, which maps the image domain to $\mathbb{R}^2$, is a smooth deformation representing the local warping of the canvas; it contains information on local thread density, local thread angles, etc. The slowly varying function $a(x)$ accounts for variations of the amplitude of the X-ray image of the canvas, e.g., due to variation in illumination conditions.

In some cases, the X-ray image fails to show canvas information in portions of the painting (e.g. when the paint layer dominates); the model (6.15) is then not uniformly valid. Because our analysis uses spatially localized information (analyzing the image patch by patch), this affects our results only locally: in those (small) portions of the image we have no good estimates for the canvas parameters. For simplicity, this exposition assumes that (6.15) is valid for the whole image.

We rewrite c by representing the weave pattern function S, periodic on $[0, 2\pi)^2$, in terms of its Fourier series,

$$c(x) = \sum_{n\in\mathbb{Z}^2} a(x)\widehat{S}(n)e^{2\pi i N n\cdot\phi(x)}. \tag{6.16}$$

This is a superposition of smoothly warped plane-waves with local wave vectors $N\nabla(n\cdot\phi(x))$. The idea of our analysis is to extract the function ϕ by exploiting that the Fourier coefficients $\{\widehat{S}(n)\}$ are dominated by a few leading terms.

6.2.3 Fourier-Space Based Canvas Analysis

Windowed Fourier transform

Because a and ϕ vary slowly with x, we can use Taylor expansions to approximate the function for x near x_0 as

$$c(x) \approx \sum_{n\in\mathbb{Z}^2} a(x_0)\widehat{S}(n)e^{2\pi i N n\cdot\phi(x_0)}e^{2\pi i N(x-x_0)\cdot\nabla_x(n\cdot\phi)(x_0)}. \tag{6.17}$$

The right hand side of (6.17) is a superposition of complex exponentials with frequencies $w = (w_1, w_2)$, with

$$w_l = \sum_{l'=1}^{2} n_{l'}(\partial_l\phi_{l'})(x_0);$$

these would stand out in a Fourier transform as peaks in the 2D Fourier spectrum. Since the approximation is accurate only near x_0, we also use a windowed Fourier transform with envelope given by, e.g., a Gaussian centered at x_0 with width σ. We have then

$$\begin{aligned} W(x_0,k) &:= \frac{1}{2\pi\sigma^2}\iint e^{-2\pi ik(x-x_0)}e^{-(x-x_0)^2/2\sigma^2}c(x)\,\mathrm{d}^2x \\ &\approx \sum_{n\in\mathbb{Z}^2} a(x_0)\widehat{S}(n)e^{2\pi i N n\phi(x_0)}e^{-2\pi^2\sigma^2[k-N\nabla_x(n\cdot\phi)(x_0)]^2}. \end{aligned} \tag{6.18}$$

Instead of being sharply peaked, the spectrum of the windowed Fourier transform is thus "spread out" around the $N\nabla_x(n\cdot\phi)(x_0)$ – a manifestation of the well-known uncertainty principle in signal processing, with a trade-off w.r.t. the parameter σ: a larger σ reduces the "spreading" at the price of a larger error in the approximation (6.17), since the Gaussian is then correspondingly wider in the real space.

The method of [100, 104] uses the local maxima of the amplitude of the windowed Fourier transform to estimate the location of $\{N\nabla(n\cdot\phi(x_0))\}$ for a selection of positions x_0 of the X-ray image (local swatches are used instead of the Gaussian envelope, but the spirit is the same). For ideal signals, (6.18) shows that the maxima of the amplitude $|W(x_0,\cdot)|$ identify the dominating wave vectors in Fourier-space, which are then used to extract information, including weave density and thread angles. Thread density is estimated by the length of the wave vectors; the weave orientation is determined by the angles. This back-of-the-envelope calculation is fairly precise when N is much larger than 1, resulting in a small $\mathcal{O}(N^{-1})$ error in the Taylor expansions and stationary phase approximations. In terms of the canvas, $N\gg 1$ means that the inverse of the average thread density must be much smaller than the length scale of the variation of the canvas texture, which is typically on the scale of the size of the painting. This is essentially a high-frequency assumption, ensuring that stationary phase approximations can be applied in the time-frequency analysis. Details can be found in standard references of time-frequency analysis, e.g., the book [69].

In more complicated scenarios, in particular, when the X-ray signal corresponding to the canvas is heavily "contaminated" by the other parts of the painting, it is desirable to have more robust and refined analysis tools at hand than locating local maxima of the Fourier spectrum. The synchrosqueezed transforms are nonlinear time-frequency analysis tools developed for this purpose, in different (1D and 2D) applications which suggests they could be suitable for canvas analysis in challenging situations. A comparison of the two methods is shown in Figure 6.21 and will be explained below.

Synchrosqueezed transforms

The crucial observation is that the phase of the complex function $W(x,k)$, obtained from the windowed Fourier transform (6.18) contains information on the local frequency (i.e., local wave vectors) of the signal. Indeed, for (x,k) such that k is close to $N\nabla_x(n\cdot\phi)$, we have

$$w_f(x,k) := \frac{1}{2\pi}\Im\left(\nabla_x \ln W(x,k)\right) = N\nabla_x(n\cdot\phi)(x) + o(N), \tag{6.19}$$

where $\Im(z)$ stands for the imaginary part of the complex number z. Motivated by this heuristic, the synchrosqueezed windowed Fourier transform "squeezes" the time-frequency spectrum by reassigning

the amplitude at (x, k) to $(x, w_f(x, k))$ as

$$T(x, \xi) := \iint |W(x, k)|^2 \delta(\xi - w_f(x, k)) \, \mathrm{d}^2 k. \tag{6.20}$$

This significantly enhances the sharpness of the time-frequency representation, leading to an estimate of the local frequency of the signal, that enjoys better properties than the windowed Fourier transform, as we illustrate below. This gives a sharpened energy distribution on phase space:

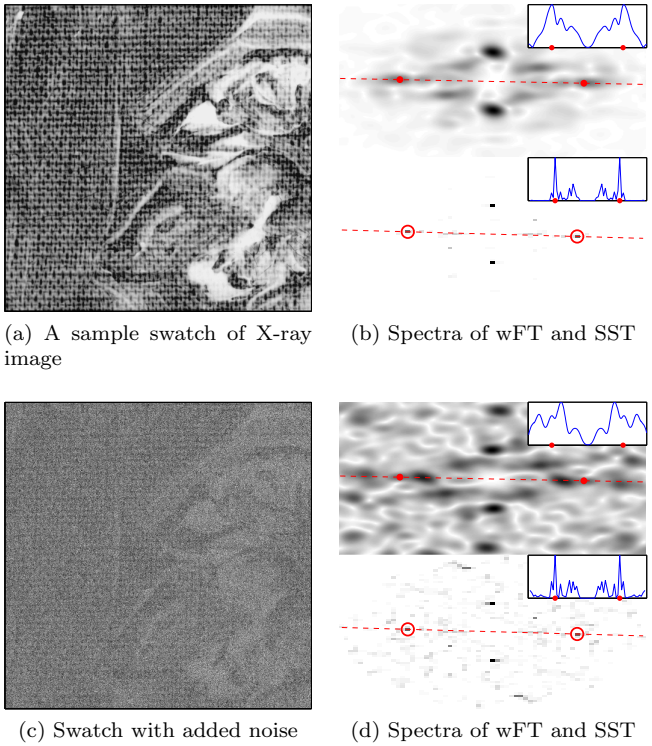
$$T(x, \xi) \approx \sum_{n \in \mathbb{Z}^2} |a(x)|^2 \left|\widehat{S}(n)\right|^2 \delta(\xi - N\nabla(n \cdot \phi(x))), \tag{6.21}$$

in the sense of distributions. See Chapter 2 or [180, 182, 184] for more details, as well as an analysis of the method. The peaks of the synchrosqueezed spectrum T then provide estimates of the $N\nabla(n \cdot \phi(x))$, determining local measurement of both the thread count and the angle. Figure 6.21 illustrates the resulting spectrum of the 2D SST, compared with the wFT for a sample X-ray image from a canvas. The reassignment carried out in (6.20), taking into account the local oscillation of the phase of a highly redundant wFT (in practice we adopt the generalized curvelet transform) rather than the maximum energy of the wFT to reduce the influence of noise, results in a much more concentrated spatial frequency portrait. As illustrated by the behavior of the estimates when extra noise is added, this leads to increased robustness for the estimates of the dominating wave vectors, which determine the thread count and angle. The performance and the robustness of the 2D SST are supported by rigorous mathematical analysis and numerical illustration in Chapter 3 and [183].

6.2.4 Applications to Art Investigations

Let us now present some results of quantitative canvas analysis using 2D SST.

The first example (Fig. 6.22a) is the painting F205 by van Gogh, the X-ray image of which is publicly available as part of the RKD dataset [144] provided by the Netherlands Institute for Art History; this was one of the first examples analyzed using the method based on the windowed Fourier transform; see [100, Figure 4] and also [104, Figure 6]. In Figure 6.23, the thread count and thread angle estimates are shown for horizontal and vertical threads. Comparing with the previous results in [100, 104], we observe that the general characteristics of the canvas agree quite well. For example, [104] reports average thread counts of 13.3 threads/cm (horizontal) and 16.0 threads/cm (vertical), while our method obtains 13.24 threads/cm (horizontal) and 15.92 threads/cm (vertical). Compared to the earlier results, the current analysis gives a more detailed spatial variation of the thread counts. In particular, it captures the oscillation of the thread count on a much finer scale. We don't know whether such fine details will have applications beyond the canvas characterization already achieved by less detailed methods, but it is interesting that they can be captured by an



(a) A sample swatch of X-ray image

(b) Spectra of wFT and SST

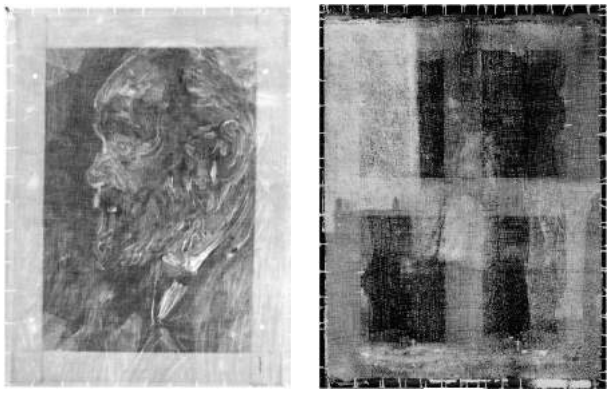
(c) Swatch with added noise

(d) Spectra of wFT and SST

Figure 6.21: (a) A sample swatch of an X-ray image, in which canvas is clearly visible (in most places) despite the paint layers on top of the canvas; (b) The spectrum of the windowed Fourier transform (wFT) (top) and SST (bottom) at one location. Local maxima (circled in red) indicate the wave vector estimates; the insets show the intensity profile on a cross section (dashed line) through two maxima; (c) The same swatch as in (a) with noise added (such that the noise level is visually comparable to the real data example in Figure 6.28) to test for robustness; (d) The wFT and SST spectra again at the same location, illustrating the more robust nature of the SST estimate (due to its taking into account phase information of the wFT in a neighborhood of the peaks of the absolute value of the wFT as well as the peak values). For comparison, the positions of the red circles are the same as in (b). The peaks are displaced in wFT due to noise, while the result of SST is not affected.

automatic method. Note that visual inspection confirms the presence of these fine details.

We next consider a painting of Vermeer, *Woman in Blue Reading a Letter* (L17), the X-ray image of which is also available as part of the RKD dataset [144]. The canvas analysis for Vermeer's paintings is considerably more challenging than that of van Gogh's [121]. This can be understood by direct comparison of the X-ray images in Figures 6.22b and 6.22a. The stretchers and nails significantly perturb the X-ray image for the Vermeer. The results are shown in Figures 6.24 and 6.25. Although the thread count and angle estimate are affected by artifacts in the X-ray image, they still provide a detailed characterization of the canvas weave. This is justified by the result



(a) X-ray image, van Gogh's F205

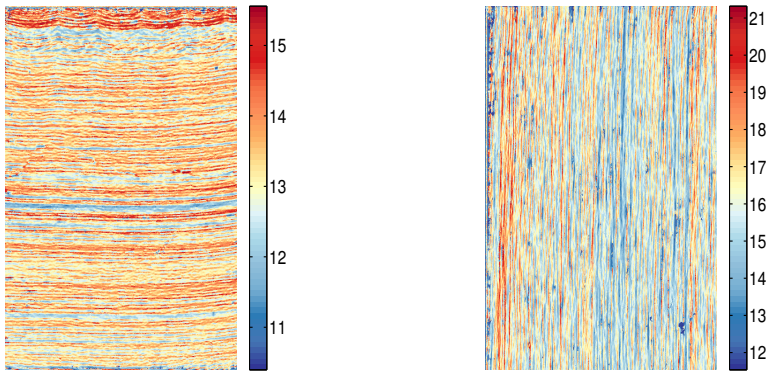
(b) X-ray image, Vermeer's L17

Figure 6.22: (a) X-ray image of van Gogh's painting *Portrait of an Old Man with Beard*, 1885, Van Gogh Museum, Amsterdam (F205); (b) X-ray image of Vermeer's painting *Woman in Blue Reading a Letter*, 1663-64, Rijksmuseum Amsterdam, Amsterdam (L17). X-ray images provided by Professor C. Richard Johnson through the RKD dataset [144].

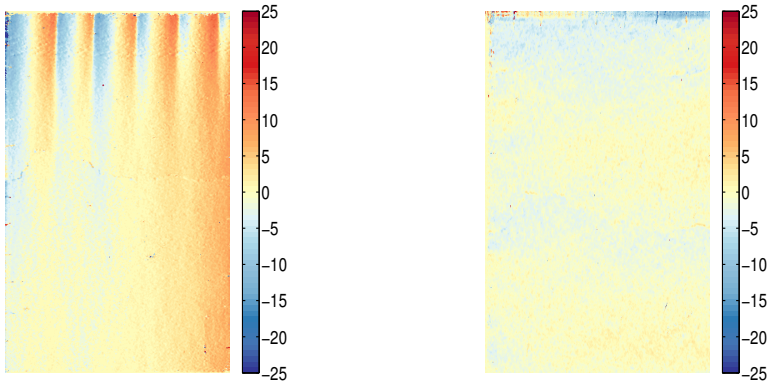
in Figure 6.25, which shows a zoom-in for the X-ray image and the vertical thread angle map. It is observed that the algorithm captures (and quantifies) detailed deviations in the vertical thread angle recognizable by visual inspection. Despite the challenges, the 2D SST-based canvas analysis performs quite well on the Vermeer example.

To test the algorithm on a different type of canvas weave, we applied it to the X-ray image of Albert P. Ryder's *The Pasture*, a painting on twill canvas. Figure 6.26 shows the result for a portion of the canvas. The twill canvas pattern is clear on the zoomed-in X-ray image. The method is still able to capture fine scale features of the canvas; the admittedly higher number of artifacts is due to the increased difficulty to "read" a twill vs. a standard weave pattern, as well as a weaker canvas signal on the X-ray.

For our final example, we apply the 2D SST-based canvas analysis to *The Peruzzi Altarpiece* by Giotto di Bondone and his assistants. The altarpiece is in the collection of the North Carolina Museum of Art; see Figure 6.27 for the altarpiece as well as the X-ray images used in the analysis. This is a painting on wood panel, but the ground of traditional white gesso was applied over a coarsely woven fabric interlayer glued to a poplar panel. We carried out a canvas analysis on the fabric interlayer, likely a hand woven linen cloth. The results of a canvas analysis based on the synchrosqueezed transform are shown in Figure 6.29. This example is much more challenging than the previous ones, since the X-ray intensity contributed by the canvas is much weaker because the ground does not contain lead; see e.g., Figure 6.28, a detail of the X-ray image of the Christ panel. The canvas is barely visible, in sharp contrast to the X-ray images in, e.g., Figures 6.21a or 6.25. All panels except the central Christ panel are cradled; the wood texture of these cradles interferes



(a) Horizontal thread count (b) Vertical thread count

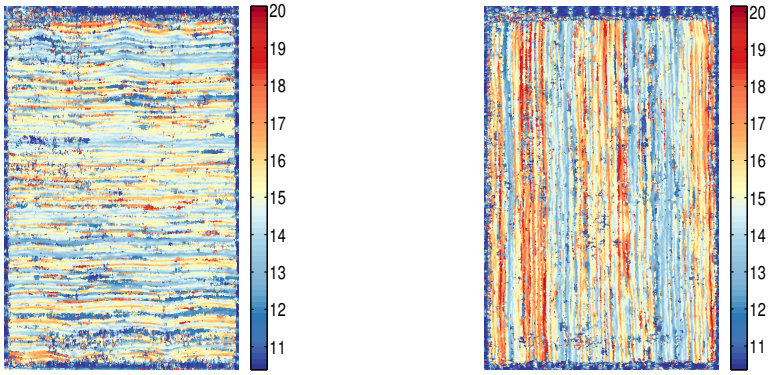


(c) Horizontal thread angle (d) Vertical thread angle

Figure 6.23: The canvas analysis results of van Gogh's F205 using the synchrosqueezed transform: (a) and (b): thread count map of the horizontal and vertical threads; (c) and (d): the estimated thread angle. Compare with [104, Figure 6].

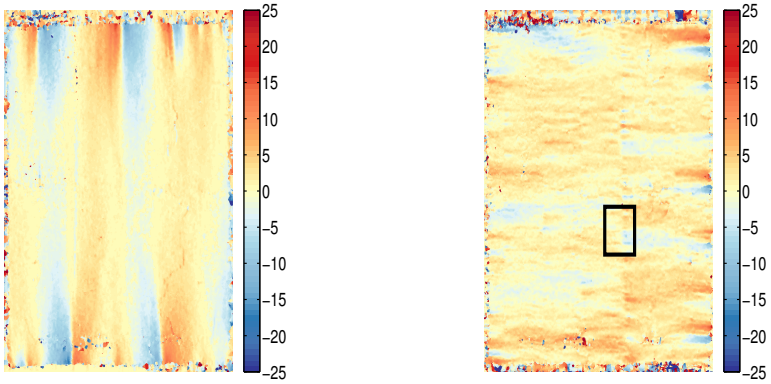
with the canvas pattern on the X-ray image, introducing an additional difficulty. This difficulty is reflected in our results: e.g., the vertical thread count for the central panel has much fewer artifacts than those of the other panels (see Figure 6.29). [In future work, we will explore carrying out a canvas analysis after signal-processing-based virtual cradle removal – see [187].]

One interesting ongoing art investigation debate concerning this altarpiece is the relative position of the panels of John the Baptist and Francis of Assisi. While the order shown in Figure 6.27, with Francis in the right-most position, and the Baptist second from right, is the most commonly accepted [149], there have been alternative arguments that the Francis panel should be instead placed next to the central panel. Typically the grain of the wood as seen in X-rays can be used to set the relative position of panels in an altarpiece painted on a single plank of wood, but because the cradle pattern obscures an accurate reading of the X-rays of the Baptist and Francis this proposed alternative orientation can not be discounted. We wondered what ordering (if any) would be suggested by the



(a) Horizontal thread count

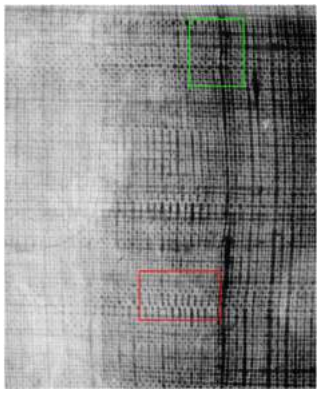
(b) Vertical thread count



(c) Horizontal thread angle

(d) Vertical thread angle

Figure 6.24: Canvas analysis results of Vermeer's L17 using the synchrosqueezed transform: (a) and (b) are thread count map of the horizontal and vertical threads; (c) and (d) show the estimated thread angle. Average thread density is 14.407 threads/cm (horizontal) and 14.817 threads/cm (vertical). The boxed region of the vertical thread angle map (panel (d)) is shown, enlarged, in Figure 6.25; it is part of a striking anomaly in the vertical angle pattern in this canvas, lining up along one vertical traversing the whole canvas.



(a) X-ray image (zoomed-in)

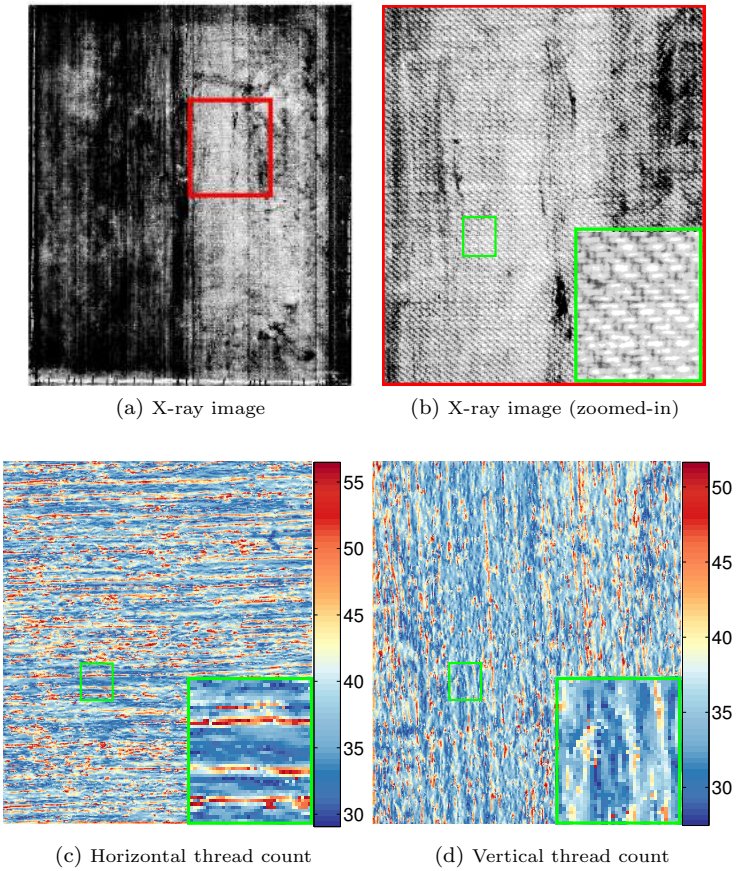
(b) Vertical thread angle (zoomed-in)

Figure 6.25: Details of the X-ray image and the corresponding vertical thread angle map for Vermeer's L17, highlighting two examples (boxed regions) of noticeable fine scale variation of the vertical thread angle, readily recognizable also by visual inspection of the corresponding zones in the X-ray image.

canvas analysis. Under the assumption that the pieces of canvas are cut off consecutively from one larger piece of cloth, we investigated which arrangement provides the best matching. One plausible arrangement of the canvas is shown in Figure 6.30. Our analysis suggests that the canvas of the central panel should be rotated for 90 degrees clockwise to match with the other panels. (The larger height of the central panel, possibly exceeding the width of the cloth roll, may have necessitated this.) Moreover, a better matching is achieved if the canvas of the panel of the Baptist is flipped horizontally (in other words, flipped front to back). Given our results, it seems unlikely that the Francis-panel canvas would fit best to the left of the Baptist-panel canvas. A better, more precise result will be possible after virtual cradle removal. Of course, even a more thorough canvas roll arrangement would not be conclusive evidence for the relative position of the panels themselves; but combined with other elements in a more exhaustive study, it can play a significant role.

6.2.5 Conclusion

We apply $2D$ synchrosqueezed transforms to quantitative canvas weave analysis for art investigations. The synchrosqueezed transforms offer a sharpened phase-space representation of the X-ray image of the paintings, which yields fine scale characterization of thread count and thread angle of the canvas. We demonstrated the effectiveness of the method on art works by van Gogh, Vermeer, and Ryder. The tool is applied to *The Peruzzi Altarpiece* by Giotto and his assistants, to provide insight into the issue of panel arrangement. A software package has been developed and available freely online to help conservators investigate canvas painting.



(a) X-ray image (b) X-ray image (zoomed-in)

(c) Horizontal thread count (d) Vertical thread count

Figure 6.26: (a) X-ray image of Albert P. Ryder's *The Pasture*, 1880-85, North Carolina Museum of Art, Raleigh. (b) is an enlargement of the red-boxed region, with clearly recognizable twill canvas weave. (c) and (d) show the thread count maps corresponding to the zoomed-in region shown in (b). Note the much higher thread counts than for plain weave canvas, typical for the finer threads used in twill weave. The bottom-right insets of (b), (c) and (d) show the further zoom-in of the green-boxed region for visual inspection. The horizontal thread count matches the changes observed in the X-ray image quite well.

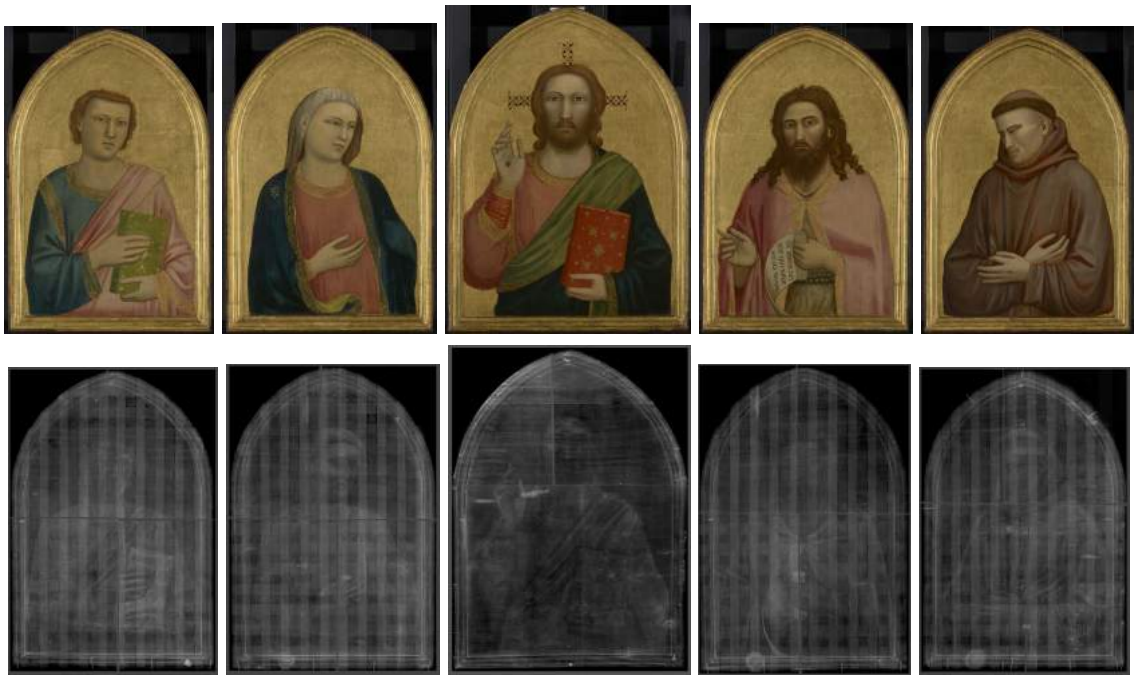


Figure 6.27: Giotto di Bondone and assistants, *The Peruzzi Altarpiece*, ca. 1310-15, North Carolina Museum of Art, Raleigh. The panels from left to right are John the Evangelist, the Virgin Mary, Christ in Majesty, John the Baptist, and Francis of Assisi. The resolution of the X-ray image used in the analysis is 300 DPI. The vertical and (less obvious) horizontal stripes on the X-ray images in all panels except the central panel of Christ are caused by cradling. Each X-ray image is a mosaic of 4 X-ray films, leading to visible boundaries of the different pieces (thin horizontal and vertical lines) on the X-ray image.

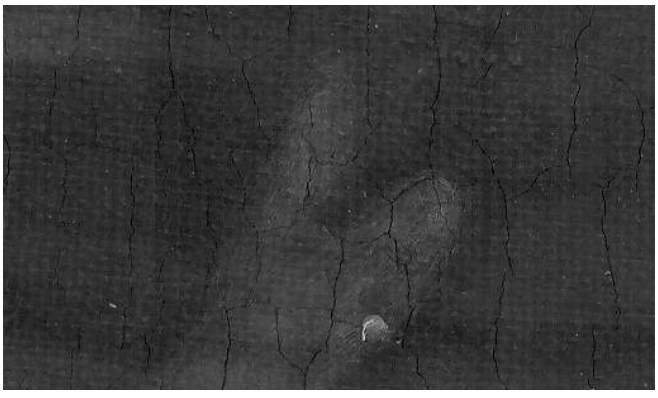


Figure 6.28: A zoomed-in X-ray image of the central panel in *The Peruzzi Altarpiece*. The canvas texture is barely visible, even though the image is scaled such that the thread density is comparable with that of the zoomed-in X-ray in Figure 6.25.

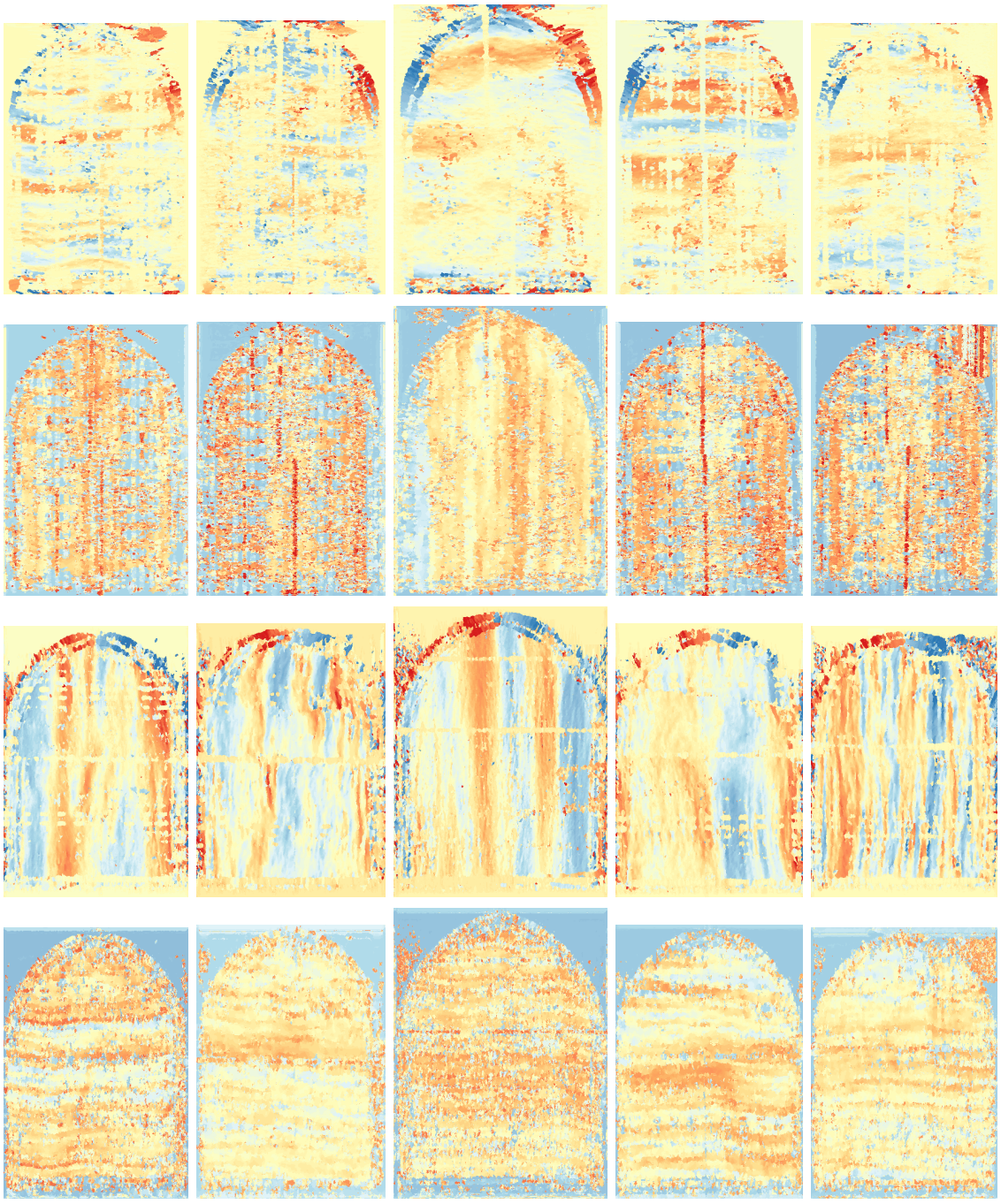


Figure 6.29: Canvas analysis result of the Giotto altarpiece. First row: deviation of vertical thread angle; second row: deviation of vertical thread count; third row: deviation of horizontal thread angle; fourth row: deviation of horizontal thread count. The panels are in the same order as in Figure 6.27.

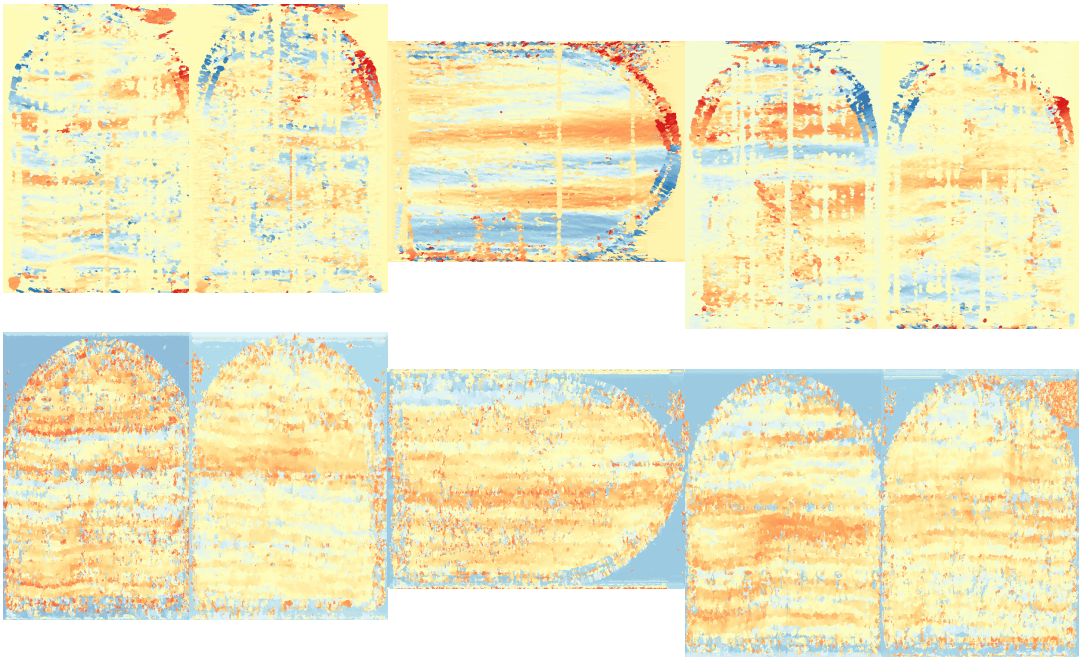


Figure 6.30: A candidate canvas matching and arrangement for *The Peruzzi Altarpiece* by Giotto and his assistants. Top row: deviation of weft thread angle. Bottom row, deviation of warp thread count. (The weft thread count and warp thread angle are not shown as they are less helpful in inferring a possible arrangement.) The canvas pieces from left to right correspond to the panels for John the Evangelist, the Virgin Mary, Christ in Majesty, John the Baptist, and Francis of Assisi (in that order). The canvas of the central panel is rotated clockwise by 90 degrees, and that of the Baptist is flipped horizontally.

Chapter 7

Conclusions of Part I

7.1 Summary

The first part of this thesis has introduced a few new algorithms that enable accurate, efficient and reasonably robust data analysis involving nonlinear wave phenomenon. These algorithms can be widely applied in various areas. In geophysics, they can be applied to problems like seismic wave field separation, ground-roll removal and seismic imaging. In biological and clinical study, they can analyze oscillatory cellular behaviors, tissue level patterns and organism clocks. In astronomy, they can help to solve the gravitational wave detection problem. In materials science, they have been applied to atomic crystal analysis. In art forensics, a new method for painting canvas analysis has been established based on our algorithms. In most cases, our algorithms are significantly faster than the existing state-of-art algorithms and obtain better results. In some difficult examples, we still achieve reliable results efficiently.

7.2 Future Work

Oscillatory data analysis has been an active research line in the past two decades and many important problems remain open. The robustness analysis in [183] suggests that a more adaptive time-frequency transform would lead to a better SST. A good alternative is the chirplet transform [20, 30, 29]. It is possible to develop a more robust method for oscillatory data analysis based on the chirplet transform. Another important issue is to develop methods to handle crossover frequencies, i.e. $\nabla\phi_k(x) = \nabla\phi_j(x)$ for some $k \neq j$ at some point x. Although this is an ill-posed problem, it is very important in real applications. In terms of statistical detection with heavy noise, there is relatively little study on multi-component detection. In real data, there are probably several oscillatory components hidden in a heavily noisy background. Hence, it is significant to establish a novel statistical theory for multi-component detection. In terms of denoising, statistics theory

and numerical tools for denoising smooth or sparse objects have been well established. However, they all focus on constructing a good approximation for the given signal. In some oscillatory data analysis, engineers are more interested in a good approximation for the phase function in the wave-like component.

Part II

Fast Algorithms for Integral Operators in Harmonic Analysis

Chapter 8

Introduction

One of the key problems in computational harmonic analysis is the rapid evaluation of the dense matrix-vector multiplication arising from integral operators in harmonic analysis of the form

$$\mathcal{L}(g(\xi))(x) = \int_{\Omega} K(x,\xi)g(\xi)d\xi \text{ for } x \in X. \tag{8.1}$$

There are numerous useful integral transforms of this form in science and engineering. Some famous examples are the Fourier transform and the Laplacian transform. Suppose N is the number of grid points in each dimension, then after discretization the problem in (8.1) becomes a matrix-vector multiplication

$$u = Kg \tag{8.2}$$

where $K \in \mathbb{C}^{N^d \times N^d}$ is the discrete analogue of the kernel function $K(x,\xi)$, $g \in \mathbb{C}^{N^d}$ is an input vector, $u \in \mathbb{C}^{N^d}$ is an output vector, and d is the dimension of the problem.

The application of these transforms may be very expensive due to mainly three reasons that could result in a large matrix-vector multiplication with a dense matrix K. Different from linear systems coming from the discretization of partial differential equation, the matrix K representing the integral kernel $K(x,\xi)$ are usually dense. In the singular integral transform for which the kernel function $K(x,\xi)$ is singular at some points, it requires a large N to guarantee good numerical accuracy to characterize the singularity. There are also a wide range of integral transforms with highly oscillatory kernels. By the Nyquist-Shannon sampling theorem, N has be sufficiently large such that the discrete analogue $u = Kg$ can approximate the continuous integral in (8.1) precisely.

Given a matrix $K \in \mathbb{C}^{N^d \times N^d}$ and a vector $g \in \mathbb{C}^{N^d}$, the direct computation of the vector $u = Kg \in \mathbb{C}^{N^d}$ takes $O(N^{2d})$ operations since each entry of K contributes to the result. The application of K becomes prohibitive when N is large. A lot of work has been devoted to performing this computation more efficiently without sacrificing the accuracy. Such a reduction in computational

complexity depends highly on the algebraic and numerical properties of the matrix K. For certain types of matrices K, such as the Fourier matrix, numerically low-rank matrices, hierarchically semi-separable (HSS) matrices [177], and hierarchical matrices [82, 122], there exist fast algorithms for computing Kg accurately in $O(N^d \log N)$ or even $O(N^d)$ operations.

In this part of the thesis, we propose several fast algorithms to accelerate the application of large K when it satisfies a special low-rank property. For such a matrix, the rows are typically indexed by a set of points, say X, and the columns by another set of points, say Ω. Both X and Ω are often point sets in $\mathbb{R}^d$ for some dimension d. Associated with X and Ω are two hierarchical trees T_X and T_Ω, respectively and both trees are assumed to have the same depth $L = O(\log N)$, with the top level being level 0 and the bottom one being level L. Such a matrix K of size $N^d \times N^d$ is said to satisfy the **complementary low-rank property** if for any level ℓ, any node A in T_X at level ℓ, and any node B in T_Ω at level $L-\ell$, the submatrix $K_{A,B}$, obtained by restricting K to the rows indexed by the points in A and the columns indexed by the points in B, is numerically low-rank, i.e., for a given precision ϵ there exists a low-rank approximation of $K_{A,B}$ with the 2-norm error bounded by ϵ and the rank bounded polynomially in $\log N$ and $\log(1/\epsilon)$. In many applications, one can even show that the rank is only bounded polynomially in $\log(1/\epsilon)$ and is independent of N. Similarly, it is straightforward to generalize the concept of the complementary low-rank property to a matrix with different row and column dimensions. A well-known example of the complementary low-rank matrices is the matrix representation of a nonuniform Fourier transform.

In general, a matrix K coming from real applications may not satisfy the complementary low-rank property in the whole domain $X \times \Omega$, but the property is true locally in $X \times \Omega$. For example, a kernel $K(x,\xi) = e^{2\pi\imath\Phi(x,\xi)}$ coming from a two-dimensional Fourier integral operator (FIO) [21, 22, 119] may not have a discrete analogue that is complementary low-rank due to the possible singularity of $\Phi(x,\xi)$ at $\xi = 0$. The FIO kernel $K(x,\xi)$ is complementary low-rank in a subdomain $X \times \Omega_j \subset X \times \Omega$ with $\xi = 0$ away from Ω_j. There are mainly two methods to deal with this irregularity at $\xi = 0$. One idea is to apply a well-designed transformation mapping the domain $X \times \Omega$ into a new domain $X \times P$ such that $\xi = 0$ is mapped to the boundary of P. After this transformation, the new kernel function defined on $X \times P$ would be complementary low-rank and our proposed fast algorithms are applicable. Another idea is to partition $X \times \Omega$ into a sequence of subdomains $X \times \Omega_j$ such that the kernel function $K(x,\xi)$ is complementary low-rank in each subdomain. Thus, we can apply the proposed fast algorithms to apply $K(x,\xi)$ efficiently in each subdomain.

We will introduce two kinds of fast algorithms for complementary low-rank matrices. In the first situation, we assume that an explicit kernel $K(x,\xi)$ is known, e.g., an FIO kernel $K(x,\xi) = e^{2\pi\imath\Phi(x,\xi)}$, and the kernel is applied to only a few input functions. In this situation, we propose a **multiscale butterfly algorithm** to efficiently evaluate

$$u(x) = \int_\Omega e^{2\pi\imath\Phi(x,\xi)} g(\xi) d\xi \quad \text{for} \quad x \in X.$$

The multiscale butterfly algorithm requires only $O(N^d \log N)$ operations and $O(N^d)$ memory to evaluate the above integral without precomputation. The complexity has a prefactor smaller than the well-known butterfly algorithm for FIOs in [22]. The multiscale butterfly algorithm can be extended to evaluate (8.1) if $K(x, \xi)$ is complementary low-rank away from $\xi = 0$.

In the second situation, we assume that the complementary low-rank matrix K is repeatedly applied to a large number of input vectors or functions. This assumption is motivated by fast sweeping methods for Helmholtz equations and iterative methods for Kirchoff migration in which K is repeatedly applied. In this situation, we propose the **butterfly factorization**, which represents K as a product of $L + 3$ sparse matrices:

$$K \approx U^L G^{L-1} \cdots G^{L/2} M^{L/2} (H^{L/2})^* \cdots (H^{L-1})^* (V^L)^*, \tag{8.3}$$

where the depth $L = O(\log N)$ of T_X and T_Ω is assumed to be even, and all factors are sparse matrices with $O(N^d)$ nonzero entries. Once the butterfly factorization has been constructed, storing and applying K only requires $O(N^d \log N)$ complexity with a small prefactor. The construction of the butterfly factorization is problem-dependent. We consider two cases that are quite common in applications:

1. Only black-box routines for computing Kg and K^*g in $O(N^d \log N)$ operations are given.

2. Only a black-box routine for evaluating any entry of the matrix K in $O(1)$ operations is given.

In the first case, the butterfly factorization can be constructed in $O(N^{1.5d} \log N)$ operations with $O(N^{1.5d})$ memory complexity. In the second case, the operation and memory complexity for the construction is $O(N^{1.5d})$ and $O(N^d \log N)$, respectively.

The rest of this thesis is organized as follows. In Chapter 9, we introduce the multiscale butterfly algorithm for FIOs. The application of this algorithm to other complementary low-rank kernels is straightforward. In Chapter 10, we start introducing the butterfly factorization for complementary low-rank matrices coming from one-dimensional problems. The butterfly factorization for multi-dimensional complementary low-rank matrices is introduced in Chapter 11. We close this part of thesis with a conclusion in Chapter 12.

Chapter 9

Multiscale Butterfly Algorithm

9.1 Introduction

This chapter is concerned with the rapid application of

$$(\mathcal{L}g)(x) = \int_{\mathbb{R}^d} K(x,\xi)g(\xi)d\xi, \tag{9.1}$$

where $g(\xi)$ is an input function, $K(x,\xi)$ is a complementary low-rank kernel or a kernel that is complementary low-rank away from $\xi = 0$. A famous example is the Fourier integral operators (FIOs), which are defined as

$$(\mathcal{L}f)(x) = \int_{\mathbb{R}^d} a(x,\xi)e^{2\pi \boldsymbol{i}\Phi(x,\xi)}\widehat{f}(\xi)d\xi, \tag{9.2}$$

where

- $a(x,\xi)$ is an amplitude function that is smooth in both x and ξ,
- $\Phi(x,\xi)$ is a phase function that is smooth in (x,ξ) for $\xi \neq 0$ and obeys the homogeneity condition of degree 1 in ξ, namely, $\Phi(x,\lambda\xi) = \lambda\Phi(x,\xi)$ for each $\lambda > 0$, and
- $\widehat{f}$ is the Fourier transform of the input f defined by

$$\widehat{f}(\xi) = \int_{\mathbb{R}^d} e^{-2\pi \boldsymbol{i}x\cdot\xi} f(x)dx.$$

We will focus on the two-dimensional FIOs to introduce the multiscale butterfly algorithm to evaluate (9.2) efficiently. This is joint work with Yingzhou Li and Lexing Ying in [119]. Extending this algorithm to other complementary low-rank kernels in (9.1) is straightforward. In a typical setting, it is often assumed that the problem is periodic (i.e., $a(x,\xi)$, $\Phi(x,\xi)$, and $f(x)$ are all periodic in x)

or the function $f(x)$ decays sufficiently fast so that one can embed the problem in a sufficiently large periodic cell. A simple discretization in two dimensions considers functions f given on a Cartesian grid

$$X = \left\{ x = \left(\frac{n_1}{N}, \frac{n_2}{N} \right), 0 \le n_1, n_2 < N \text{ with } n_1, n_2 \in \mathbb{Z} \right\} \tag{9.3}$$

in a unit square and defines the discrete Fourier integral operator by

$$(Lf)(x) = \sum_{\xi \in \Omega} a(x,\xi) e^{2\pi \boldsymbol{i} \Phi(x,\xi)} \widehat{f}(\xi), \quad x \in X,$$

where

$$\Omega = \left\{ \xi = (n_1, n_2), -\frac{N}{2} \le n_1, n_2 < \frac{N}{2} \text{ with } n_1, n_2 \in \mathbb{Z} \right\}, \tag{9.4}$$

and $\widehat{f}$ is the discrete Fourier transform of f

$$\widehat{f}(\xi) = \frac{1}{N^2} \sum_{x \in X} e^{-2\pi \boldsymbol{i} x \cdot \xi} f(x).$$

In most examples, since $a(x,\xi)$ is a smooth symbol of order zero and type $(1,0)$ [11, 21, 51, 88, 181], $a(x,\xi)$ is numerically low-rank in the joint X and Ω domain and its numerical treatment is relatively easy. Therefore, we will simplify the problem by assuming $a(x,\xi) = 1$ in the following analysis and the algorithm description. We refer the reader to [22] for discussion on how to deal with a nonconstant amplitude function. Under this assumption, the discrete FIO discussed in this chapter takes the following form:

$$(Lf)(x) = \sum_{\xi \in \Omega} e^{2\pi \boldsymbol{i} \Phi(x,\xi)} \widehat{f}(\xi), \quad x \in X. \tag{9.5}$$

A direct computation of (9.5) takes $O(N^4)$ operations, which is quadratic in the number of degrees of freedom, N^2. Hence, a practical need is to design efficient and accurate algorithms to evaluate (9.5). This research topic is of great interest for computing wave equations especially in geophysics [52, 95, 157, 186].

9.1.1 Previous Work

An earlier method for the rapid computation of general FIOs is the algorithm for two-dimensional problems proposed in [21]. This method starts by partitioning the frequency domain Ω into $O(\sqrt{N})$ wedges of equal angle. The integral (9.5) *restricted to each wedge* is then factorized into two components, both of which can be handled efficiently. The first one has a low-rank structure that leads to an $O(N^2 \log N)$ fast computation, while the second one is a non-uniform Fourier transform which can be evaluated in $O(N^2 \log N)$ steps with the algorithms developed in [3, 60, 140]. Summing the

computational cost over all $O(\sqrt{N})$ wedges gives an $O(N^{2.5}\log N)$ computational cost.

Shortly after, an algorithm with quasilinear complexity for general FIOs was proposed in [22] using the framework of the butterfly algorithms in [132, 138]. This approach introduces a polar coordinate transformation in the frequency domain to remove the singularity of $\Phi(x,\xi)$ at $\xi = 0$, proves the existence of low-rank separated approximations between certain pairs of spatial and frequency domains, and implements the low-rank approximations with oscillatory Chebyshev interpolations. The resulting algorithm evaluates (9.5) with $O(N^2 \log N)$ operations and $O(N^2)$ memory, both essentially linear in terms of the number of unknowns.

Another related research direction seeks sparse representations of the FIOs using modern basis functions from harmonic analysis. A sparse representation allows fast matrix-vector products in the transformed domain. Local Fourier transforms [10, 15, 40], wavelet-packet transforms [99], the curvelet transform [19, 23, 24, 25], the wave atom frame [49, 50], and the wave packet frame [4, 45] have been investigated for the purpose of operator sparsification. In spite of favorable asymptotic behaviors, the actual representations of the FIOs typically have a large prefactor constant in terms of both the computational time and the memory requirement. This makes them less competitive compared to the approaches in [21, 22].

9.1.2 Motivation

The main motivation of the current work is to improve the performance of the butterfly algorithm in [22]. As we pointed out earlier, this algorithm starts by applying a polar coordinate transformation in the frequency domain to remove the singularity of the phase function at $\xi = 0$. For this reason, we refer the reader to this algorithm as the **polar butterfly algorithm**. More precisely, the polar butterfly algorithm introduces a polar-Cartesian coordinate transformation $T : (p_1, p_2) \to (\xi_1, \xi_2)$ such that

$$\xi = (\xi_1, \xi_2) = \frac{\sqrt{2}}{2} N p_1 e^{2\pi \boldsymbol{i} p_2}, \quad e^{2\pi \boldsymbol{i} p_2} = (\cos 2\pi p_2, \sin 2\pi p_2). \tag{9.6}$$

Let $P = T^{-1}(\Omega)$. By definition, each point $p = (p_1, p_2) \in P$ belongs to $[0,1]^2$. The new phase function $\Psi(x,p)$ in the p variable is now given by

$$\Psi(x,p) := \frac{1}{N}\Phi(x,\xi) = \frac{\sqrt{2}}{2}\Phi(x, e^{2\pi \boldsymbol{i} p_2})p_1, \tag{9.7}$$

where the last identity comes from the homogeneity of $\Phi(x,\xi)$ in ξ. Thus, computing (9.5) is equivalent to evaluate

$$(Lf)(x) = \sum_{\xi \in \Omega} e^{2\pi \boldsymbol{i} \Phi(x,\xi)} \widehat{f}(\xi) = \sum_{p \in P} e^{2\pi \boldsymbol{i} N \Psi(x,p)} \widehat{f}(T(p)). \tag{9.8}$$

The new phase function $\Psi(x,p)$ is smooth in the whole domain $(x,p) \subset [0,1]^2 \times [0,1]^2$, since $\Phi(x,\xi)$ is smooth in (x,ξ) for $\xi \neq 0$. This smoothness guarantees a low-rank separated approximation of $e^{2\pi \boldsymbol{i} N \Psi(x,p)}$ when x and p are properly restricted to certain subdomains in $X \times P$ under certain geometric configuration. This low-rank property allows for the application of the butterfly algorithm in [188] and results in a fast algorithm with an $O(N^2 \log N)$ computational complexity and an $O(N^2)$ memory complexity.

However, the application of this polar-Cartesian transformation comes with several drawbacks, which result in a large prefactor of the computational complexity. First, due to the polar grid in the frequency domain, the points in P for the butterfly algorithm are irregularly distributed and a separate Chebyshev interpolation matrix is required for the evaluation at each point. In order to avoid the memory bottleneck from storing these interpolation matrices, the polar butterfly algorithm generates these interpolation matrices on-the-fly during the evaluation. This turns out to be expensive in the operation count. Second, since the amplitude and phase functions are often written in the Cartesian coordinates, the polar butterfly algorithm applies the polar-Cartesian transformation for each kernel evaluation. Finally, in order to maintain a reasonable accuracy, the polar butterfly algorithm divides the frequency domain into multiple parts and applies the same butterfly algorithm to each part separately. This also increases the actual running time by a nontrivial constant factor.

9.1.3 Our Contribution

Those drawbacks of the polar butterfly algorithm motivate us to propose a multiscale butterfly algorithm using a Cartesian grid both in the spatial and the frequency domains. To deal with the singularity of the kernel $\Phi(x,\xi)$ at $\xi = 0$, we hierarchically decompose the frequency domain into a union of nonoverlapping Cartesian coronas with a common center $\xi = 0$ (see Figure 9.1). More precisely, define

$$\Omega_j = \left\{ (n_1, n_2) : \frac{N}{2^{j+1}} < \max(|n_1|, |n_2|) \leq \frac{N}{2^j} \right\} \cap \Omega$$

for $j = 1, \ldots, \log N - s$, where s is just a small constant integer. The domain $\Omega_d = \Omega \setminus \cup_j \Omega_j$ is the remaining square grid at the center of constant size. Following this decomposition of the frequency domain, one can write (9.5) accordingly as

$$(Lf)(x) = \sum_j \left(\sum_{\xi \in \Omega_j} e^{2\pi \boldsymbol{i} \Phi(x,\xi)} \widehat{f}(\xi) \right) + \sum_{\xi \in \Omega_d} e^{2\pi \boldsymbol{i} \Phi(x,\xi)} \widehat{f}(\xi). \tag{9.9}$$

The kernel function of (9.9) is smooth in each subdomain Ω_j and a Cartesian butterfly algorithm is applied to evaluate the contribution from Ω_j. For the center square Ω_d, since it contains only a constant number of points, a direct summation is used. Because of the multiscale nature of the frequency domain decomposition, we refer to this algorithm as **the multiscale butterfly**

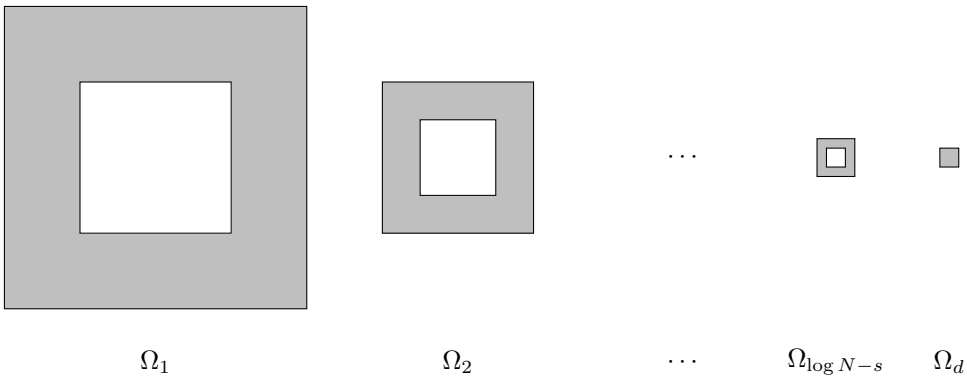


Figure 9.1: This figure shows the frequency domain decomposition of Ω. Each subdomain Ω_j, $j = 1, \ldots, \log N - s$, is a corona and Ω_d is a small square domain near the origin.

algorithm. As we shall see, the computational and memory complexity of the multiscale butterfly algorithm are still $O(N^2 \log N)$ and $O(N^2)$, respectively. On the other hand, the prefactors are much smaller, since the multiscale butterfly is based on the Cartesian grids and requires no polar-Cartesian transformation.

9.1.4 Organization

The rest of this chapter is organized as follows. Section 9.2 presents the overall structure of a butterfly algorithm. Section 9.3 proves a low-rank property that is essential to the multiscale butterfly algorithm. Section 9.4 combines the results of the previous two sections and describes the multiscale butterfly algorithm in detail. In Section 9.5, numerical results of several examples are provided to demonstrate the efficiency of the multiscale butterfly algorithm. Finally, we conclude this chapter with some discussion in Section 9.6.

9.2 The Butterfly Algorithm

This section provides brief description of the overall structure of the butterfly algorithm. In this section, X and Ω refer to two general sets of M points in $\mathbb{R}^2$, respectively. We assume the points in these two sets are distributed quasi-uniformly but they are not necessarily the sets defined in (9.3) and (9.4).

Given an input $\{g(\xi), \xi \in \Omega\}$, the goal is to compute the potentials $\{u(x), x \in X\}$ defined by

$$u(x) = \sum_{\xi \in \Omega} K(x, \xi) g(\xi), \quad x \in X,$$

where $K(x, \xi)$ is a kernel function. Let $D_X \supset X$ and $D_\Omega \supset \Omega$ be two square domains containing X

and Ω respectively. The main data structure of the butterfly algorithm is a pair of quadtrees T_X and T_Ω. Having D_X as its root box, the tree T_X is built by recursive dyadic partitioning of D_X until each leaf box contains only a few points. The tree T_Ω is constructed by recursively partitioning in the same way. With the convention that a root node is at level 0, a leaf node is at level $L = O(\log M)$ under the quasi-uniformity condition about the point distributions, where M is the number of points in X and Ω. Throughout, we shall use A and B to denote the square boxes of T_X and T_Ω with ℓ_A and ℓ_B denoting their levels, respectively.

At the heart of the butterfly algorithm is a special low-rank property. Consider any pair of boxes $A \in T_X$ and $B \in T_\Omega$ obeying the condition $\ell_A + \ell_B = L$. The butterfly algorithm assumes that the submatrix $\{K(x,\xi)\}_{x\in A,\xi\in B}$ to be approximately of a constant rank. More precisely, for any ϵ, there exist a constant r_ϵ independent of M and two sets of functions $\{\alpha_t^{AB}(x)\}_{1\le t\le r_\epsilon}$ and $\{\beta_t^{AB}(\xi)\}_{1\le t\le r_\epsilon}$ such that the following holds:

$$\left| K(x,\xi) - \sum_{t=1}^{r_\epsilon} \alpha_t^{AB}(x)\beta_t^{AB}(\xi) \right| \le \epsilon, \quad \forall x \in A, \forall \xi \in B. \tag{9.10}$$

The number r_ϵ is called the *ϵ-separation rank*. The exact form of the functions $\{\alpha_t^{AB}(x)\}_{1\le t\le r_\epsilon}$ and $\{\beta_t^{AB}(\xi)\}_{1\le t\le r_\epsilon}$ of course depends on the problem to which the butterfly algorithm is applied.

For a given square B in D_Ω, define $u^B(x)$ to be the *restricted potential* over the sources $\xi \in B$

$$u^B(x) = \sum_{\xi\in B} K(x,\xi) g(\xi).$$

The low-rank property gives a compact expansion for $\{u^B(x)\}_{x\in A}$, as summing (9.10) over $\xi \in B$ with weights $g(\xi)$ gives

$$\left| u^B(x) - \sum_{t=1}^{r_\epsilon} \alpha_t^{AB}(x) \left(\sum_{\xi\in B} \beta_t^{AB}(\xi) g(\xi) \right) \right| \le \left(\sum_{\xi\in B} |g(\xi)| \right) \epsilon, \quad \forall x \in A.$$

Therefore, if one can find coefficients $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ obeying

$$\delta_t^{AB} \approx \sum_{\xi\in B} \beta_t^{AB}(\xi) g(\xi), \quad 1 \le t \le r_\epsilon, \tag{9.11}$$

then the restricted potential $\{u^B(x)\}_{x\in A}$ admits a compact expansion

$$\left| u^B(x) - \sum_{t=1}^{r_\epsilon} \alpha_t^{AB}(x) \delta_t^{AB} \right| \le \left(\sum_{\xi\in B} |g(\xi)| \right) \epsilon, \quad \forall x \in A.$$

A key point of the butterfly algorithm is that for each pair (A, B), the number of terms in the

expansion is independent of M.

Computing $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ by means of (9.11) for all pairs A, B is not efficient when B is a large box because for each B there are many paired boxes A. The butterfly algorithm, however, comes with an efficient way of computing $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ recursively. The general structure of the algorithm consists of a top down traversal of T_X and a bottom up traversal of T_Ω, carried out simultaneously.

1. Construct the trees T_X and T_Ω with root nodes D_X and D_Ω.

2. Let A be the root of T_X. For each leaf box B of T_Ω, construct the expansion coefficients $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ for the potential $\{u^B(x)\}_{x\in A}$ by simply setting

$$\delta_t^{AB} = \sum_{\xi\in B} \beta_t^{AB}(\xi) g(\xi), \quad 1 \le t \le r_\epsilon. \tag{9.12}$$

3. For $\ell = 1, 2, \ldots, L$, visit level ℓ in T_X and level $L-\ell$ in T_Ω. For each pair (A, B) with $\ell_A = \ell$ and $\ell_B = L-\ell$, construct the expansion coefficients $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ for the potential $\{u^B(x)\}_{x\in A}$ using the low-rank representation constructed at the previous level ($\ell = 0$ is the initialization step). Let P be A's parent and C be a child of B. Throughout, we shall use the notation $C \succ B$ when C is a child of B. At level $\ell - 1$, the expansion coefficients $\{\delta_s^{PC}\}_{1\le s\le r_\epsilon}$ of $\{u^C(x)\}_{x\in P}$ are readily available and we have

$$\left| u^C(x) - \sum_{s=1}^{r_\epsilon} \alpha_s^{PC}(x)\delta_s^{PC} \right| \le \left(\sum_{\xi\in C} |g(\xi)| \right) \epsilon, \quad \forall x \in P.$$

Since $u^B(x) = \sum_{C\succ B} u^C(x)$, the previous inequality implies that

$$\left| u^B(x) - \sum_{C\succ B} \sum_{s=1}^{r_\epsilon} \alpha_s^{PC}(x)\delta_s^{PC} \right| \le \left(\sum_{\xi\in B} |g(\xi)| \right) \epsilon, \quad \forall x \in P.$$

Since $A \subset P$, the above approximation is of course true for any $x \in A$. However, since $\ell_A + \ell_B = L$, the sequence of restricted potentials $\{u^B(x)\}_{x\in A}$ also has a low-rank approximation of size r_ϵ, namely,

$$\left| u^B(x) - \sum_{t=1}^{r_\epsilon} \alpha_t^{AB}(x)\delta_t^{AB} \right| \le \left(\sum_{\xi\in B} |g(\xi)| \right) \epsilon, \quad \forall x \in A.$$

Combining the last two approximations, we obtain that $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ should obey

$$\sum_{t=1}^{r_\epsilon} \alpha_t^{AB}(x)\delta_t^{AB} \approx \sum_{C\succ B} \sum_{s=1}^{r_\epsilon} \alpha_s^{PC}(x)\delta_s^{PC}, \quad \forall x \in A. \tag{9.13}$$

This represents an overdetermined linear system for $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ in cases when $\{\delta_s^{PC}\}_{1\le s\le r_\epsilon, C\succ B}$

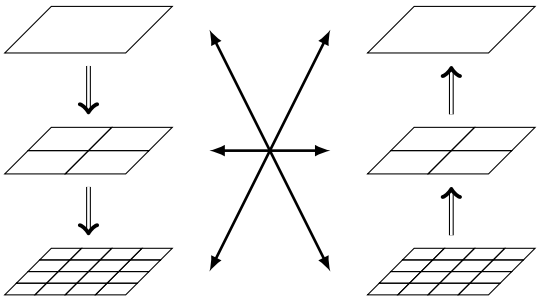


Figure 9.2: Hierarchical domain trees of the two-dimensional butterfly algorithm. Left: T_X for the spatial domain D_X. Right: T_Ω for the frequency domain D_Ω. The interactions between subdomains $A \subset D_X$ and $B \subset D_\Omega$ are represented by left-right arrow lines.

are available. Instead of computing $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ with a least-square method, the butterfly algorithm typically uses an efficient linear transformation approximately mapping $\{\delta_s^{PC}\}_{1\le s\le r_\epsilon, C\succ B}$ into $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$. The actual implementation of this step is very much application-dependent.

4. Finally, let $\ell = L$ and set B to be the root node of T_Ω. For each leaf box $A \in T_X$, use the constructed expansion coefficients $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ to evaluate $u(x)$ for each $x \in A$,

$$u(x) = \sum_{t=1}^{r_\epsilon} \alpha_t^{AB}(x)\delta_t^{AB}. \tag{9.14}$$

A schematic illustration of this algorithm is provided in Figure 9.2. We would like to emphasize that the strict balance between the levels of the target boxes A and source boxes B maintained throughout this procedure is the key to obtaining the accurate low-rank separated approximations.

9.3 Low-Rank Approximations

In this section, the sets X and Ω refer to the sets defined in (9.3) and (9.4). In order to apply the algorithm in Section 9.5, one would require the existence of the following low-rank separated representation:

$$e^{2\pi i\Phi(x,\xi)} \approx \sum_{t=1}^{r_\epsilon} \alpha_t^{AB}(x)\beta_t^{AB}(\xi)$$

for any pair of boxes A and B such that $\ell_A + \ell_B = L$. However, this is not true for a general FIO kernel $e^{2\pi i\Phi(x,\xi)}$ due to the singularity of $\Phi(x,\xi)$ at the origin $\xi = 0$, i.e., when the square B in Ω is close to the origin of the frequency domain. However, if the frequency domain B is well separated from the origin $\xi = 0$ in a relative sense, one can prove a low-rank separated representation.

In order to make it more precise, for two given squares $A \subset X$ and $B \subset \Omega$, we introduce a new

function called the residue phase function

$$R^{AB}(x,\xi) := \Phi(x,\xi) - \Phi(c_A,\xi) - \Phi(x,c_B) + \Phi(c_A,c_B), \tag{9.15}$$

where c_A and c_B are the centers of A and B respectively. Using this new definition, the kernel can be written as

$$xe^{2\pi \boldsymbol{i}\Phi(x,\xi)} = e^{2\pi \boldsymbol{i}\Phi(c_A,\xi)} e^{2\pi \boldsymbol{i}\Phi(x,c_B)} e^{-2\pi \boldsymbol{i}\Phi(c_A,c_B)} e^{2\pi \boldsymbol{i}R^{AB}(x,\xi)}. \tag{9.16}$$

Theorem 9.3.1. *Suppose $\Phi(x,\xi)$ is a phase function that is real analytic for x and ξ away from $\xi = 0$. There exist positive constants ϵ_0 and N_0 such that the following is true. Let A and B be two squares in X and Ω, respectively, obeying $w_A w_B \le 1$ and dist$(B,0) \ge \frac{N}{4}$. For any positive $\epsilon \le \epsilon_0$ and $N \ge N_0$, there exists an approximation*

$$\left| e^{2\pi \boldsymbol{i}R^{AB}(x,\xi)} - \sum_{t=1}^{r_\epsilon} \tilde{\alpha}_t^{AB}(x)\tilde{\beta}_t^{AB}(\xi) \right| \le \epsilon$$

for $x \in A$ and $\xi \in B$ with $r_\epsilon \lesssim \log^4(\frac{1}{\epsilon})$. Moreover,

- *when $w_B \le \sqrt{N}$, the functions $\{\tilde{\beta}_t^{AB}(\xi)\}_{1\le t\le r_\epsilon}$ can all be chosen as monomials in $(\xi - c_B)$ with a degree not exceeding a constant times $\log^2(1/\epsilon)$,*
- *and when $w_A \le 1/\sqrt{N}$, the functions $\{\tilde{\alpha}_t^{AB}(x)\}_{1\le t\le r_\epsilon}$ can all be chosen as monomials in $(x - c_A)$ with a degree not exceeding a constant times $\log^2(1/\epsilon)$.*

In Theorem 9.3.1, w_A and w_B denote the side lengths of A and B, respectively; dist$(B,0)$ denotes the distance between the square B and the origin 0 in the frequency domain. The distance is given by $\text{dist}(B,0) = \min_{\xi\in B} \|\xi - 0\|$. Throughout this chapter, when we write $O(\cdot)$, $\lesssim$ and $\gtrsim$, the implicit constant is independent of N and ϵ.

Proof. Since $w_A w_B \le 1$, we have either $w_A \le 1/\sqrt{N}$ or $w_B \le \sqrt{N}$, or we have both.

Let us first consider the case $w_B \le \sqrt{N}$. Then

$$\begin{aligned} R^{AB}(x,\xi) &= \Phi(x,\xi) - \Phi(c_A,\xi) - \Phi(x,c_B) + \Phi(c_A,c_B) \\ &= [\Phi(x,\xi) - \Phi(c_A,\xi)] - [\Phi(x,c_B) - \Phi(c_A,c_B)] \\ &= H(x,\xi) - H(x,c_B), \end{aligned}$$

where $H(x,\xi) := \Phi(x,\xi) - \Phi(c_A,\xi)$. The function $R^{AB}(x,\xi)$ inherits the smoothness from $\Phi(x,\xi)$. Applying the multivariable Taylor expansion of degree k in ξ centered at c_B gives

$$R^{AB}(x,\xi) = \sum_{1\le|i|<k} \frac{\partial_\xi^i H(x,c_B)}{i!}(\xi - c_B)^i + \sum_{|i|=k} \frac{\partial_\xi^i H(x,\xi^*)}{i!}(\xi - c_B)^i, \tag{9.17}$$

where ξ^* is a point in the segment between c_B and ξ. Here $i = (i_1, i_2)$ is a multi-index with $i! = i_1! i_2!$, and $|i| = i_1 + i_2$. Let us first choose the degree k so that the second sum in (9.17) is bounded by $\epsilon/(4\pi)$. For each i with $|i| = k$, the definition of $H(x, \xi)$ gives

$$\partial_\xi^i H(x, \xi^*) = \sum_{|j|=1} \partial_x^j \partial_\xi^i \Phi(x^*, \xi^*)(x - c_A)^j,$$

for some point x^* in the segment between c_A and x. Using the fact that $\Phi(x, \xi)$ is real-analytic over $|\xi| = 1$ gives that there exists a radius R such that

$$|\partial_x^j \partial_\xi^i \Phi(x, \xi)| \le C i! j! \frac{1}{R^{|i+j|}} = C i! j! \frac{1}{R^{k+1}},$$

for ξ with $|\xi| = 1$. Here the constant C is independent of k. Since $\Phi(x, \xi)$ is homogeneous of degree 1 in ξ, a scaling argument shows that

$$|\partial_x^j \partial_\xi^i \Phi(x^*, \xi^*)| \le C i! j! \frac{1}{R^{k+1} |\xi^*|^{k-1}}.$$

Since $\mathrm{dist}(B, 0) \ge N/4$ and $w_A w_B \le 1$, we have

$$\left| \frac{\partial_\xi^i H(x, \xi^*)}{i!} (\xi - c_B)^i \right| \le \frac{2C i! j!}{i!} \frac{1}{R^{k+1} |\xi^*|^{k-1}} w_A w_B^k \le \frac{2C}{R^{k+1}} \left(\frac{4}{\sqrt{N}} \right)^{k-1}.$$

Combining this with (9.17) gives

$$\left| R^{AB}(x, \xi) - \sum_{1 \le |i| < k} \frac{\partial_\xi^i H(x, c_B)}{i!} (\xi - c_B)^i \right| = \left| \sum_{|i|=k} \frac{\partial_\xi^i H(x, \xi^*)}{i!} (\xi - c_B)^i \right| \le \frac{2C(k+1)}{R^{k+1}} \left(\frac{4}{\sqrt{N}} \right)^{k-1}.$$

Therefore, for a sufficient large $N_0(R)$, if $N > N_0(R)$, choosing $k = k_\epsilon = O(\log(1/\epsilon))$ ensures that the difference is bounded by $\epsilon/(4\pi)$.

The special case $k = 1$ results in the following bound for $R^{AB}(x, \xi)$

$$|R^{AB}(x, \xi)| \le \frac{4C}{R^2}.$$

To simplify the notation, we define

$$R_\epsilon^{AB}(x, \xi) := \sum_{1 \le |i| < k_\epsilon} \frac{\partial_\xi^i H(x, c_B)}{i!} (\xi - c_B)^i,$$

i.e., the first sum on the right-hand side of (9.17) with $k = k_\epsilon$. The choice of k_ϵ together with (9.17)

implies the bound

$$|R_\epsilon^{AB}(x,\xi)| \le \frac{4C}{R^2} + \epsilon.$$

Since $R_\epsilon^{AB}(x,\xi)$ is bounded, a direct application of Lemma 3.2 of [22] gives

$$\left| e^{2\pi \boldsymbol{i} R_\epsilon^{AB}(x,\xi)} - \sum_{p=0}^{d_\epsilon} \frac{(2\pi \boldsymbol{i} R_\epsilon^{AB}(x,\xi))^p}{p!} \right| \le \epsilon/2, \tag{9.18}$$

where $d_\epsilon = O(\log(1/\epsilon))$. Since $R_\epsilon^{AB}(x,\xi)$ is a polynomial in $(\xi - c_B)$, the sum in (9.18) is also a polynomial in $(\xi - c_B)$ with degree bounded by $k_\epsilon d_\epsilon = O(\log^2(1/\epsilon))$. Since our problem is in two dimensions, there are at most $O(\log^4(1/\epsilon))$ possible monomials in $(\xi - c_B)$ with degree bounded by $k_\epsilon d_\epsilon$. Grouping the terms with the same multi-index in ξ results in an $O(\log^4(1/\epsilon))$ term ϵ-accurate separated approximation for $e^{2\pi \boldsymbol{i} R_\epsilon^{AB}(x,\xi)}$ with the factors $\{\tilde{\beta}_t^{AB}(\xi)\}_{1\le t\le r_\epsilon}$ being monomials of $(\xi - c_B)$.

Finally, from the inequality $|e^{\imath a} - e^{\imath b}| \le |a - b|$, it is clear that a separated approximation for $e^{2\pi \boldsymbol{i} R_\epsilon^{AB}(x,\xi)}$ with accuracy $\epsilon/2$ is also one for $e^{2\pi \boldsymbol{i} R^{AB}(x,\xi)}$ with accuracy $\epsilon/2 + \epsilon/2 = \epsilon$. This completes the proof for the case $w_B \le \sqrt{N}$.

The proof for the case $w_A \le 1/\sqrt{N}$ is similar. The only difference is that we now group with

$$R^{AB}(x,\xi) = [\Phi(x,\xi) - \Phi(x,c_B)] - [\Phi(c_A,\xi) - \Phi(c_A,c_B)]$$

and apply the multivariable Taylor expansion in x centered at c_A instead. This results in an $O(\log^4(1/\epsilon))$ term ϵ-accurate separated approximation for $e^{2\pi \boldsymbol{i} R^{AB}(x,\xi)}$ with the factors $\{\tilde{\alpha}_t^{AB}(x)\}_{1\le t\le r_\epsilon}$ being monomials of $(x - c_A)$. □

Though the above proof is constructive, it is cumbersome to construct the separated approximation this way. On the other hand, the proof shows that when $w_B \le \sqrt{N}$, the ξ-dependent factors in the low-rank approximation of $e^{2\pi \boldsymbol{i} R^{AB}(x,\xi)}$ are all monomials in $(\xi - c_B)$. Similarly, when $w_A \le 1/\sqrt{N}$, the x-dependent factors are monomials in $(x - c_A)$. This suggests using Chebyshev interpolation in x when $w_A \le 1/\sqrt{N}$ and in ξ when $w_B \le \sqrt{N}$. For this purpose, we associate with each box a Chebyshev grid as follows.

For a fixed integer q, the Chebyshev grid of order q on $[-1/2, 1/2]$ is defined by

$$\left\{ z_i = \frac{1}{2}\cos\left(\frac{i\pi}{q-1}\right) \right\}_{0\le i\le q-1}.$$

A tensor-product grid *adapted to a square* with center c and side length w is then defined via shifting and scaling as

$$\{c + w(z_i, z_j)\}_{i,j=0,1,\ldots,q-1}$$

In what follows, M_t^B is the two-dimensional Lagrange interpolation polynomial on the Chebyshev

grid adapted to the square B (i.e., using $c = c_B$ and $w = w_B$).

Theorem 9.3.2. *Let A and B be as in Theorem 9.3.1. Then for any $\epsilon \le \epsilon_0$ and $N \ge N_0$ where ϵ_0 and N_0 are the constants in Theorem 9.3.1, there exists $q_\epsilon \lesssim \log^2(1/\epsilon)$ such that the following hold:*

- *when $w_B \le \sqrt{N}$, the Lagrange interpolation of $e^{2\pi \boldsymbol{i} R^{AB}(x,\xi)}$ in ξ on a $q_\epsilon \times q_\epsilon$ Chebyshev grid $\{g_t^B\}_{1\le t\le r_\epsilon}$ adapted to B obeys*

$$\left| e^{2\pi \boldsymbol{i} R^{AB}(x,\xi)} - \sum_{t=1}^{r_\epsilon} e^{2\pi \boldsymbol{i} R^{AB}(x,g_t^B)} M_t^B(\xi) \right| \le \epsilon, \quad \forall x \in A, \forall \xi \in B, \tag{9.19}$$

- *when $w_A \le 1/\sqrt{N}$, the Lagrange interpolation of $e^{2\pi \boldsymbol{i} R^{AB}(x,\xi)}$ in x on a $q_\epsilon \times q_\epsilon$ Chebyshev grid $\{g_t^A\}_{1\le t\le r_\epsilon}$ adapted to A obeys*

$$\left| e^{2\pi \boldsymbol{i} R^{AB}(x,\xi)} - \sum_{t=1}^{r_\epsilon} M_t^A(x) e^{2\pi \boldsymbol{i} R^{AB}(g_t^A,\xi)} \right| \le \epsilon, \quad \forall x \in A, \forall \xi \in B. \tag{9.20}$$

Both (9.19) *and* (9.20) *provide a low-rank approximation with $r_\epsilon = q_\epsilon^2 \lesssim \log^4(1/\epsilon)$ terms.*

The proof for this follows exactly that of Theorem 3.3 in [22].

Finally, we are ready to construct the low-rank approximation for the kernel $e^{2\pi \boldsymbol{i}\Phi(x,\xi)}$, i.e.,

$$e^{2\pi \boldsymbol{i}\Phi(x,\xi)} \approx \sum_{t=1}^{r_\epsilon} \alpha_t^{AB}(x) \beta_t^{AB}(\xi). \tag{9.21}$$

When $w_B \le \sqrt{N}$, one multiply (9.19) with $e^{2\pi \boldsymbol{i}\Phi(c_A,\xi)} e^{2\pi \boldsymbol{i}\Phi(x,c_B)} e^{-2\pi \boldsymbol{i}\Phi(c_A,c_B)}$, which gives that $\forall x \in A, \forall \xi \in B$

$$\left| e^{2\pi \boldsymbol{i}\Phi(x,\xi)} - \sum_{t=1}^{r_\epsilon} e^{2\pi \boldsymbol{i}\Phi(x,g_t^B)} \left(e^{-2\pi \boldsymbol{i}\Phi(c_A,g_t^B)} M_t^B(\xi) e^{2\pi \boldsymbol{i}\Phi(c_A,\xi)} \right) \right| \le \epsilon.$$

In terms of the notation in (9.21), the expansion functions are given by

$$\alpha_t^{AB}(x) = e^{2\pi \boldsymbol{i}\Phi(x,g_t^B)}, \quad \beta_t^{AB}(\xi) = e^{-2\pi \boldsymbol{i}\Phi(c_A,g_t^B)} M_t^B(\xi) e^{2\pi \boldsymbol{i}\Phi(c_A,\xi)}, \quad 1 \le t \le r_\epsilon. \tag{9.22}$$

This is a special interpolant of the function $e^{2\pi \boldsymbol{i}\Phi(x,\xi)}$ in the ξ variable, which prefactors the oscillation, performs the interpolation, and then remodulates the outcome. When $w_A \le 1/\sqrt{N}$, multiply (9.20) with $e^{2\pi \boldsymbol{i}\Phi(c_A,\xi)} e^{2\pi \boldsymbol{i}\Phi(x,c_B)} e^{-2\pi \boldsymbol{i}\Phi(c_A,c_B)}$ and obtain that $\forall x \in A, \forall \xi \in B$

$$\left| e^{2\pi \boldsymbol{i}\Phi(x,\xi)} - \sum_{t=1}^{r_\epsilon} \left(e^{2\pi \boldsymbol{i}\Phi(x,c_B)} M_t^A(x) e^{-2\pi \boldsymbol{i}\Phi(g_t^A,c_B)} \right) e^{2\pi \boldsymbol{i}\Phi(g_t^A,\xi)} \right| \le \epsilon.$$

The expansion functions are now

$$\alpha_t^{AB}(x) = e^{2\pi \boldsymbol{i} \Phi(x, c_B)} M_t^A(x) e^{-2\pi \boldsymbol{i} \Phi(g_t^A, c_B)}, \quad \beta_t^{AB}(\xi) = e^{2\pi \boldsymbol{i} \Phi(g_t^A, \xi)}, \quad 1 \le t \le r_\epsilon. \tag{9.23}$$

Due to the presence of the demodulation and remodulation steps in the definitions (9.22) and (9.23), we refer to them as *oscillatory Chebyshev interpolations*.

9.4 Multiscale Butterfly Algorithm

In this section, we combine the low-rank approximations described in Section 9.3 with the butterfly algorithm in Section 9.2. Due to the restriction on the distance between B and the origin, we decompose (9.5) into a multiscale summation

$$(Lf)(x) = \sum_{\xi \in \Omega_d} e^{2\pi \boldsymbol{i} \Phi(x,\xi)} \widehat{f}(\xi) + \sum_j \sum_{\xi \in \Omega_j} e^{2\pi \boldsymbol{i} \Phi(x,\xi)} \widehat{f}(\xi), \tag{9.24}$$

where

$$\Omega_j = \left\{ (n_1, n_2) : \frac{N}{2^{j+1}} < \max(|n_1|, |n_2|) \le \frac{N}{2^j} \right\} \cap \Omega$$

for $j = 1, \ldots, \log N - s$, s is a constant, and $\Omega_d = \Omega \setminus \cup_j \Omega_j$.

The term of Ω_d can be evaluated directly since $|\Omega_d| = O(1)$. Let us now fix an Ω_j. Since any square B in Ω_j always stays away from the origin, the results in Section 9.3 apply to the term for Ω_j in (9.24). Therefore, the butterfly algorithm as described in Section 9.2 can be adapted to evaluate

$$\sum_{\xi \in \Omega_j} e^{2\pi \boldsymbol{i} \Phi(x,\xi)} \widehat{f}(\xi)$$

for the Cartesian domains X and Ω_j. In contrast to the polar butterfly algorithm that works in the polar coordinates for Ω, we refer to this one as *the Cartesian butterfly algorithm*.

9.4.1 Cartesian Butterfly Algorithm

To make it more explicit, let us first consider the interaction between (X, Ω_1), with the low-rank approximation implemented using the oscillatory Chebyshev interpolation discussed in Section 9.3.

1. *Preliminaries.* Construct two quadtrees T_X and T_{Ω_1} for X and Ω_1 by uniform hierarchical partitioning. Let b be a constant greater than or equal to 4 and define $N_1 = N$.

2. *Initialization.* For each square $A \in T_X$ of width $1/b$ and each square $B \in T_{\Omega_1}$ of width b, the low-rank approximation functions are

$$\alpha_t^{AB}(x) = e^{2\pi \boldsymbol{i} \Phi(x, g_t^B)}, \quad \beta_t^{AB}(\xi) = e^{-2\pi \boldsymbol{i} \Phi(c_A, g_t^B)} M_t^B(\xi) e^{2\pi \boldsymbol{i} \Phi(c_A, \xi)}, \quad 1 \le t \le r_\epsilon. \tag{9.25}$$

Hence, we can define the expansion weights $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ with

$$\delta_t^{AB} := \sum_{\xi\in B} \beta_t^{AB}(\xi)\widehat{f}(\xi) = e^{-2\pi \boldsymbol{i}\Phi(c_A,g_t^B)} \sum_{\xi\in B}\left(M_t^B(\xi)e^{2\pi\boldsymbol{i}\Phi(c_A,\xi)}\widehat{f}(\xi)\right). \tag{9.26}$$

3. *Recursion.* Go up in tree T_{Ω_1} and down in tree T_X at the same time until we reach the level such that $w_B = \sqrt{N_1}$. At each level, visit all the pairs (A, B). We apply the Chebyshev interpolation in variable ξ and still define the approximation functions given in (9.25). Let $\{\delta_s^{PC}\}_{1\le s\le r_\epsilon}$ denote the expansion coefficients available in previous steps, where P is A's parent, C is a child of B, and s indicates the Chebyshev grid points in previous domain pairs. We define the new expansion coefficients $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ as

$$\delta_t^{AB} := e^{-2\pi\boldsymbol{i}\Phi(c_A,g_t^B)} \sum_{C\succ B}\sum_{s=1}^{r_\epsilon} M_t^B(g_s^C)e^{2\pi\boldsymbol{i}\Phi(c_A,g_s^C)}\delta_s^{PC}, \tag{9.27}$$

where we recall that the notation $C \succ B$ means that C is a child of B.

4. *Switch.* For the levels visited, the Chebyshev interpolation is applied in variable ξ, while the interpolation is applied in variable x for levels $l > \log(N_1)/2$. Hence, we are switching the interpolation method at this step. Now we are still working on level $l = \log(N_1)/2$ and the same domain pairs (A, B) in the last step. Let δ_s^{AB} denote the expansion weights obtained by Chebyshev interpolation in variable ξ in the last step. Correspondingly, $\{g_s^B\}_s$ are the grid points in B in the last step. We take advantage of the interpolation in variable x in A and generate grid points $\{g_t^A\}_{1\le t\le r_\epsilon}$ in A. Then we can define new expansion weights

$$\delta_t^{AB} := \sum_{s=1}^{r_\epsilon} e^{2\pi\boldsymbol{i}\Phi(g_t^A,g_s^B)}\delta_s^{AB}.$$

5. *Recursion.* Go up in tree T_{Ω_1} and down in tree T_X at the same time until we reach the level such that $w_B = N_1/b$. We construct the approximation functions by Chebyshev interpolation in variable x as follows:

$$\alpha_t^{AB}(x) = e^{2\pi\boldsymbol{i}\Phi(x,c_B)}M_t^A(x)e^{-2\pi\boldsymbol{i}\Phi(g_t^A,c_B)}, \quad \beta_t^{AB}(\xi) = e^{2\pi\boldsymbol{i}\Phi(g_t^A,\xi)}. \tag{9.28}$$

We define the new expansion coefficients $\{\delta_t^{AB}\}_{1\le t\le r_\epsilon}$ as

$$\delta_t^{AB} := \sum_{C\succ B} e^{2\pi\boldsymbol{i}\Phi(g_t^A,c_C)}\sum_{s=1}^{r_\epsilon}\left(M_s^P(g_t^A)e^{-2\pi\boldsymbol{i}\Phi(g_s^P,c_C)}\delta_s^{PC}\right), \tag{9.29}$$

where again P is A's parent and C is a child box of B.

6. *Termination.* Finally, we reach the level that $w_B = N_1/b$. For each B on this level and for each square $A \in T_X$ of width b/N_1, we apply the approximation functions given by (9.28) and obtain

$$u^B(x) := e^{2\pi \boldsymbol{i} \Phi(x, c_B)} \sum_{t=1}^{r_\epsilon} \left(M_t^A(x) e^{-2\pi \boldsymbol{i} \Phi(g_t^A, c_B)} \delta_t^{AB} \right) \tag{9.30}$$

for each $x \in A$. Finally, summing over all B on this level, we have

$$u^{\Omega_1}(x) := \sum_B u^B(x) \tag{9.31}$$

for each $x \in A$.

We would like to emphasize that the center part of the tree T_{Ω_j} is always empty since Ω_j is a corona. Accordingly, the algorithm skips this empty part.

For a general Ω_j, the interaction between (X, Ω_j) follows a similar algorithm, except that we replace Ω_1 with Ω_j, $u^{\Omega_1}(x)$ with $u^{\Omega_j}(x)$, and N_1 with $N_j = N/2^{j-1}$, and stop at the level that $w_B = N_j/b$.

Finally, (9.24) is evaluated via

$$(Lf)(x) = u^{\Omega_d}(x) + \sum_j u^{\Omega_j}(x). \tag{9.32}$$

9.4.2 Complexity Analysis

The cost of evaluating the term of Ω_d takes at most $O(N^2)$ steps since $|\Omega_d| = O(1)$. Let us now consider the cost of the terms associated with $\{\Omega_j\}$.

For the interaction between X and Ω_1, the computation consists of two parts: the recursive evaluation of $\{\delta_t^{AB}\}$ and the final evaluation of $u^{\Omega_1}(x)$. The recursive part takes $O(q^3 N^2 \log N)$ since there are at most $O(N^2 \log N)$ pairs of squares (A, B) and the evaluation of $\{\delta_t^{AB}\}$ for each pair takes $O(q^3)$ steps via dimension-wise Chebyshev interpolation. The final evaluation of $u^{\Omega_1}(x)$ clearly takes $O(q^2 N^2)$ steps as we spend $O(q^2)$ on each point $x \in X$.

For the interaction between X and Ω_j, the analysis is similar. The recursive part now takes $O(q^3 N_j^2 \log N_j)$ steps (with $N_j = N/2^{j-1}$) as there are at most $O(N_j^2 \log N_j)$ pairs of squares involved. The final evaluation still takes $O(q^2 N^2)$ steps.

Summing these contributions together results in the total computational complexity

$$O(q^3 N^2 \log N) + O(q^2 N^2 \log N) = O(q^3 N^2 \log N) = O(r_\epsilon^{3/2} N^2 \log N).$$

The multiscale butterfly algorithm is also highly efficient in terms of memory as the Cartesian butterfly algorithm is applied sequentially to evaluate (9.30) for each Ω_j. The overall memory complexity is $O(\frac{N^2}{b^2})$, only $\frac{1}{b^2}$ of that the original Cartesian butterfly algorithm.

9.5 Numerical Results

This section presents several numerical examples to demonstrate the effectiveness of the multiscale butterfly algorithm introduced above. In truth, FIOs usually have nonconstant amplitude functions. Nevertheless, the main computational difficulty is the oscillatory phase term. We refer to [22] for detailed fast algorithms to deal with nonconstant amplitude functions. Our MATLAB implementation can be found on the authors' personal homepages. The numerical results were obtained on a desktop with a 3.5 GHz CPU and 32 GB of memory. Let $\{u^d(x), x \in X\}$, $\{u^m(x), x \in X\}$ and $\{u^p(x), x \in X\}$ be the results of a discrete FIO computed by a direct matrix-vector multiplication, the multiscale butterfly algorithm and the polar butterfly algorithm [22], respectively. To report on the accuracy, we randomly select a set S of 256 points from X and evaluate the relative errors of the multiscale butterfly algorithm and the polar butterfly algorithm by

$$\epsilon^m = \sqrt{\frac{\sum_{x\in S}|u^d(x) - u^m(x)|^2}{\sum_{x\in S}|u^d(x)|^2}} \text{ and } \epsilon^p = \sqrt{\frac{\sum_{x\in S}|u^d(x) - u^p(x)|^2}{\sum_{x\in S}|u^d(x)|^2}}. \tag{9.33}$$

According to the description of the multiscale butterfly algorithm in Section 9.4, we recursively divide Ω into $\Omega_j, j = 1, 2, \ldots, \log N - s$, where s is 5 in the following examples. This means that the center square Ω_d is of size $2^5 \times 2^5$ and the interaction from Ω_d is evaluated via a direct matrix-vector multiplication. Suppose q_ϵ is the number of Chebyshev points in each dimension. There is no sense in using butterfly algorithms to construct $\{\delta_t^{AB}\}$ when the number of points in B is fewer than q_ϵ^2. Hence, the recursion step in butterfly algorithms starts from the squares B that are a couple of levels away from the bottom of T_Ω such that each square contains at least q_ϵ^2 points. Similarly, the recursion stops at the squares in T_X that are the same number of levels away from the bottom. In the following examples, we start from level $\log N - 3$ and stop at level 3 (corresponding to $b = 2^3$ defined in Section 9.4) which matches with q_ϵ (4 to 11).

In order to make a fair comparison, we compare the MATLAB versions of the polar butterfly algorithm and the multiscale butterfly algorithm. Hence, the running time of the polar butterfly algorithm here is slower than that in [22], which was implemented in C++.

Example 1. The first example is a generalized Radon transform whose kernel is given by

$$\begin{aligned} \Phi(x,\xi) &= x \cdot \xi + \sqrt{c_1^2(x)\xi_1^2 + c_2^2(x)\xi_2^2}, \\ c_1(x) &= (2 + \sin(2\pi x_1)\sin(2\pi x_2))/3, \\ c_2(x) &= (2 + \cos(2\pi x_1)\cos(2\pi x_2))/3. \end{aligned} \tag{9.34}$$

We assume the amplitude of this example is a constant 1. Now the FIO models an integration over ellipses where $c_1(x)$ and $c_2(x)$ are the axis lengths of the ellipse centered at the point $x \in X$. Table 9.1 summarize the results of this example given by the polar butterfly algorithm and the

multiscale butterfly algorithm.

Multiscale Butterfly			Polar Butterfly			
N, q_ϵ	ϵ^m	$T_m(sec)$	N, q_ϵ	ϵ^p	$T_p(sec)$	T_p/T_m
256,5	7.89e-02	6.96e+01	256,5	4.21e-02	4.84e+02	6.96e+00
512,5	9.01e-02	3.62e+02	512,5	5.54e-02	2.34e+03	6.46e+00
1024,5	9.13e-02	1.81e+03	1024,5	4.26e-02	1.14e+04	6.31e+00
2048,5	9.47e-02	8.79e+03	2048,5	-	-	-
256,7	6.95e-03	8.20e+01	256,7	5.66e-03	5.97e+02	7.28e+00
512,7	8.43e-03	4.16e+02	512,7	5.89e-03	2.82e+03	6.79e+00
1024,7	8.45e-03	2.03e+03	1024,7	4.84e-03	1.35e+04	6.64e+00
2048,7	8.42e-03	1.04e+04	2048,7	-	-	-
256,9	3.90e-04	1.10e+02	256,9	8.25e-04	7.74e+02	7.04e+00
512,9	3.42e-04	5.39e+02	512,9	6.78e-04	3.57e+03	6.61e+00
1024,9	7.61e-04	2.74e+03	1024,9	4.18e-04	1.67e+04	6.09e+00
2048,9	4.82e-04	1.25e+04	2048,9	-	-	-
256,11	2.15e-05	1.84e+02	256,11	3.69e-05	1.15e+03	6.27e+00
512,11	1.89e-05	8.60e+02	512,11	5.53e-05	5.10e+03	5.93e+00
1024,11	1.96e-05	4.27e+03	1024,11	2.042e-05	2.30e+04	5.39e+00
2048,11	1.50e-05	1.82e+04	2048,11	-	-	-

Table 9.1: Comparison of the multiscale butterfly algorithm and the polar butterfly algorithm for the phase function in (9.34). T_m is the running time of the multiscale butterfly algorithm; T_a is the running time of the polar butterfly algorithm; and T_m/T_p is the speedup factor.

Example 2. Next, we provide an FIO example with a smooth amplitude function,

$$u(x) = \sum_{\xi\in\Omega} a(x,\xi) e^{2\pi \boldsymbol{i}\Phi(x,\xi)} \widehat{f}(\xi), \tag{9.35}$$

where the amplitude and phase functions are given by

$$\begin{aligned} a(x,\xi) &= (J_0(2\pi\rho(x,\xi)) + \boldsymbol{i}Y_0(2\pi\rho(x,\xi)))e^{-\pi\boldsymbol{i}\rho(x,\xi)}, \\ \Phi(x,\xi) &= x\cdot\xi + \rho(x,\xi), \\ \rho(x,\xi) &= \sqrt{c_1^2(x)\xi_1^2 + c_2^2(x)\xi_2^2}, \\ c_1(x) &= (2+\sin(2\pi x_1)\sin(2\pi x_2))/3, \\ c_2(x) &= (2+\cos(2\pi x_1)\cos(2\pi x_2))/3. \end{aligned}$$

Here, J_0 and Y_0 are Bessel functions of the first and second kinds. We refer the reader to [21] for more details of the derivation of these formulas. As discussed in [22], we compute the low-rank approximation of the amplitude functions $a(x,\xi)$ first:

$$a(x,\xi) \approx \sum_{t=1}^{s_\epsilon} g_t(x) h_t(\xi).$$

In the second step, we apply the multiscale butterfly algorithm to compute

$$u_t(x) = \sum_{\xi \in \Omega} e^{2\pi \boldsymbol{i} \Phi(x,\xi)} \widehat{f}(\xi) h_t(\xi),$$

and sum up all $g_t(x)u_t(x)$ to evaluate

$$u(x) = \sum_t g_t(x) u_t(x).$$

Table 9.2 summarizes the results of this example given by the direct method and the multiscale butterfly algorithm.

N, q_ϵ	ϵ^m	$T_d(sec)$	$T_m(sec)$	T_d/T_m
256,7	5.10e-03	3.78e+03	6.07e+02	6.23e+00
512,7	7.29e-03	3.71e+04	3.50e+03	1.06e+01
1024,7	6.16e-03	6.42e+05	1.70e+04	3.77e+01
256,9	4.49e-04	2.34e+03	7.88e+02	2.97e+00
512,9	4.04e-04	3.66e+04	4.64e+03	7.90e+00
1024,9	3.88e-04	6.21e+05	2.17e+04	2.86e+01
256,11	1.86e-05	2.48e+03	1.33e+03	1.86e+00
512,11	1.80e-05	3.60e+04	6.94e+03	5.18e+00
1024,11	2.39e-05	5.96e+05	2.83e+04	2.11e+01

Table 9.2: Numerical results given by the multiscale butterfly algorithm for the FIO in (9.35). T_d is the running time of the direct evaluation; T_m is the running time of the multiscale butterfly algorithm; and T_d/T_m is the speedup factor.

Note that the accuracy of the multiscale butterfly algorithm is well controlled by the number of Chebyshev points q_ϵ. This indicates that our algorithm is numerically stable. Another observation is that the relative error improves on average by a factor of 12 every time q_ϵ is increased by a factor of 2. As we can see in those tables, for a fixed kernel and a fixed q_ϵ, the accuracy is almost independent of N. Hence, in practical applications, one can increase the value of q_ϵ until a desired accuracy is reached in the problem with a small N. In the comparison in Table 9.1, the multiscale butterfly algorithm and the polar butterfly algorithm use $q_\epsilon = \{5, 7, 9, 11\}$ and achieve comparable accuracy. Meanwhile, as we observed from Table 9.1, the relative error decreasing rate of the multiscale butterfly algorithm is larger than the decreasing rate of the polar butterfly algorithm. This means if a high accuracy is desired, the multiscale butterfly algorithm requires a smaller q_ϵ to achieve it comparing to the polar butterfly algorithm.

The second concern about the algorithm is the asymptotic complexity. From the T_m column of Table 9.1 and 9.2, we see that T_m almost quadrupled when the problem size doubled under the same q_ϵ. According to this, we are convinced that the empirical running time of the multiscale butterfly algorithm follows the $O\left(N^2 \log N\right)$ asymptotic complexity. Note that the speedup factor over the

polar butterfly algorithm is about 6 and the multiscale butterfly algorithm obtains better accuracy. This makes the multiscale butterfly algorithm quite attractive to practitioners who are interested in evaluating an FIO with a large N.

Example 3. Extending the multiscale butterfly algorithm to higher dimensions is straightforward. There are two main modifications: higher-dimensional multiscale domain decomposition and Chebyshev interpolation. In three dimensions, the frequency domain is decomposed into cubic shells instead of coronas. The kernel interpolation is applied on a three-dimensional Chebyshev grids. We apply our three-dimensional multiscale butterfly algorithm to a simple example integrating over spheres with different radii. We assume a constant amplitude function and the kernel function is given by

$$\Phi(x,\xi) = x\cdot\xi + c(x)\sqrt{\xi_1^2+\xi_2^2}, \quad c(x) = (3+\sin(2\pi x_1)\sin(2\pi x_2)\sin(2\pi x_3))/4. \tag{9.36}$$

Table 9.3 summarizes the results of this example given by the direct method and the multiscale butterfly algorithm.

N, q_ϵ	ϵ^m	$T_d(sec)$	$T_m(sec)$	T_d/T_m
64,5	9.41e-02	1.82e+04	2.50e+03	7.31e+00
128,5	7.57e-02	6.21e+05	2.42e+04	2.57e+01
256,5	8.23e-02	3.91e+07	2.35e+05	1.66e+02
64,7	1.20e-02	1.83e+04	7.32e+03	2.50e+00
128,7	1.03e-02	6.03e+05	4.48e+04	1.35e+01
256,7	8.13e-03	4.39e+07	3.81e+05	1.15e+02

Table 9.3: Numerical results given by the multiscale butterfly algorithm for the phase function in (9.36).

9.6 Conclusion

A simple and efficient multiscale butterfly algorithm for evaluating FIOs is introduced in this chapter. This method hierarchically decomposes the frequency domain into multiscale coronas in order to avoid possible singularity of the phase function $\Phi(x,\xi)$ at $\xi = 0$. A Cartesian butterfly algorithm is applied to evaluate the FIO over each corona. Many drawbacks of the original butterfly algorithm based on a polar-Cartesian transform in [22] can be avoided. The new multiscale butterfly algorithm has a quasilinear operation complexity with a smaller prefactor, while it keeps the same linear memory complexity. This algorithm can be extended to other integral operators that have a complementary low-rank kernel $K(x,\xi)$ or $K(x,\xi)$ is complementary low-rank away from $\xi = 0$.

Chapter 10

One-Dimensional Butterfly Factorization

10.1 Introduction

10.1.1 Complementary Low-Rank Matrices and Butterfly Algorithm

This chapter is concerned with one-dimensional complementary low-rank matrices. For such a matrix, the rows are typically indexed by a set of points, say X, and the columns by another set of points, say Ω. Both X and Ω are often point sets in $\mathbb{R}^1$. Associated with X and Ω are two trees T_X and T_Ω, respectively and both trees are assumed to have the same depth $L = O(\log N)$, with the top level being level 0 and the bottom one being level L. Recall that such a matrix K of size $N \times N$ is said to satisfy the complementary low-rank property if for any level ℓ, any node A in T_X at level ℓ, and any node B in T_Ω at level $L - \ell$, the submatrix $K_{A,B}$, obtained by restricting K to the rows indexed by the points in A and the columns indexed by the points in B, is numerically low-rank, i.e., for a given precision ϵ there exists a low-rank approximation of $K_{A,B}$ with the 2-norm error bounded by ϵ and the rank bounded polynomially in $\log N$ and $\log(1/\epsilon)$. In many applications, one can even show that the rank is only bounded polynomially in $\log(1/\epsilon)$ and is independent of N. While it is straightforward to generalize the concept of the complementary low-rank property to a matrix with different row and column dimensions, the following discussion is restricted to the square matrices for simplicity.

A simple yet important example is the Fourier matrix K of size $N \times N$, where

$$X = \Omega = \{0, \ldots, N-1\},$$
$$K = \left(\exp(2\pi\imath jk/N)\right)_{0 \le j,k < N}.$$

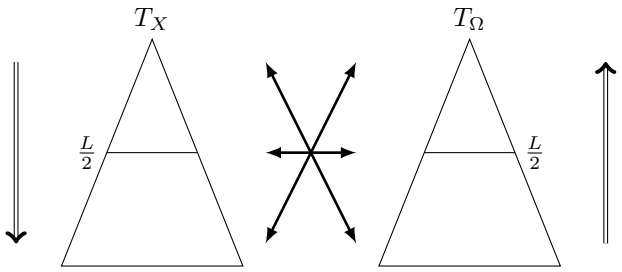


Figure 10.1: Trees of the row and column indices. Left: T_X for the row indices X. Right: T_Ω for the column indices Ω. The interaction between $A \in T_X$ and $B \in T_\Omega$ starts at the root of T_X and the leaves of T_Ω.

Here the trees T_X and T_Ω are generated by bisecting the sets X and Ω recursively. Both trees have the same depth $L = \log_2 N$. For each pair of nodes $A \in T_X$ and $B \in T_\Omega$ with A at level ℓ and B at level $L - \ell$, the numerical rank of the submatrix $K_{A,B}$ for a fixed precision ϵ is bounded by a number that is independent of N and scales linearly with respect to $\log(1/\epsilon)$ [138].

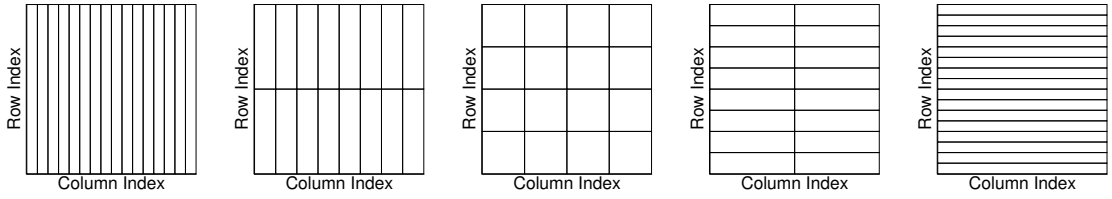


Figure 10.2: Hierarchical decomposition of the row and column indices of a 16×16 matrix. The trees T_X and T_Ω have roots containing 16 column and row indices and leaves containing a single column and row index. The rectangles above indicate the submatrices satisfying the complementary low-rank property.

For complementary low-rank matrices, the matrix-vector multiplication can be carried out efficiently via the butterfly algorithm, which was initially proposed in [132] and later extended in [138]. For a general matrix K of this type, the butterfly algorithm consists of two stages: the off-line stage and the on-line stage. In the off-line stage, it conducts simultaneously a top down traversal of T_X and a bottom up traversal of T_Ω (see Figure 10.1 for an interpretation of data flows) to recursively compress all complementary low-rank submatrices (see Figure 10.2 for an example of necessary submatrices). This typically takes $O(N^2)$ operations [138, 146] for a general complementary low-rank matrix K. In the on-line stage, the butterfly algorithm then evaluates $u = Kg$ for a given input vector $g \in \mathbb{C}^N$ in $O(N \log N)$ operations. While the on-line application cost is essentially linear, the $O(N^2)$ off-line precomputation cost appears to be a major bottleneck for many calculations. For certain complementary low-rank matrices, such as the ones obtained from the Fourier integral operators (FIOs) [22, 119, 141], the sparse Fourier transforms [188], and the numerical solutions of

acoustic wave equations [53], the off-line precomputation cost can be reduced to nearly linear or even totally eliminated. However, in all these cases, the reduction heavily relies on strong assumptions on the analytic properties of the kernel function of K. When such detailed information is not available, we are then forced to fall back on the $O(N^2)$ off-line precomputation algorithm.

10.1.2 Motivations and Significance

A natural question is whether it is still possible to reduce the cost of the precomputation stage if the analytic properties of the kernel are not accessible. The following two cases are quite common in applications:

1. Only black-box routines for computing Kg and K^*g in $O(N \log N)$ operations are given.
2. Only a black-box routine for evaluating any entry of the matrix K in $O(1)$ operations is given.

To answer this question, this chapter introduces the **butterfly factorization**, which represents K as a product of $L + 3$ sparse matrices:

$$K \approx U^L G^{L-1} \cdots G^h M^h (H^h)^* \cdots (H^{L-1})^* (V^L)^*, \tag{10.1}$$

where the depth $L = O(\log N)$ of T_X and T_Ω is assumed to be even, $h = L/2$ is a middle level index, and all factors are sparse matrices with $O(N)$ nonzero entries. This is joint work with Yingzhou Li, Eileen R. Martin, Kenneth L. Ho and Lexing Ying in [117].

The construction of the butterfly factorization proceeds as follows in two stages. The first stage is to construction a preliminary middle level factorization that is associated with the middle level of T_X and T_Ω

$$K \approx U^h M^h (V^h)^*, \tag{10.2}$$

where U^h and V^h are block diagonal matrices and M^h is a weighted permutation matrix. In the first case, this is achieved by applying K to a set of $O(N^{1/2})$ structured random vectors and then applying the randomized singular value decomposition (SVD) to the result. This typically costs $O(N^{3/2} \log N)$ operations. In the second case, (10.2) is built via the randomized sampling method proposed in [66, 181] for computing approximate SVDs. This randomized sampling needs to make the assumption that the columns and rows of middle level blocks of K to be incoherent with respect to the delta functions and it typically takes only $O(N^{3/2})$ operations in practice.

Once the middle level factorization (10.2) is available, the second stage is a sequence of truncated SVDs that further factorize each of U^h and V^h into a sequence of sparse matrices, resulting in the final factorization (10.1). The operation count of this stage is $O(N^{3/2})$ and the total memory complexity for constructing butterfly factorization is $O(N^{3/2})$.

When the butterfly factorization (10.1) is constructed, the cost of applying K to a given vector $g \in \mathbb{C}^N$ is $O(N \log N)$ because (10.1) is a sequence of $O(\log N)$ sparse matrices, each with $O(N)$

non-zero entries. Although we shall limit our discussion to one-dimensional problems in this chapter, the proposed butterfly factorization, along with its construction algorithm, can be easily generalized to higher dimensions.

This work is motivated by problems that require repeated applications of a butterfly algorithm. In several applications, such as inverse scattering [157, 186] and fast spherical harmonic transform (SHT) [158], the butterfly algorithm is called repeatedly either in an iterative process of minimizing some regularized objective function or to a large set of different input vectors. Therefore, it becomes important to reduce the constant prefactor of the butterfly algorithm to save actual runtime. For example in [22], Chebyshev interpolation is applied to recover low-rank structures of submatrices with a sufficiently large number of interpolation points. The recovered rank is far from the optimum. Hence, the prefactor of the corresponding butterfly algorithm in [22] is large. The butterfly factorization can further compress this butterfly algorithm to obtain nearly optimal low-rank approximations resulting in a much smaller prefactor, as will be shown in the numerical results. Therefore, it is more efficient to construct the butterfly factorization using this butterfly algorithm and then apply the butterfly factorization repeatedly. In this sense, the butterfly factorization can be viewed as a compression of certain butterfly algorithms.

Another important application is the computation of a composition of several FIOs. A direct method to construct the composition takes $O(N^3)$ operations, while the butterfly factorization provides a data-sparse representation of this composition in $O(N^{3/2} \log N)$ operations, once the fast algorithm for applying each FIO is available. After the construction, the application of the butterfly factorization is independent of the number of FIOs in the composition, which is significant when the number of FIOs is large.

Recently, there has also been a sequence of papers on recovering a structured matrix via applying it to (structured) random vectors. For example, the randomized SVD algorithms [83, 120, 169] recover a low-rank approximation to an unknown matrix when it is numerically low-rank. The work in [130] constructs a sparse representation for an unknown HSS matrix. More recently, [122] considers the more general problem of constructing a sparse representation of an unknown $\mathcal{H}$-matrix. To our best knowledge, the present work is the first to address such matrix recovery problem if the unknown matrix satisfies the complementary low-rank property.

10.1.3 Content

The rest of this chapter is organized as follows. Section 10.2 briefly reviews some basic tools that shall be used repeatedly in Sections 10.3. Section 10.3 describes in detail the butterfly factorization and its construction algorithm. In Section 10.4, numerical examples are provided to demonstrate the efficiency of the proposed algorithms. Finally, Section 10.5 lists several directions for future work.

10.2 Preliminaries

For a matrix $Z \in \mathbb{C}^{m\times n}$, we define a rank-$r$ approximate singular value decomposition (SVD) of Z as

$$Z \approx U_0 \Sigma_0 V_0^*,$$

where $U_0 \in \mathbb{C}^{m\times r}$ is unitary, $\Sigma_0 \in \mathbb{R}^{r\times r}$ is diagonal, and $V_0 \in \mathbb{C}^{n\times r}$ is unitary. A straightforward method to obtain the optimal rank-r approximation of Z is to compute its truncated SVD, where U_0 is the matrix with the first r left singular vectors, Σ_0 is a diagonal matrix with the first r singular values in decreasing order, and V_0 is the matrix with the first r right singular vectors.

A typical computation of the truncated SVD of Z takes $O(mn\min(m,n))$ operations, which can be quite expensive when m and n are large. Therefore, a lot of research has been devoted to faster algorithms for computing approximate SVDs, especially for matrices with fast decaying singular values. In Sections 10.2.1 and 10.2.2, we will introduce two randomized algorithms for computing approximate SVDs for numerically low-rank matrices Z: the first one [83] is based on applying the matrix to random vectors while the second one [66, 181] relies on sampling the matrix entries randomly.

Once an approximate SVD $Z \approx U_0\Sigma_0 V_0^*$ is computed, it can be written in several equivalent ways, each of which is convenient for certain purposes. First, one can write

$$Z \approx USV^*,$$

where

$$U = U_0\Sigma_0,\ S = \Sigma_0^{-1} \text{ and } V^* = \Sigma_0 V_0^*. \tag{10.3}$$

This construction is analogous to the well-known CUR decomposition [127] in the sense that the left and right factors in both factorization methods inherit similar singular values of the original numerical low-rank matrix. Here, the middle matrix S in (10.3) can be carefully constructed to ensure numerical stability, since the singular values in Σ_0 can be computed to nearly full relative precision.

As we shall see, sometimes it is also convenient to write the approximation as

$$Z \approx UV^*$$

where

$$U = U_0 \text{ and } V^* = \Sigma_0 V_0^*, \tag{10.4}$$

or

$$U = U_0\Sigma_0 \text{ and } V^* = V_0^*. \tag{10.5}$$

Here, one of the factors U and V share the singular values of Z.

10.2.1 SVD via Random Matrix-Vector Multiplication

One popular approach is the randomized algorithm in [83] that reduces the cubic complexity to $O(rmn)$ complexity. We briefly review this following [83] for constructing a rank-r approximation SVD $Z \approx U_0\Sigma_0 V_0^*$ below.

Algorithm 10.2.1. *Randomized SVD*

1. *Generate two tall skinny random Gaussian matrices $R_{col} \in \mathbb{C}^{n\times(r+p)}$ and $R_{row} \in \mathbb{C}^{m\times(r+p)}$, where $p = O(1)$ is an additive oversampling parameter that increases the approximation accuracy.*

2. *Apply the pivoted QR factorization to ZR_{col} and let Q_{col} be the matrix of the first r columns of the Q matrix. Similarly, apply the pivoted QR factorization to Z^*R_{row} and let Q_{row} be the matrix of the first r columns of the Q matrix.*

3. *Generate a tiny middle matrix $M = Q_{col}^* Z Q_{row}$ and compute its rank-r truncated SVD: $M \approx U_M\Sigma_M V_M^*$.*

4. *Let $U_0 = Q_{col}U_M$, $\Sigma_0 = \Sigma_M$, and $V_0^* = V_M^* Q_{row}^*$. Then $Z \approx U_0\Sigma_0 V_0^*$.*

The dominant complexity comes from the application of Z to $O(r)$ random vectors. If fast algorithms for applying Z are available, the quadratic complexity can be further reduced.

Once the approximate SVD of Z is ready, the equivalent forms in (10.3), (10.4), and (10.5) can be constructed easily. Under the condition that the singular values of Z decay sufficiently rapidly, the approximation error of the resulting rank-r is nearly optimal with an overwhelming probability. Typically, the additive over-sampling parameter $p = 5$ is sufficient to obtain an accurate rank-r approximation of Z.

For most applications, the goal is to construct a low-rank approximation up to a fixed relative precision ϵ, rather than a fixed rank r. The above procedure can then be embedded into an iterative process that starts with a relatively small r, computes a rank-r approximation, estimates the error probabilistically, and repeats the steps with doubled rank $2r$ if the error is above the threshold ϵ [83].

10.2.2 SVD via Random Sampling

The above algorithm relies only on the product of the matrix $Z \in \mathbb{C}^{m\times n}$ or its transpose with given random vectors. If one is allowed to access the individual entries of Z, the following randomized sampling method for low-rank approximations introduced in [66, 181] can be more efficient. This

method only visits $O(r)$ columns and rows of Z and hence only requires $O(r(m+n))$ operations and memory.

Here, we adopt the standard notation for a submatrix: given a row index set I and a column index set J, $Z_{I,J} = Z(I, J)$ is the submatrix with entries from rows in I and columns in J; we also use " : " to denote the entire columns or rows of the matrix, i.e., $Z_{I,:} = Z(I, :)$ and $Z_{:,J} = Z(:, J)$. With these handy notations, we briefly introduce the randomized sampling algorithm to construct a rank-r approximation of $Z \approx U_0\Sigma_0 V_0^*$.

Algorithm 10.2.2. *Randomized sampling for low-rank approximation*

1. *Let Π_{col} and Π_{row} denote the important columns and rows of Z that are used to form the column and row bases. Initially $\Pi_{col} = \emptyset$ and $\Pi_{row} = \emptyset$.*

2. *Randomly sample rq rows and denote their indices by S_{row}. Let $I = S_{row} \cup \Pi_{row}$. Here $q = O(1)$ is a multiplicative oversampling parameter. Perform a pivoted QR decomposition of $Z_{I,:}$ to get*

$$Z_{I,:}P = QR,$$

where P is the resulting permutation matrix and $R = (r_{ij})$ is an $O(r) \times n$ upper triangular matrix. Define the important column index set Π_{col} to be the first r columns picked within the pivoted QR decomposition.

3. *Randomly sample rq columns and denote their indices by S_{col}. Let $J = S_{col} \cup \Pi_{col}$. Perform a pivoted LQ decomposition of $Z_{:,J}$ to get*

$$PZ_{:,J} = LQ,$$

where P is the resulting permutation matrix and $L = (l_{ij})$ is an $m \times O(r)$ lower triangular matrix. Define the important row index set Π_{row} to be the first r rows picked within the pivoted LQ decomposition.

4. *Repeat steps 2 and 3 a few times to ensure Π_{col} and Π_{row} sufficiently sample the important columns and rows of Z.*

5. *Apply the pivoted QR factorization to $Z_{:,\Pi_{col}}$ and let Q_{col} be the matrix of the first r columns of the Q matrix. Similarly, apply the pivoted QR factorization to $Z^*_{\Pi_{row},:}$ and let Q_{row} be the matrix of the first r columns of the Q matrix.*

6. *We seek a middle matrix M such that $Z \approx Q_{col}MQ^*_{row}$. To solve this problem efficiently, we approximately reduce it to a least-squares problem of a smaller size. Let S_{col} and S_{row} be the index sets of a few extra randomly sampled columns and rows. Let $J = \Pi_{col} \cup S_{col}$ and*

$I = \Pi_{row} \cup S_{row}$. *A simple least-squares solution to the problem*

$$\min_M \|Z_{I,J} - (Q_{col})_{I,:} M (Q_{row}^*)_{:,J}\|$$

gives $M = (Q_{col})_{I,:}^{\dagger} Z_{I,J} (Q_{row}^*)_{:,J}^{\dagger}$, *where* $(\cdot)^{\dagger}$ *stands for the pseudo-inverse.*

7. *Compute an SVD* $M \approx U_M \Sigma_M V_M^*$. *Then the low-rank approximation of* $Z \approx U_0 S_0 V_0^*$ *is given by*

$$U_0 = Q_{col} U_M, \quad \Sigma_0 = \Sigma_M, \quad V_0^* = V_M^* Q_{row}^*. \tag{10.6}$$

We have not been able to quantify the error and success probability rigorously for this procedure at this point. On the other hand, when the columns and rows of K are incoherent with respect to "delta functions" (i.e., vectors that have only one significantly larger entry), this procedure works well in our numerical experiments. Here, a vector u is said to be incoherent with respect to a vector v if $\mu = |u^T v|/(\|u\|_2 \|v\|_2)$ is small. In the typical implementation, the multiplicative oversampling parameter q is equal to 3 and Steps 2 and 3 are iterated no more than three times. These parameters are empirically sufficient to achieve accurate low-rank approximations and are used throughout numerical examples in Section 10.4.

As we mentioned above, for most applications the goal is to construct a low-rank approximation up to a fixed relative error ϵ, rather than a fixed rank. This process can also be embedded into an iterative process to achieve the desired accuracy.

10.3 Butterfly Factorization

This section presents the butterfly factorization algorithm for a matrix $K \in \mathbb{C}^{N \times N}$. For simplicity let $X = \Omega = \{1, \ldots, N\}$. The trees T_X and T_Ω are complete binary trees with $L = \log_2 N - O(1)$ levels. We assume that L is an even integer and the number of points in each leaf node of T_X and T_Ω is bounded by a uniform constant.

At each level ℓ, $\ell = 0, \ldots, L$, we denote the ith node at level ℓ in T_X as A_i^ℓ for $i = 0, 1, \ldots, 2^\ell - 1$ and the jth node at level $L - \ell$ in T_Ω as $B_j^{L-\ell}$ for $j = 0, 1, \ldots, 2^{L-\ell} - 1$. These nodes naturally partition K into $O(N)$ submatrices $K_{A_i^\ell, B_j^{L-\ell}}$. For simplicity, we write $K_{i,j}^\ell := K_{A_i^\ell, B_j^{L-\ell}}$, where the superscript is used to indicate the level (in T_X). The butterfly factorization utilizes rank-r approximations of all submatrices $K_{i,j}^\ell$ with $r = O(1)$.

The butterfly factorization of K is built in two stages. In the first stage, we compute a rank-r approximations of each submatrix $K_{i,j}^h$ at the level $\ell = h = L/2$ and then organize them into an initial factorization:

$$K \approx U^h M^h (V^h)^*,$$

where U^h and V^h are block diagonal matrices and M^h is a weighted permutation matrix. This is

referred as the **middle level factorization** and is described in detail in Section 10.3.1.

In the second stage, we recursively factorize $U^\ell \approx U^{\ell+1}G^\ell$ and $(V^\ell)^* \approx (H^\ell)^*(V^{\ell+1})^*$ for $\ell = h, h+1, \ldots, L-1$, since U^ℓ and $(V^\ell)^*$ inherit the complementary low-rank property from K, i.e., the low-rank property of U^ℓ comes from the low-rank property of $K^\ell_{i,j}$ and the low-rank property of V^ℓ results from the one of $K^{L-\ell}_{i,j}$. After this recursive factorization, one reaches at the butterfly factorization of K

$$K \approx U^L G^{L-1} \cdots G^h M^h (H^h)^* \cdots (H^{L-1})^* (V^L)^*, \tag{10.7}$$

where all factors are sparse matrices with $O(N)$ nonzero entries. We refer to this stage as the **recursive factorization** and it is discussed in detail in Section 10.3.2.

10.3.1 Middle Level Factorization

The first step of the middle level factorization is to compute a rank-r approximation to every $K^h_{i,j}$. Recall that we consider one of the following two cases.

1. Only black-box routines for computing Kg and K^*g in $O(N \log N)$ operations are given.
2. Only a black-box routine for evaluating any entry of the matrix K in $O(1)$ operations is given.

The actual computation of this step proceeds differently depending on which case is under consideration. Through the discussion, $m = 2^h = O(N^{1/2})$ is the number of nodes in the middle level $h = L/2$ and we assume without loss of generality that N/m is an integer.

- In the first case, the rank-r approximation of each $K^h_{i,j}$ is constructed with the SVD algorithm via random matrix-vector multiplication in Section 10.2.1. This requires us to apply $K^h_{i,j}$ and its adjoint to random Gaussian matrices of size $(N/m) \times (r+p)$, where r is the desired rank and p is an oversampling parameter. In order to take advantage of the fast algorithm for multiplying K, we construct a matrix C of size $N \times m(r+p)$. C is partitioned into an $m \times m$ blocks with each block C_{ij} for $i, j = 0, 1, \ldots, m-1$ of size $(N/m) \times (r+p)$. In additional, C is block-diagonal and its diagonal blocks are random Gaussian matrices. This is equivalent to applying each $K^h_{i,j}$ to the same random Gaussian matrix C_{jj} for all i. We then use the fast algorithm to apply K to each column of C and store the results. Similarly, we form another random block diagonal matrix R similar to C and use the fast algorithm of applying K^* to R. This is equivalent to applying each $(K^h_{i,j})^*$ to an $(N/m) \times (r+p)$ Gaussian random matrix R_{ii} for all $j = 0, 1, \ldots, m-1$. With $K^h_{i,j}C_{jj}$ and $(K^h_{i,j})^*R_{ii}$ ready, we can compute the rank-r approximate SVD of $K^h_{i,j}$ following the procedure described in Section 10.2.1.
- In the second case, it is assumed that an arbitrary entry of K can be calculated in $O(1)$ operations. We simply apply the SVD algorithm via random sampling in Section 10.2.2 to each $K^h_{i,j}$ to construct a rank-r approximate SVD.

In either case, once the approximate SVD of $K^h_{i,j}$ is ready, it is transformed in the form

$$K^h_{i,j} \approx U^h_{i,j} S^h_{i,j} (V^h_{j,i})^*$$

following (10.3). We would like to emphasize that the columns of $U^h_{i,j}$ and $V^h_{j,i}$ are scaled with the singular values of the approximate SVD so that they keep track of the importance of these columns in approximating $K^h_{i,j}$.

After calculating the approximate rank-r factorization of each $K^h_{i,j}$, we assemble these factors into three block matrices U^h, M^h and V^h as follows:

$$\begin{aligned} K \approx & \begin{pmatrix} U^h_{0,0} S^h_{0,0} (V^h_{0,0})^* & U^h_{0,1} S^h_{0,1} (V^h_{1,0})^* & \cdots & U^h_{0,m-1} S^h_{0,m-1} (V^h_{m-1,0})^* \\ U^h_{1,0} S^h_{1,0} (V^h_{0,1})^* & U^h_{1,1} S^h_{1,1} (V^h_{1,1})^* & & U^h_{1,m-1} S^h_{1,m-1} (V^h_{m-1,1})^* \\ \vdots & & \ddots & \\ U^h_{m-1,0} S^h_{m-1,0} (V^h_{0,m-1})^* & U^h_{m-1,1} S^h_{m-1,1} (V^h_{1,m-1})^* & & U^h_{m-1,m-1} S^h_{m-1,m-1} (V^h_{m-1,m-1})^* \end{pmatrix} \\ = & \begin{pmatrix} U^h_0 & & & \\ & U^h_1 & & \\ & & \ddots & \\ & & & U^h_{m-1} \end{pmatrix} \begin{pmatrix} M^h_{0,0} & M^h_{0,1} & \cdots & M^h_{0,m-1} \\ M^h_{1,0} & M^h_{1,1} & & M^h_{1,m-1} \\ \vdots & & \ddots & \\ M^h_{m-1,0} & M^h_{m-1,1} & & M^h_{m-1,m-1} \end{pmatrix} \begin{pmatrix} V^h_0 & & & \\ & V^h_1 & & \\ & & \ddots & \\ & & & V^h_{m-1} \end{pmatrix}^* \\ = & U^h M^h (V^h)^*, \end{aligned} \tag{10.8}$$

where

$$U^h_i = \begin{pmatrix} U^h_{i,0} & U^h_{i,1} & \cdots & U^h_{i,m-1} \end{pmatrix} \in \mathbb{C}^{(N/m) \times mr}, \quad V^h_j = \begin{pmatrix} V^h_{j,0} & V^h_{j,1} & \cdots & V^h_{j,m-1} \end{pmatrix} \in \mathbb{C}^{(N/m) \times mr}, \tag{10.9}$$

and $M^h \in \mathbb{C}^{(m^2 r) \times (m^2 r)}$ is a weighted permutation matrix. Each submatrix $M^h_{i,j}$ is itself an $m \times m$ block matrix with block size $r \times r$ where all blocks are zero except that the (j, i) block is equal to the diagonal matrix $S^h_{i,j}$. It is obvious that there are only $O(N)$ nonzero entries in M^h. See Figure 10.3 for an example of a middle level factorization of a 64×64 matrix with $r = 1$.

10.3.2 Recursive Factorization

In this section, we will recursively factorize

$$U^\ell \approx U^{\ell+1} G^\ell \tag{10.10}$$

for $\ell = h, h+1, \ldots, L-1$ and

$$(V^\ell)^* \approx (H^\ell)^* (V^{\ell+1})^* \tag{10.11}$$

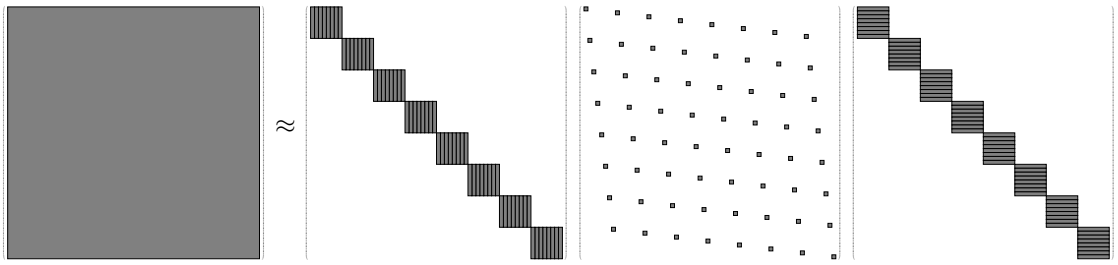


Figure 10.3: The middle level factorization of a 64×64 complementary low-rank matrix $K \approx U^3 M^3 (V^3)^*$ assuming $r = 1$. Grey blocks indicate nonzero blocks. U^3 and V^3 are block-diagonal matrices with 8 blocks. The diagonal blocks of U^3 and V^3 are assembled according to Equation (10.9) as indicated by black rectangles. M^3 is a 8×8 block matrix with each block $M^3_{i,j}$ itself an 8×8 block matrix containing diagonal weights matrix on the (j, i) block.

for $\ell = h, h+1, \ldots, L-1$. After these recursive factorizations, we can obtain the following butterfly factorization by substituting these factorizations into (10.8):

$$K \approx U^L G^{L-1} \cdots G^h M^h (H^h)^* \cdots (H^{L-1})^* (V^L)^*. \tag{10.12}$$

Recursive factorization of U^h

Each factorization at level ℓ in (10.10) results from the low-rank property of $K^\ell_{i,j}$ for $\ell \geq L/2$. When $\ell = h$, recall that

$$U^h = \begin{pmatrix} U^h_0 & & & \\ & U^h_1 & & \\ & & \ddots & \\ & & & U^h_{m-1} \end{pmatrix}$$

and

$$U^h_i = \begin{pmatrix} U^h_{i,0} & U^h_{i,1} & \cdots & U^h_{i,m-1} \end{pmatrix}$$

with each $U^h_{i,j} \in \mathbb{C}^{(N/m) \times r}$. We split U^h_i and each $U^h_{i,j}$ into halves by row, i.e.,

$$U^h_i = \begin{pmatrix} U^{h,t}_i \\ \hline U^{h,b}_i \end{pmatrix} \text{ and } U^h_{i,j} = \begin{pmatrix} U^{h,t}_{i,j} \\ \hline U^{h,b}_{i,j} \end{pmatrix},$$

where the superscript t denotes the top half and b denotes the bottom half of a matrix. Then we have

$$U^h_i = \left(\begin{array}{cccc} U^{h,t}_{i,0} & U^{h,t}_{i,1} & \cdots & U^{h,t}_{i,m-1} \\ \hline U^{h,b}_{i,0} & U^{h,b}_{i,1} & \cdots & U^{h,b}_{i,m-1} \end{array}\right). \tag{10.13}$$

Notice that, for each $i = 0, 1, \ldots, m-1$ and $j = 0, 1, \ldots, m/2-1$, the columns of

$$\begin{pmatrix} U_{i,2j}^{h,t} & U_{i,2j+1}^{h,t} \end{pmatrix} \text{ and } \begin{pmatrix} U_{i,2j}^{h,b} & U_{i,2j+1}^{h,b} \end{pmatrix} \tag{10.14}$$

in (10.13) are in the column space of $K_{2i,j}^{h+1}$ and $K_{2i+1,j}^{h+1}$, respectively. By the complementary low-rank property of the matrix K, $K_{2i,j}^{h+1}$ and $K_{2i+1,j}^{h+1}$ are numerical low-rank. Hence $\begin{pmatrix} U_{i,2j}^{h,t} U_{i,2j+1}^{h,t} \end{pmatrix}$ and $\begin{pmatrix} U_{i,2j}^{h,b} U_{i,2j+1}^{h,b} \end{pmatrix}$ are numerically low-rank matrices in $\mathbb{C}^{(N/2m)\times 2r}$. Compute their rank-$r$ approximations by the standard truncated SVD, transform it into the form of (10.5) and denote them as

$$\begin{pmatrix} U_{i,2j}^{h,t} & U_{i,2j+1}^{h,t} \end{pmatrix} \approx U_{2i,j}^{h+1} G_{2i,j}^{h} \text{ and } \begin{pmatrix} U_{i,2j}^{h,b} & U_{i,2j+1}^{h,b} \end{pmatrix} \approx U_{2i+1,j}^{h+1} G_{2i+1,j}^{h} \tag{10.15}$$

for $i = 0, 1, \ldots, m-1$ and $j = 0, 1, \ldots, m/2-1$. The matrices in (10.15) can be assembled into two new sparse matrices, such that

$$U^h \approx U^{h+1} G^h = \begin{pmatrix} U_0^{h+1} & & & \\ & U_1^{h+1} & & \\ & & \ddots & \\ & & & U_{2m-1}^{h+1} \end{pmatrix} \begin{pmatrix} G_0^h & & & \\ & G_1^h & & \\ & & \ddots & \\ & & & G_{m-1}^h \end{pmatrix},$$

where

$$U_i^{h+1} = \begin{pmatrix} U_{i,0}^{h+1} & U_{i,1}^{h+1} & \cdots & U_{i,m/2-1}^{h+1} \end{pmatrix}$$

for $i = 0, 1, \ldots, 2m-1$, and

$$G_i^h = \left(\begin{array}{cccc} G_{2i,0}^h & & & \\ & G_{2i,1}^h & & \\ & & \ddots & \\ & & & G_{2i,m/2-1}^h \\ \hline G_{2i+1,0}^h & & & \\ & G_{2i+1,1}^h & & \\ & & \ddots & \\ & & & G_{2i+1,m/2-1}^h \end{array}\right)$$

for $i = 0, 1, \ldots, m-1$.

Since there are $O(1)$ nonzero entries in each $G_{i,j}^h$ and there are $O(N)$ such submatrices, there are only $O(N)$ nonzero entries in G^h. See Figure 10.4 top for an example of the factorization $U^h \approx U^{h+1} G^h$ for the left factor U^h with $L = 6$, $h = 3$ and $r = 1$ in Figure 10.3.

Similarly, for any ℓ between h and $L-1$, we can factorize $U^\ell \approx U^{\ell+1} G^\ell$, because the columns

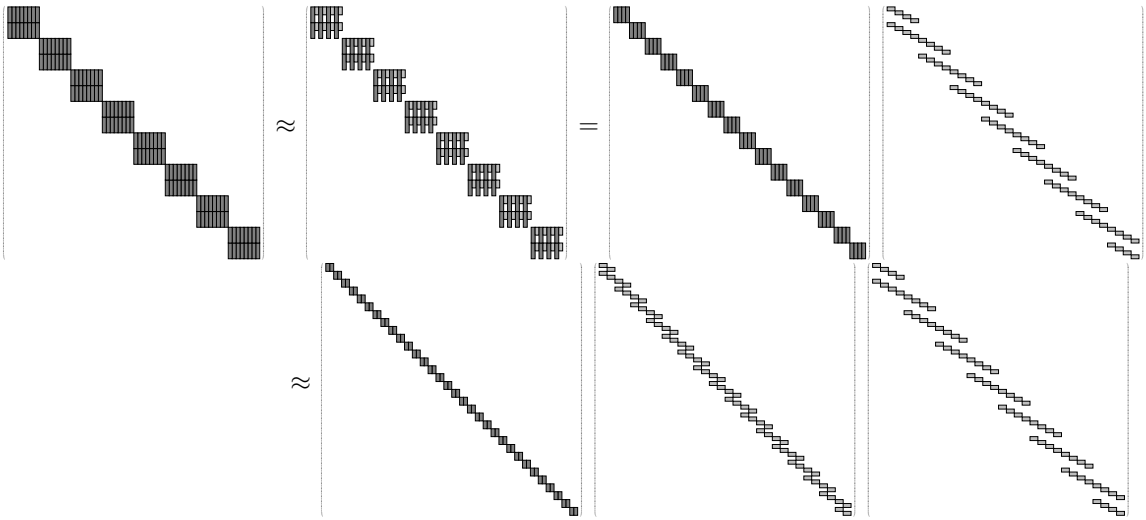


Figure 10.4: The recursive factorization of U^3 in Figure 10.3. Gray factors are matrices inheriting the low-rank property of the butterfly matrix. Top: left matrix: U^3 with each diagonal block partitioned into smaller blocks according to Equation (10.13) as indicated by black rectangles; middle-left matrix: low-rank approximations of submatrices in U^3 given by Equation (10.15); middle right matrix: U^4; right matrix: G^3. Bottom: U^4 in the first row is further factorized into $U^4 \approx U^5 G^4$, giving $U^3 \approx U^5 G^4 G^3$.

in $\left(U^{\ell,t}_{i,2j} U^{\ell,t}_{i,2j+1}\right)$ and $\left(U^{\ell,b}_{i,2j} U^{\ell,b}_{i,2j+1}\right)$ are in the column space of the numerically low-rank matrices $K^{\ell+1}_{2i,j}$ and $K^{\ell+1}_{2i+1,j}$, respectively. Computing the rank-r approximations via the standard truncated SVD and transforming them into the form of (10.5) give

$$\left(U^{\ell,t}_{i,2j} \quad U^{\ell,t}_{i,2j+1}\right) \approx U^{\ell+1}_{2i,j} G^{\ell}_{2i,j} \text{ and } \left(U^{\ell,b}_{i,2j} \quad U^{\ell,b}_{i,2j+1}\right) \approx U^{\ell+1}_{2i+1,j} G^{\ell}_{2i+1,j} \tag{10.16}$$

for $i = 0, 1, \ldots, 2^\ell - 1$ and $j = 0, 1, \ldots, 2^{L-\ell-1} - 1$. After assembling these factorizations together, we obtain

$$U^\ell \approx U^{\ell+1} G^\ell = \begin{pmatrix} U^{\ell+1}_0 & & & \\ & U^{\ell+1}_1 & & \\ & & \ddots & \\ & & & U^{\ell+1}_{2^{\ell+1}-1} \end{pmatrix} \begin{pmatrix} G^\ell_0 & & & \\ & G^\ell_1 & & \\ & & \ddots & \\ & & & G^\ell_{2^\ell-1} \end{pmatrix},$$

where

$$U^{\ell+1}_i = \begin{pmatrix} U^{\ell+1}_{i,0} & U^{\ell+1}_{i,1} & \cdots & U^{\ell+1}_{i,2^{L-\ell-1}-1} \end{pmatrix}$$

for $i = 0, 1, \ldots, 2^{\ell+1} - 1$, and

$$G_i^\ell = \left(\begin{array}{cccc} G_{2i,0}^\ell & & & \\ & G_{2i,1}^\ell & & \\ & & \ddots & \\ & & & G_{2i,2^{L-\ell-1}-1}^\ell \\ \hline G_{2i+1,0}^\ell & & & \\ & G_{2i+1,1}^\ell & & \\ & & \ddots & \\ & & & G_{2i+1,2^{L-\ell-1}-1}^\ell \end{array}\right)$$

for $i = 0, 1, \ldots, 2^\ell - 1$.

After $L - h$ steps of recursive factorizations

$$U^\ell \approx U^{\ell+1} G^\ell$$

for $\ell = h, h+1, \ldots, L-1$, we obtain the recursive factorization of U^h as

$$U^h \approx U^L G^{L-1} \cdots G^h. \tag{10.17}$$

See Figure 10.4 bottom for an example of a recursive factorization for the left factor U^h with $L = 6$, $h = 3$ and $r = 1$ in Figure 10.3.

Similar to the analysis of G^h, it is also easy to check that there are only $O(N)$ nonzero entries in each G^ℓ in (10.17). Since there are $O(N)$ diagonal blocks in U^L and each block contains $O(1)$ entries, there is $O(N)$ nonzero entries in U^L.

Recursive factorization of V^h

The recursive factorization of V^h is similar to the one of U^h. In each step of the factorization

$$(V^\ell)^* \approx (H^\ell)^* (V^{\ell+1})^*,$$

we take advantage of the low-rank property of the row space of $K_{i,2j}^{L-\ell-1}$ and $K_{i,2j+1}^{L-\ell-1}$ to obtain rank-r approximations. Applying the exact same procedure of Section 10.3.2 now to V^ℓ leads to the recursive factorization $V^h \approx V^L H^{L-1} \cdots H^h$, or equivalently

$$(V^h)^* \approx (H^h)^* \cdots (H^{L-1})^* (V^L)^*, \tag{10.18}$$

with all factors containing only $O(N)$ nonzero entries. See Figure 10.5 for an example of a recursive factorization $(V^h)^* \approx (H^h)^* \cdots (H^{L-2})^* (V^{L-1})^*$ for the left factor V^h with $L = 6$, $h = 3$ and $r = 1$ in Figure 10.3.

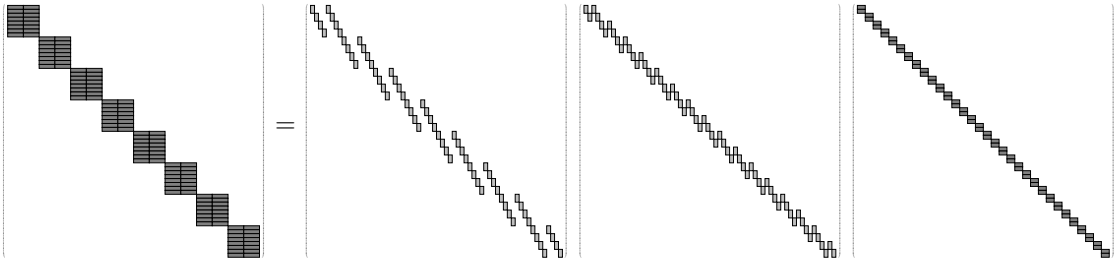


Figure 10.5: The recursive factorization $(V^3)^* \approx (H^3)^* (H^4)^* (V^5)^*$ of $(V^3)^*$ in Figure 10.3.

Given the recursive factorization of U^h and $(V^h)^*$ in (10.17) and (10.18), we reach the butterfly factorization

$$K \approx U^L G^{L-1} \cdots G^h M^h (H^h)^* \cdots (H^{L-1})^* (V^L)^*, \tag{10.19}$$

where all factors are sparse matrices with $O(N)$ nonzero entries. For a given input vector $g \in \mathbb{C}^N$, the $O(N^2)$ matrix-vector multiplication $u = Kg$ can be approximated by a sequence of $O(\log N)$ sparse matrix-vector multiplications given by the butterfly factorization.

10.3.3 Complexity Analysis

The complexity analysis of the construction of a butterfly factorization naturally consists of two parts: the middle level factorization and the recursive factorization.

The complexity of the middle level factorization depends on which one of the cases is under consideration.

- For the first case, the approximate SVDs are determined by the application of K and K^* to Gaussian random matrices in $\mathbb{C}^{N \times N^{1/2}(r+p)}$ and the rank-r approximations of K_{ij}^h for each (i, j) pair. Assume that each matrix-vector multiplication by K or K^* via the given black-box routines requires $O(C_K(N))$ operations (which is at least $O(N)$). Then the dominant cost is due to applying K and K^* $O(N^{1/2})$ times, which yields an overall computational complexity of $O(C_K(N)N^{1/2})$.

- In the second case, the approximate SVDs are computed via random sampling for each K_{ij}^h of the $O(N)$ pairs (i, j). The complexity of performing randomized sampling for each such block is $O(N^{1/2})$. Hence, the overall computational complexity is $O(N^{3/2})$.

In the recursive factorization, U^ℓ at level ℓ consists of $O(2^\ell)$ diagonal blocks of size $O(N/2^\ell) \times O(N/2^\ell)$. In each diagonal block, there are $O(N/2^\ell)$ factorizations in (10.16). Since the operation

complexity of performing one factorization in (10.16) is $O(N/2^\ell)$, it takes $O(N^2/2^\ell)$ operations to factorize U^ℓ. Summing up the operations at all levels gives the total complexity for recursively factorizing U^h:

$$\sum_{\ell=h}^{L-1} O(N^2/2^\ell) = O(N^{3/2}). \tag{10.20}$$

Similarly, the operation complexity for recursively compressing V^h is also $O(N^{3/2})$.

The memory peak of the butterfly factorization occurs in the middle level factorization since we have to store the initial factorization in (10.8). There are $O(N^{3/2})$ nonzero entries in U^h and V^h, and $O(N)$ in M^h. Hence, the total memory complexity is $O(N^{3/2})$. The total operation complexity for constructing the butterfly factorization is summarized in Table 10.1.

		Randomized SVD	Randomized sampling
Factorization Complexity	Middle level factorization	$O(C_K(N)N^{1/2})$	$O(N^{3/2})$
	Recursive factorization	$O(N^{3/2})$	
	Total	$O(C_K(N)N^{1/2})$	$O(N^{3/2})$
Memory Complexity		$O(N^{3/2})$	$O(N \log N)$
Application Complexity		$O(N \log N)$	

Table 10.1: Computational complexity and memory complexity of the butterfly factorization. $C_K(N)$ is the operation complexity of one application of K or K^*. In most of the cases encountered, $C_K(N) = O(N \log N)$.

It is worth pointing out that the memory complexity can be reduced to $O(N \log N)$, when we apply the randomized sampling method to construct each block in the initial factorization in (10.8) separately. Instead of factorizing U^h and V^h at the end of the middle level factorization, we can factorize the left and right factors U_i^h and V_i^h in (10.8) on the fly to avoid storing all factors in (10.8). For a fixed i, we generate U_i^h from K_{ij}^h for all j, and recursively factorize U_i^h. The memory cost is $O(N)$ for storing U_i^h and $O(N^{1/2} \log N)$ for storing the sparse matrices after its recursive factorization. Repeating this process for $i = 1, \ldots, N^{1/2}$ gives the complete factorization of U^h. The factorization of V^h is conducted similarly. The total memory complexity is $O(N \log N)$.

The operation and memory complexity for the application of the butterfly factorization are governed by the number of nonzero entries in the factorization: $O(N \log N)$.

10.4 Numerical Results

This section presents three numerical examples to demonstrate the effectiveness of the algorithms proposed above. The first example is an FIO in [22] and the second example is a special function transform in [138]. Both examples provide an explicit kernel function that becomes a one-dimensional complementary low-rank matrix after discretization. This allows us to apply the butterfly factorization construction algorithm with random sampling. The computational complexity and the memory cost are $O(N^{3/2})$ and $O(N \log N)$ in this case.

The third example is a composition of two FIOs for which an explicit kernel function of their composition is not available. Since we can apply either the butterfly algorithm in [22] or the butterfly factorization to evaluate these FIOs one by one, a fast algorithm for computing the composition is available. We apply the butterfly factorization construction algorithm with random matrix-vector multiplication to this example which requires $O(N^{3/2} \log N)$ operations and $O(N^{3/2})$ memory cost.

Our implementation is in MATLAB. The numerical results were obtained on a server computer with a 2.0 GHz CPU. The additive oversampling parameter is $p = 5$ and the multiplicative oversampling parameter is $q = 3$.

Let $\{u^d(x), x \in X\}$ and $\{u^a(x), x \in X\}$ denote the results given by the direct matrix-vector multiplication and the butterfly factorization. The accuracy of applying the butterfly factorization algorithm is estimated by the following relative error

$$\epsilon^a = \sqrt{\frac{\sum_{x \in S} |u^a(x) - u^d(x)|^2}{\sum_{x \in S} |u^d(x)|^2}}, \tag{10.21}$$

where S is a point set of size 256 randomly sampled from X.

Example 1. Our first example is to evaluate a one-dimensional FIO of the following form:

$$u(x) = \int_{\mathbb{R}} e^{2\pi\imath\Phi(x,\xi)} \widehat{f}(\xi) d\xi, \tag{10.22}$$

where $\widehat{f}$ is the Fourier transform of f, and $\Phi(x, \xi)$ is a phase function given by

$$\Phi(x, \xi) = x \cdot \xi + c(x)|\xi|, \quad c(x) = (2 + \sin(2\pi x))/8. \tag{10.23}$$

The discretization of (10.22) is

$$u(x_i) = \sum_{\xi_j} e^{2\pi\imath\Phi(x_i,\xi_j)} \widehat{f}(\xi_j), \quad i, j = 1, 2, \ldots, N, \tag{10.24}$$

where $\{x_i\}$ and $\{\xi_j\}$ are uniformly distributed points in $[0,1)$ and $[-N/2, N/2)$ following

$$x_i = (i-1)/N \text{ and } \xi_j = j - 1 - N/2. \tag{10.25}$$

(10.24) can be represented in a matrix form as $u = Kg$, where $u_i = u(x_i)$, $K_{ij} = e^{2\pi\imath\Phi(x_i,\xi_j)}$ and $g_j = \widehat{f}(\xi_j)$. The matrix K satisfies the complementary low-rank property as proved in [22, 119]. The explicit kernel function of K allows us to use the construction algorithm with random sampling. Table 10.2 summarizes the results of this example for different grid sizes N and truncation ranks r.

N, r	ϵ^a	$T_{Factor}(min)$	$T_d(sec)$	$T_a(sec)$	T_d/T_a
1024,4	2.49e-05	2.92e-01	2.30e-01	3.01e-02	7.65e+00
4096,4	4.69e-05	1.62e+00	2.64e+00	4.16e-02	6.35e+01
16384,4	5.77e-05	1.22e+01	2.28e+01	1.84e-01	1.24e+02
65536,4	6.46e-05	8.10e+01	2.16e+02	1.02e+00	2.12e+02
262144,4	7.13e-05	4.24e+02	3.34e+03	4.75e+00	7.04e+02
1024,6	1.57e-08	1.81e-01	1.84e-01	1.20e-02	1.54e+01
4096,6	3.64e-08	1.55e+00	2.56e+00	6.42e-02	3.98e+01
16384,6	6.40e-08	1.25e+01	2.43e+01	3.01e-01	8.08e+01
65536,6	6.53e-08	9.04e+01	2.04e+02	1.77e+00	1.15e+02
262144,6	6.85e-08	5.45e+02	3.68e+03	8.62e+00	4.27e+02
1024,8	5.48e-12	1.83e-01	1.78e-01	1.63e-02	1.09e+01
4096,8	1.05e-11	1.98e+00	2.71e+00	8.72e-02	3.11e+01
16384,8	2.09e-11	1.41e+01	3.34e+01	5.28e-01	6.33e+01
65536,8	2.62e-11	1.17e+02	2.10e+02	2.71e+00	7.75e+01
262144,8	4.13e-11	6.50e+02	3.67e+03	1.52e+01	2.42e+02

Table 10.2: Numerical results for the FIO given in (10.24). N is the size of the matrix; r is the fixed rank in the low-rank approximations; T_{Factor} is the factorization time of the butterfly factorization; T_d is the running time of the direct evaluation; T_a is the application time of the butterfly factorization; T_d/T_a is the speedup factor.

Example 2. Next, we provide an example of a special function transform. This example can be further applied to accelerate the Fourier-Bessel transform that is important in many real applications. Following the standard notation, we denote the Hankel function of the first kind of order m by $H_m^{(1)}$. When m is an integer, $H_m^{(1)}$ has a singularity at the origin and a branch cut along the negative real axis. We are interested in evaluating the sum of Hankel functions over different orders,

$$u(x_i) = \sum_{j=1}^{N} H_{j-1}^{(1)}(x_i) g_j, \quad i = 1, 2, \ldots, N, \tag{10.26}$$

which is analogous to expansion in orthogonal polynomials. The points x_i are defined via the formula,

$$x_i = N + \frac{2\pi}{3}(i-1) \tag{10.27}$$

which are bounded away from zero. It is demonstrated in [138] that (10.26) can be represented via $u = Kg$ where K satisfies the complementary low-rank property, $u_i = u(x_i)$ and $K_{ij} = H^{(1)}_{j-1}(x_i)$. The entries of matrix K can be calculated efficiently and the construction algorithm with random sampling is applied to accelerate the evaluation of the sum (10.26). Table 10.3 summarizes the results of this example for different grid sizes N and truncation ranks r.

N, r	ϵ^a	$T_{Factor}(min)$	$T_d(sec)$	$T_a(sec)$	T_d/T_a
1024,4	2.35e-06	8.78e-01	8.30e-01	1.06e-02	7.86e+01
4096,4	5.66e-06	5.02e+00	5.30e+00	2.83e-02	1.87e+02
16384,4	6.86e-06	3.04e+01	5.51e+01	1.16e-01	4.76e+02
65536,4	7.04e-06	2.01e+02	7.59e+02	6.38e-01	1.19e+03
1024,6	2.02e-08	4.31e-01	7.99e-01	9.69e-03	8.25e+01
4096,6	4.47e-08	6.61e+00	5.41e+00	4.52e-02	1.20e+02
16384,6	5.95e-08	4.19e+01	5.62e+01	1.61e-01	3.48e+02
65536,6	7.86e-08	2.76e+02	7.60e+02	1.01e+00	7.49e+02

Table 10.3: Numerical results with the matrix given by (10.26).

Example 3. We provide another example of a special function transform that is related to the spherical harmonic transform (SHT). The SHT is an analogue of the Fourier transform for functions defined on the two-dimensional surface of the unit sphere in $\mathbb{R}^3$. Similar to the Fourier series expansion that represents a function on a unit interval with the eigenfunctions of the Laplacian operator, the spherical harmonic expansion represents a function on a unit sphere with the eigenfunctions of the Laplacian operator on the sphere. For a band-limited function f on the surface of the sphere, its spherical harmonic expansion is

$$f(\theta, \psi) = \sum_{k=0}^{2N-1} \sum_{m=-k}^{k} \alpha_k^m \bar{P}_k^{|m|}(\cos\theta) e^{im\psi}, \tag{10.28}$$

where $\theta \in (0, \pi)$ and $\psi \in (0, 2\pi)$. A standard discretization is to let

$$-1 < \cos\theta_0 < \cos\theta_1 < \cdots < \cos\theta_{2N-1} < 1$$

be the Gauss-Legendre quadrature nodes of degree $2N$, i.e.,

$$\bar{P}^0_{2N}(\cos\theta_k) = 0,$$

and let

$$\psi_j = \frac{2\pi(j+\frac{1}{2})}{4N-1}$$

be equispaced on $(0, 2\pi)$. We can rewrite (10.28) into

$$f(\theta,\psi) = \sum_{m=-2N+1}^{2N-1} e^{im\psi} \sum_{k=|m|}^{2N-1} \alpha_k^m \bar{P}_k^{|m|}(\cos\theta) = \sum_{m=-2N+1}^{2N-1} e^{im\psi} g(m,\theta), \tag{10.29}$$

where

$$g(m,\theta) := \sum_{k=|m|}^{2N-1} \alpha_k^m \bar{P}_k^{|m|}(\cos\theta). \tag{10.30}$$

Since we can apply the fast Fourier transform and its inverse to compute the transformation between $g(m,\theta)$ and $f(\theta,\psi)$ in (10.29), the main computational issue is the transformation between $\{\alpha_k^m\}$ and $g(m,\theta)$ in (10.30).

In [158], it has been shown numerically that $\bar{P}_k^{|m|}(\cos\theta)$ is complementary low-rank in variables k and θ for a fixed m. The author in [158] proposed a butterfly algorithm that applies the interpolative decomposition [33, 80] to evaluate the transformation in (10.30) in $O(N\log N)$ operations for a fixed m. However, the precomputation of this algorithm is $O(N^2)$.

We apply the butterfly factorization in this chapter to design another $O(N\log N)$ algorithm for the transformation in (10.30). The precomputation of the our method is $O(N^{1.5})$ once the matrices $\bar{P}_k^{|m|}(\cos\theta)$ is given. For the transform from $\{\alpha_k^m\}$ to $g(m,\theta)$, we compute the butterfly factorization

$$\bar{P}_k^{|m|}(\cos\theta) \approx U^{m,L} G^{m,L-1} \cdots G^{m,h} M^{m,h} (H^{m,h})^* \cdots (H^{m,L-1})^* (V^{m,L})^*. \tag{10.31}$$

By the fact that $\bar{P}_k^{|m|}(x)$ are orthogonal polynomials, the columns in the matrix $\bar{P}_k^{|m|}(\cos\theta)$ (a column can be considered as a function of θ) are orthogonal. Hence, the inverse of $\bar{P}_k^{|m|}(\cos\theta)$ is just a product of its transpose and a diagonal weight matrix. This leads to an $O(N\log N)$ algorithm from $g(m,\theta)$ to $\{\alpha_k^m\}$.

Table 10.4 summarizes the results of the butterfly factorization of $\bar{P}_k^0(\cos\theta)$ for computing $g(0,\theta)$ from a given Gaussian random vector $\{\alpha_k^0\}$ for different degrees N and truncation ranks r.

N, r	ϵ^a	$T_{Factor}(min)$	$T_d(sec)$	$T_a(sec)$	T_d/T_a
256,4	2.24e-12	2.72e-02	5.13e-03	9.65e-04	5.34e+00
512,4	1.44e-10	7.64e-02	1.63e-02	2.45e-03	6.65e+00
1024,4	3.09e-10	1.70e-01	6.46e-02	6.30e-03	1.03e+01
256,6	1.52e-13	1.84e-02	4.76e-03	8.34e-04	5.71e+00
512,6	3.46e-13	4.16e-02	1.97e-02	2.62e-03	7.52e+00
1024,6	8.16e-12	1.50e-01	6.42e-02	6.07e-03	1.06e+01

Table 10.4: Numerical results with the matrix given by $\bar{P}_k^0(\cos\theta)$.

From Table 10.2, 10.3 and 10.4, we note that the accuracy of the butterfly factorization is well controlled by the max rank r. For a fixed rank r, the accuracy is almost independent of N. In practical applications, one can set the desired ϵ ahead and increase the truncation rank r until the relative error reaches ϵ.

The tables for Example 1, 2 and 3 also provide numerical evidence for the asymptotic complexity of the proposed algorithms. The construction algorithm based on random sampling is of computational complexity $O(N^{3/2})$. When we quadruple the problem size, the running time of the construction sextuples and is better than we expect. The reason is that in the random sampling method, the computation of a middle matrix requires pseudo-inverses of $r \times r$ matrices whose complexity is $O(r^3)$ with a large prefactor. Hence, when N is not large, the running time will be dominated by the $O(r^3 N)$ computation of middle matrices. The numbers also show that the application complexity of the butterfly factorization is $O(N \log N)$ with a prefactor much smaller than the butterfly algorithm with Chebyshev interpolation [22]. In example 1, when the relative error is $\epsilon \approx 10^{-5}$, the butterfly factorization truncates the low-rank submatrices with rank 4 whereas the butterfly algorithm with Chebyshev interpolation uses 9 Chebyshev grid points. The speedup factors are 200 on average.

Example 4. In this example, we consider a composition of two FIOs, which is the discretization of the following operator

$$u(x) = \int_{\mathbb{R}} e^{2\pi\imath\Phi_2(x,\eta)} \int_{\mathbb{R}} e^{-2\pi\imath y\eta} \int_{\mathbb{R}} e^{2\pi\imath\Phi_1(y,\xi)} \widehat{f}(\xi) d\xi dy d\eta. \tag{10.32}$$

For simplicity, we consider the same phase function $\Phi_1 = \Phi_2 = \Phi$ as given by (10.23). By the discussion of Example 1 for one FIO, we know the discrete analog of the composition (10.32) can be represented as

$$u = KFKFf =: KFKg, \quad \text{with } g = Ff,$$

where F is the standard Fourier transform in matrix form, K is the same matrix as in Example 1, $u_i = u(x_i)$, and $g_j = \widehat{f}(\xi_j)$. Under mild assumptions as discussed in [88], the composition of two FIOs is an FIO. Hence, the new kernel matrix $\tilde{K} = KFK$ again satisfies the complementary low-rank property, though typically with slightly increased ranks.

Notice that it is not reasonable to compute the matrix $\tilde{K}$ directly. However, we have the fast Fourier transform (FFT) to apply F and the butterfly factorization that we have built for K in Example 1 to apply K. Therefore, the construction algorithm with random matrix-vector multiplication is applied to factorize $\tilde{K}$.

Since the direct evaluation of each u_i takes $O(N^2)$ operations, the exact solution $\{u_i^d\}_{i\in S}$ for a selected set S is unfeasible for large N. We apply the butterfly factorization of K and the FFT to evaluate $\{u_i\}_{i\in S}$ as an approximation to the exact solution $\{u_i^d\}_{i\in S}$. These approximations

are compared to the results $\{u_i^a\}_{i\in S}$ that are given by applying the butterfly factorization of $\tilde{K}$. Table 10.5 summarizes the results of this example for different grid sizes N and truncation ranks r.

N, r	ϵ^a	$T_{Factor}(min)$	$T_d(sec)$	$T_a(sec)$	T_d/T_a
1024,4	1.40e-02	3.26e-01	3.64e-01	4.74e-03	7.69e+01
4096,4	1.96e-02	4.20e+00	6.59e+00	2.52e-02	2.62e+02
16384,4	2.34e-02	4.65e+01	3.75e+01	1.15e-01	3.25e+02
65536,4	2.18e-02	4.33e+02	3.73e+02	6.79e-01	5.49e+02
1024,8	6.62e-05	3.65e-01	3.64e-01	8.25e-03	4.42e+01
4096,8	8.67e-05	4.94e+00	6.59e+00	5.99e-02	1.10e+02
16384,8	1.43e-04	6.23e+01	3.75e+01	3.47e-01	1.08e+02
65536,8	1.51e-04	6.91e+02	3.73e+02	1.76e+00	2.12e+02
1024,12	1.64e-08	4.79e-01	3.64e-01	1.48e-02	2.46e+01
4096,12	1.05e-07	6.35e+00	6.59e+00	1.12e-01	5.88e+01
16384,12	2.55e-07	7.58e+01	3.75e+01	7.64e-01	4.91e+01
65536,12	2.69e-07	7.63e+02	3.73e+02	4.39e+00	8.49e+01

Table 10.5: Numerical results for the composition of two FIOs.

Table 10.5 shows the numerical results of the butterfly factorization of $\tilde{K}$. The accuracy improves as we increase the truncation rank r. Comparing Table 10.5 with Table 10.2, we notice that, for a fixed accuracy, the rank used in the butterfly factorization of the composition of FIOs should be larger than the rank used in a single FIO butterfly factorization. This is expected since the composition is in general more complicated than the individual FIOs. T_{Factor} grows on average by a factor of ten when we quadruple the problem size. This agrees with the estimated $O(N^{3/2}\log N)$ computational complexity for constructing the butterfly factorization. The column T_a shows that the empirical application time of our factorization is close to the estimated complexity $O(N\log N)$.

10.5 Conclusion

This chapter introduces the one-dimensional butterfly factorization as a data-sparse approximation of one-dimensional complementary low-rank matrices. More precisely, it represents such an $N\times N$ dense matrix as a product of $O(\log N)$ sparse matrices. The factorization can be built efficiently if either a fast algorithm for applying the matrix and its adjoint is available or an explicit expression for the entries of the matrix is given. The butterfly factorization gives rise to highly efficient matrix-vector multiplications with $O(N\log N)$ operation and memory complexity. The butterfly factorization is also useful when an existing butterfly algorithm is repeatedly applied, because the application of the butterfly factorization is significantly faster than pre-existing butterfly algorithms.

Chapter 11

Multi-Dimensional Butterfly Factorization

11.1 Introduction

In Chapter 10, we have introduced a one-dimensional butterfly factorization for efficient kernel evaluation of the form,

$$u(x) = \sum_{\xi \in \Omega} K(x, \xi) g(\xi), \quad x \in X, \tag{11.1}$$

where $K(x, \xi)$ is a kernel function that satisfies a complementary low-rank property, and X, Ω are point sets in $\mathbb{R}^1$. This chapter introduces a multi-dimensional butterfly factorization for point sets X and $\Omega \subset \mathbb{R}^d$ with $d \geq 2$ to accelerate the evaluation of (11.1). This is joint work with Yingzhou Li and Lexing Ying in [118]. The concept of the complementary low-rank property in one-dimensional space can be extended to multi-dimensional spaces. With no loss of generality, we assume the points in X and Ω are uniformly distributed with N points in each dimension. Let T_X and T_Ω be two d-dimensional dyadic trees associated with domains X and Ω, respectively. They have the same depth $L = O(\log N)$ with X and Ω as roots on the zero level. A kernel $K(x, \xi)$ (or its matrix representation) satisfies the complementary low-rank property if for any level $\ell = 0, 1, \ldots, L$, any node $A \in T_X$ on the ℓ-th level and any node $B \in T_\Omega$ on the $(L - \ell)$-th level, the submatrix $K_{A,B} = \{K(x_i, \xi_j)\}_{x_i \in A, \xi_j \in B}$ is numerically low-rank with rank bounded by a uniform constant independent of N. A well-known example is the (nonuniform) Fourier transform. As a discrete analogue, we say a matrix K is complementary low-rank if it is the matrix representation of a complementary low-rank kernel $K(x, \xi)$. The multi-dimensional butterfly factorization factorizes K into

$$K \approx U^L G^{L-1} \cdots G^{L/2} M^{L/2} \left(H^{L/2}\right)^* \cdots \left(H^{L-1}\right)^* \left(V^L\right)^*, \tag{11.2}$$

where the depth L is assumed to be an even number and all factors are sparse matrices with $O(N^d)$ non-zero entries. Here the superscript of a matrix denotes the level of the factor other than the power of a matrix. After factorization, it only takes $O(N^d \log N)$ memory and operation complexity to store and apply K.

In general, a multi-dimensional kernel $K(x, \xi)$ coming from real applications may not satisfy the complementary low-rank property in the whole domain $X \times \Omega$, but the property is true locally in $X \times \Omega$. For example, a multi-dimensional kernel $K(x, \xi)$ coming from a Fourier integral operator (FIO) [21, 22, 119] would have irregularity at $\xi = 0$. The FIO kernel $K(x, \xi)$ is complementary low-rank in a subdomain $X \times \Omega_j \subset X \times \Omega$ with $\xi = 0$ away from Ω_j. There are mainly two methods to deal with this irregularity at $\xi = 0$. One idea is to apply a well-designed polar transformation mapping the domain $X \times \Omega$ into a new domain $X \times P$ such that $\xi = 0$ is mapped to the boundary of P. After this transformation, the new kernel function defined on $X \times P$ would be complementary low-rank. Another idea is to partition the whole domain $X \times \Omega$ into a sequence of subdomains $X \times \Omega_j$ such that $K(x, \xi)$ is complementary low-rank in each subdomain. The corresponding kernel matrix K can be represented as a sequence of a few butterfly factorizations

$$K_{X,\Omega_j} \approx U^{j,L_j} G^{j,L_j-1} \cdots G^{j,\frac{L_j}{2}} M^{j,\frac{L_j}{2}} \left(H^{j,\frac{L_j}{2}}\right)^* \cdots \left(H^{j,L_j-1}\right)^* \left(V^{j,L_j}\right)^*, \tag{11.3}$$

and the total number of nonzero entries in the above factorizations is $O(N^d \log N)$. Hence, this factorization admits $O(N^d \log N)$ memory and operation complexity to store and apply the matrix K as well. Since the domain partition is application-dependent, we would focus on the FIO kernel as an illustration of the multi-dimensional butterfly factorization with irregularity.

Similar to the one-dimensional butterfly factorization, the multi-dimensional butterfly factorization can be constructed in two ways:

(i) A black-box routine for computing Kg and K^*g in $O(N^d \log N)$ operations is given;

(ii) A black-box routine for evaluating any entry of K in $O(1)$ operations is given.

We will first introduce the multi-dimensional butterfly factorization when the kernel $K(x, \xi)$ is complementary low-rank in the whole domain $X \times \Omega$. In the case of point irregularity at $\xi = 0$ for FIO kernels, we presents two kinds of factorizations: one is based on transforming $X \times \Omega$ to a new domain $X \times P$ via a polar-Cartesian transformation; another one is based on a multiscale partition of $X \times \Omega$ in a Cartesian grid. We denote these algorithms as *polar butterfly factorization (PBF)*, and *multiscale butterfly factorization (MBF)*, respectively. The idea of PBF and MBF comes from the polar butterfly algorithm in [22] and the multiscale butterfly algorithm in [119] or Chapter 9. PBF and MBF are essentially matrix representations of these fast algorithms.

For simplicity, we will introduce the multi-dimensional butterfly factorizations for $d = 2$. The

butterfly factorization can be constructed in a similar way for $d > 2$. We also assume that

$$X = \left\{ x = \left(\frac{n_1}{N}, \frac{n_2}{N} \right), 0 \leq n_1, n_2 < N \text{ with } n_1, n_2 \in \mathbb{Z} \right\} \tag{11.4}$$

in a unit square and defines

$$\Omega = \left\{ \xi = (n_1, n_2), -\frac{N}{2} \leq n_1, n_2 < \frac{N}{2} \text{ with } n_1, n_2 \in \mathbb{Z} \right\}. \tag{11.5}$$

11.2 Preliminaries

Let us briefly review the ideas of fast low-rank approximations that we have used in Chapter 10, the polar butterfly algorithm in [22] and multiscale butterfly algorithm in [119] and Chapter 9.

11.2.1 Randomized Low-Rank Factorization

For a matrix $Z \in \mathbb{C}^{m \times n}$, its rank-$r$ approximation in 2-norm can be computed via the truncated singular value decomposition (SVD),

$$Z \approx U_0 \Sigma_0 V_0^*, \tag{11.6}$$

where $U_0 \in \mathbb{C}^{m \times r}$ and $V_0 \in \mathbb{C}^{n \times r}$ are unitary matrices, $\Sigma_0 \in \mathbb{R}^{r \times r}$ is a diagonal matrix with the first r singular values of Z in decreasing order.

Once the rank-r SVD, $Z \approx U_0 \Sigma_0 V_0^*$, is available, we have three different ways to assign singular values to obtain a rank-r approximation of Z:

$$Z \approx USV^*, \quad U = U_0 \Sigma_0, \quad S = \Sigma_0^{-1}, \text{ and } V^* = \Sigma_0 V_0^*, \tag{11.7}$$

$$\text{or } Z \approx UV^*, \quad U = U_0 \Sigma_0, \text{ and } V^* = V_0^*, \tag{11.8}$$

$$\text{or } Z \approx UV^*, \quad U = U_0, \text{ and } V^* = \Sigma_0 V_0^*, \tag{11.9}$$

each of which results in particular benefit in the butterfly factorization.

As discussed in Section 10.2, the rank-r SVD can be constructed via either the randomized SVD in [84] or the randomized sampling algorithm in [67, 181]. If we denote the application complexity of Z and its adjoint as $O(C(m, n))$, then the construction complexity of the rank-r SVD via the randomized SVD is $O(C(m, n)r + \max(m, n)r^2)$, while the construction complexity via the randomized sampling algorithm is $O(\max(m, n)r^2)$.

11.2.2 Polar Butterfly Algorithm

The polar butterfly algorithm is initially designed for multi-dimensional Fourier integral operators

$$u(x) = \sum_{\xi \in \Omega} e^{2\pi\imath\Phi(x,\xi)} g(\xi), \quad x \in X, \tag{11.10}$$

where the phase function $\Phi(x,\xi)$ is assumed to be smooth in (x,ξ) for $\xi \neq 0$, and obeys an homogeneity condition of degree 1 in ξ, namely, $\Phi(x,\lambda\xi) = \lambda\Phi(x,\xi)$ for all $\lambda > 0$.

Due to the fact that the phase function $\Phi(x,\xi)$ might be singular at $\xi = 0$, the derivative of $\Phi(x,\xi)$ near $\xi = 0$ may not be bounded. The numerical rank of the kernel $e^{2\pi\imath\Phi(x,\xi)}$ in a domain containing $\xi = 0$ can be very large. Hence, $K(x,\xi) = e^{2\pi\imath\Phi(x,\xi)}$ does not satisfy the complementary low-rank property over the entire domain $X \times \Omega$. In [22], the authors introduce a polar coordinate transform on Ω:

$$\xi = (\xi_1, \xi_2) = \frac{\sqrt{2}}{2} N p_1 e^{2\pi\imath p_2}, \quad e^{2\pi\imath p_2} = (\cos 2\pi p_2, \sin 2\pi p_2), \tag{11.11}$$

for $\xi \in \Omega$ and $p = (p_1, p_2) \in [0,1]^2$. In the rest of this chapter, points in a polar coordinate are denoted by p, and the set of all points p transformed from Ω is denoted by P. Accordingly, we can introduce a new phase function $\Psi(x,p)$ in variables x and p satisfying

$$\Psi(x,p) = \frac{1}{N}\Phi(x,\xi) = \frac{\sqrt{2}}{2}\Phi(x, e^{2\pi\imath p_2})p_1, \tag{11.12}$$

where the last equality comes from the fact that $\Phi(x,\xi)$ is homogeneous of degree 1 in ξ. After the polar transform, the new phase function $\Psi(x,p)$ is smooth in the whole domain $X \times P$. Hence, $e^{2\pi\imath N\Psi(x,p)}$ satisfies the complementary low-rank property almost over the whole domain $X \times P$ and the matrix representation of the new kernel $e^{2\pi\imath N\Psi(x,p)}$ in $X \times P$ is complementary low-rank.

Recall that $X \times P \subset [0,1]^2 \times [0,1]^2$. By dyadic partition of $[0,1]^2$, we can construct two quadtrees T_X and T_P of depth $L = O(\log N)$ associated with X and P, respectively. The following theorem is rephrased from Theorem 3.1 in [22]. It supports the complementary low-rank property of $e^{2\pi\imath N\Psi(x,p)}$ as observed above. We denote $f \lesssim g$ if $f \leq Cg$ for some constant C independent of N and a given small ϵ.

Theorem 11.2.1. *Suppose A is a node in T_X on level ℓ and B is a node in T_P on level $L - \ell$. Let c_A and c_B be the center of A and B, respectively. Given an FIO kernel function $e^{2\pi\imath N\Psi(x,p)}$, there exist $\epsilon_0 > 0$ and $N_0 > 0$ such that for any $\epsilon \leq \epsilon_0$ and $N \geq N_0$, there exist r_ϵ pairs of functions $\{\alpha_t^{A,B}(x), \beta_t^{A,B}(p)\}_{1 \leq t \leq r_\epsilon}$ satisfying that*

$$\left| e^{2\pi\imath N\Psi(x,p)} - e^{-2\pi\imath N\Psi(c_A,c_B)} \sum_{t=1}^{r_\epsilon} e^{2\pi\imath N\Psi(x,c_B)} \alpha_t^{A,B}(x) \beta_t^{A,B}(p) e^{2\pi\imath N\Psi(c_A,p)} \right| \leq \epsilon,$$

for $x \in A$ and $p \in B$ with $r_\epsilon \lesssim \log^4(1/\epsilon)$.

The polar butterfly algorithm in [22] applies the butterfly algorithm detailed in Section 9.2 to evaluate an FIO via a summation in the form

$$u(x) = \sum_{p \in P} e^{2\pi \imath N \Psi(x,p)} g(p), \quad x \in X. \tag{11.13}$$

It simultaneously traverses in T_X level by level from the root X to leaves, and traverses in T_P from leaves to the root P. At each step of the traverse, suppose we are on level ℓ in T_X and on level $L - \ell$ in T_P, the algorithm constructs the low-rank approximation of $e^{2\pi \imath N \Psi(x,p)}$ for each pair of $A \times B \in T_X \times T_P$ on the current level. The approximation functions $\{\alpha_t^{A,B}(x), \beta_t^{A,B}(p)\}_{1 \le t \le r_\epsilon}$ are constructed via efficient Lagrange interpolation on the Chebyshev grid similar to the method described in Section 9.4.1.

The polar butterfly algorithm is highly efficient. It evaluates (11.13) with $O(N^2 \log N)$ operation complexity and $O(N^2)$ memory complexity. However, it is not suitable for the problem addressed in this chapter when the kernel function is not available explicitly.

11.2.3 Multiscale Butterfly Algorithm

The multiscale butterfly algorithm in [119] has been introduced in Chapter 9. We briefly recall its theory and ideas here. The key idea of the multiscale butterfly algorithm is to hierarchically partition the domain Ω into subdomains excluding the singular point $\xi = 0$. This multiscale partition is illustrated in Figure 11.1. The FIO kernel $e^{2\pi \imath \Phi(x,\xi)}$ satisfies the complementary low-rank property when it is restricted in each subdomain $X \times \Omega_j$. This is supported by the following theorem rephrased from Theorem 9.3.1 in Chapter 9. Recall that $\operatorname{dist}(B, 0) = \min_{\xi \in B} \|\xi - 0\|$ is the distance between the square B and the origin $\xi = 0$ in Ω.

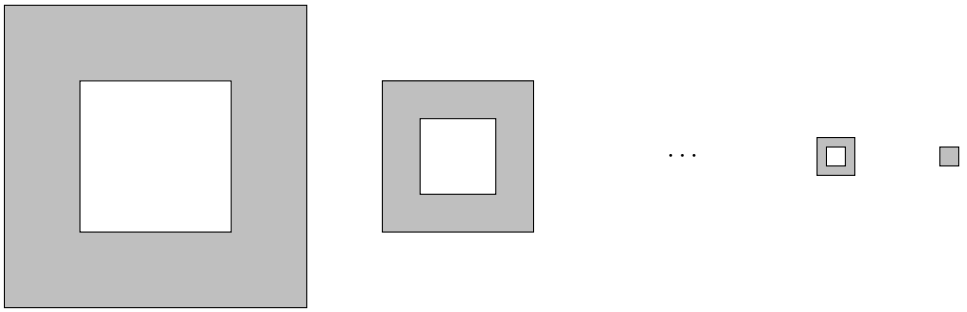


Figure 11.1: This figure shows the domain partition of Ω. Each subdomain Ω_j, $j = 1, \ldots, \log N - s$, is a corona and Ω_d is a small square domain near the origin.

Theorem 11.2.2. *Given an FIO kernel function $e^{2\pi\imath\Phi(x,p)}$, there exist a constant $N_0 > 0$ and a small constant ϵ_0 such that the following statement holds. Let A and B be two squares in X and Ω centered at c_A and c_B with length w_A and w_B, respectively. Suppose $w_A w_B \leq 1$ and dist$(B, 0) \geq \frac{N}{4}$. For any positive $\epsilon \leq \epsilon_0$ and $N \geq N_0$, there exist r_ϵ pairs of functions $\{\alpha_t^{A,B}(x), \beta_t^{A,B}(p)\}_{1\leq t\leq r_\epsilon}$ satisfying that*

$$\left| e^{2\pi\imath\Phi(x,\xi)} - e^{-2\pi\imath\Phi(c_A,c_B)} \sum_{t=1}^{r_\epsilon} e^{2\pi\imath\Phi(x,c_B)} \alpha_t^{A,B}(x) \beta_t^{A,B}(\xi) e^{2\pi\imath\Phi(c_A,\xi)} \right| \leq \epsilon,$$

for $x \in A$ and $\xi \in B$ with $r_\epsilon \lesssim \log^4(1/\epsilon)$.

According to the low-rank property in Theorem 11.2.2, the multiscale butterfly algorithm rewrites (11.10) as a multiscale summation,

$$u(x) = \sum_{\xi\in\Omega_d} e^{2\pi\imath\Phi(x,\xi)} g(\xi) + \sum_j \sum_{\xi\in\Omega_j} e^{2\pi\imath\Phi(x,\xi)} g(\xi), \tag{11.14}$$

where

$$\Omega_j = \{(\xi_1, \xi_2) : \frac{N}{2^{j+2}} < \max(|\xi_1|, |\xi_2|) \leq \frac{N}{2^{j+1}}\} \cap \Omega, \tag{11.15}$$

for $j = 0, 1, \ldots, \log_2 N - s$, s is a small constant, and $\Omega_d = \Omega \setminus \cup_j \Omega_j$. Equation (11.15) is a corona decomposition of Ω, where each Ω_j is a corona and Ω_d is a disk at the center containing $O(1)$ points.

For each j, the multiscale butterfly algorithm evaluates $u_j(x) = \sum_{\xi\in\Omega_j} e^{2\pi\imath\Phi(x,\xi)} g(\xi)$ via the Cartesian butterfly algorithm with Lagrange interpolation on Chebyshev grid points detailed in Section 9.4. The summation $u_d(x) = \sum_{\xi\in\Omega_d} e^{2\pi\imath\Phi(x,\xi)} g(\xi)$ is directly computed in $O(N)$ operations. Finally, $u(x)$ is a simple summation

$$u(x) = u_d(x) + \sum_j u_j(x), \quad x \in X. \tag{11.16}$$

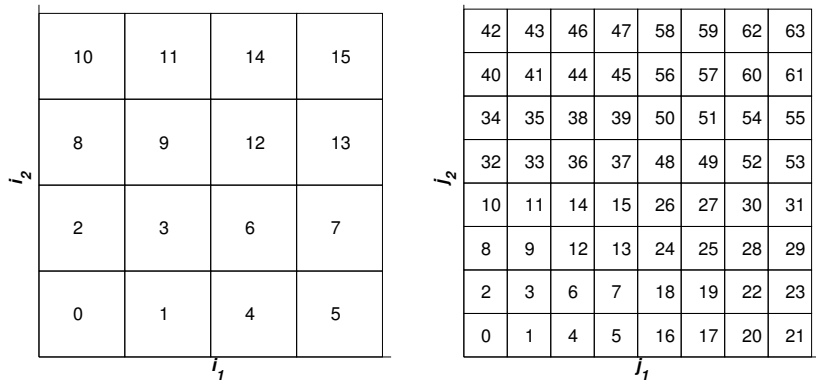
The multiscale butterfly algorithm enjoys $O(N^2 \log N)$ operation complexity and $O(N^2)$ memory complexity as analyzed in Section 9.4. However, it is still not suitable for the problem addressed in this chapter when the kernel function is not available explicitly.

11.3 Multi-Dimensional Butterfly Factorization

This section presents the multidimensional butterfly factorization for a kernel $K(x, \xi)$ that satisfies the complementary low-rank property in $X \times \Omega$. Recall that T_X and T_Ω are complete quadtrees with $L = O(\log N)$ levels. Without loss of generality, we assume L is an even integer.

We basically adopt the notation introduced in Chapter 10 for one-dimensional butterfly factorizations, but adapt them to high dimensional problems in this chapter by using vector indices in a

bold front. On the ℓ-th level of a quadtree, $\ell = 0, 1, \ldots, L$, we denote the $\boldsymbol{i}$th node of T_X as $A^\ell_{\boldsymbol{i}}$ for $\boldsymbol{i} = (i_1, i_2)$, and $i_1, i_2 = 0, 1, \ldots, 2^\ell - 1$. At each level ℓ, these 4^ℓ different vector indices can be ordered in a special zigzag way as $\boldsymbol{i}_0, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_{4^\ell-1}$ as illustrated in Figure 11.2. If we introduce four vectors $e_0 = (0, 0)$, $e_1 = (0, 1)$, $e_2 = (1, 0)$ and $e_3 = (1, 1)$, then a domain $A^\ell_{\boldsymbol{i}}$ at level ℓ has four child domains $A^{\ell+1}_{2\boldsymbol{i}+e_t}$ for $t = 0, \ldots, 3$. The ordered numbers plotted in Figure 11.2 for $\ell = 2$ (left) and 3 (right) illustrate the relation of indices and their orders at different levels. In what follows, we sometimes use vector indices rather than its ordered scalar equivalent for the purpose of convenience. Similarly, on level $L - \ell$ of T_Ω, the $\boldsymbol{j}$th node is denoted as $B^{L-\ell}_{\boldsymbol{j}}$ for $\boldsymbol{j} = (j_1, j_2)$ and $j_1, j_2 = 0, 1, \ldots, 2^{L-\ell} - 1$. The kernel matrix K is naturally partitioned into $O(N^2)$ submatrices $\{K_{A^\ell_{\boldsymbol{i}}, B^{L-\ell}_{\boldsymbol{j}}}\}_{\boldsymbol{i},\boldsymbol{j}}$. For simplicity, we write $K^\ell_{\boldsymbol{i},\boldsymbol{j}} = K_{A^\ell_{\boldsymbol{i}}, B^{L-\ell}_{\boldsymbol{j}}}$ where the superscript ℓ denotes the level in T_X. Based on the complementary low-rank property, every submatrix $K^\ell_{\boldsymbol{i},\boldsymbol{j}}$ is numerically low-rank with a rank bounded by r independent of N.



Subdomains $A^2_{\boldsymbol{i}}$ in T_X Subdomains $B^3_{\boldsymbol{j}}$ in T_Ω

Figure 11.2: Two examples of ordering vector indices $\boldsymbol{i}$ and $\boldsymbol{j}$. Left: At level 2, the domain X is divided into 4×4 subdomains $A^2_{\boldsymbol{i}}$ with $\boldsymbol{i} = (i_1, i_2)$ for $i_1, i_2 = 0, \ldots, 3$. These 16 vector indices can be ordered in a special zigzag way as $\boldsymbol{i}_0, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_{15}$. For each $\boldsymbol{i} = (\boldsymbol{i}_1, \boldsymbol{i}_2)$, its order is plotted at position $(\boldsymbol{i}_1, \boldsymbol{i}_2)$ in the left figure. Right: At level 3, the domain Ω is divided into 8×8 subdomains $B^3_{\boldsymbol{j}}$ with $\boldsymbol{j} = (j_1, j_2)$ for $j_1, j_2 = 0, \ldots, 7$. These 64 different vector indices $\boldsymbol{j}$ can be ordered similarly.

The two-dimensional butterfly factorization consists of two stages. In the first stage, we factorize $K^h_{\boldsymbol{i},\boldsymbol{j}} \approx U^h_{\boldsymbol{i},\boldsymbol{j}} S^h_{\boldsymbol{i},\boldsymbol{j}} \left(V^h_{\boldsymbol{j},\boldsymbol{i}}\right)^*$ for all $\boldsymbol{i}$ and $\boldsymbol{j}$ at level $\ell = h = L/2$ using fast low-rank factorizations and (11.7). These factorizations can be assembled into three sparse matrices, U^h, M^h and V^h such that

$$K \approx U^h M^h (V^h)^*. \tag{11.17}$$

This stage is referred as *middle level factorization* described in Section 11.3.1. In the second stage, we recursively factorize the left and right factors, U^h and V^h since their submatrices $U^h_{\boldsymbol{i},\boldsymbol{j}}$ and $V^h_{\boldsymbol{j},\boldsymbol{i}}$ have a special low-rank property as discussed later. After the recursive factorization, we assemble

all factors into sparse matrices and get an approximate factorization of K,

$$K \approx U^L G^{L-1} \cdots G^h M^h \left(H^h\right)^* \cdots \left(H^{L-1}\right)^* \left(V^L\right)^*, \tag{11.18}$$

where all sparse matrices have $O(N^2)$ non-zero entries. This stage is referred as *recursive factorization* discussed in Section 11.3.2.

11.3.1 Middle Level Factorization

Recall that we consider two cases as follows.

(i) A black-box routine for computing Kg and K^*g in $O(N^2 \log N)$ operations is given.

(ii) A black-box routine for evaluating any entry of K in $O(1)$ operations is given.

In case (i), we apply the randomized SVD method [84] to construct the rank-r SVD of each $K^h_{\boldsymbol{i},\boldsymbol{j}} \in \mathbb{R}^{N\times N}$. This requires applying $K^h_{\boldsymbol{i},\boldsymbol{j}}$ and its adjoint to a Gaussian random matrix $C_s \in \mathbb{C}^{N\times(r+k)}$, where r is the numerical rank of $K^h_{\boldsymbol{i},\boldsymbol{j}}$ and k is an oversampling parameter (in most cases $k = 5$ is sufficient). Note that only a black box routine for applying the whole matrix K and its adjoint is available. We construct a larger random matrix $C_\ell \in \mathbb{C}^{N^2\times(r+k)}$ by padding zero rows in C_s such that KC_ℓ will only touch the desired columns in K corresponding to the submatrix $K^h_{\boldsymbol{i},\boldsymbol{j}}$. Hence, $K^h_{\boldsymbol{i},\boldsymbol{j}}C_s$ can be read off from the rows in KC_ℓ corresponding to $K^h_{\boldsymbol{i},\boldsymbol{j}}$. A similar method gives $(K^h_{\boldsymbol{i},\boldsymbol{j}})^*C_s$. Finally, we compute the rank-r SVD of $K^h_{\boldsymbol{i},\boldsymbol{j}}$ from $K^h_{\boldsymbol{i},\boldsymbol{j}}C_s$ and $\left(K^h_{\boldsymbol{i},\boldsymbol{j}}\right)^* C_s$ following the randomized SVD method in [84] for all $\boldsymbol{i}$ and $\boldsymbol{j}$.

In the more flexible case (ii), we can accelerate the rank-r SVD of $K^h_{\boldsymbol{i},\boldsymbol{j}}$ via the randomized sampling algorithm [67, 181] as what we do in the one-dimensional butterfly factorization.

In either case, we can obtain a low-rank factorization

$$K^h_{\boldsymbol{i},\boldsymbol{j}} \approx U^h_{\boldsymbol{i},\boldsymbol{j}} S^h_{\boldsymbol{i},\boldsymbol{j}} \left(V^h_{\boldsymbol{j},\boldsymbol{i}}\right)^* \tag{11.19}$$

from the rank-r SVD of $K^h_{\boldsymbol{i},\boldsymbol{j}}$ via (11.7). The factorization method via (11.7) scales the left and right factors $U^h_{\boldsymbol{i},\boldsymbol{j}}$ and $V^h_{\boldsymbol{j},\boldsymbol{i}}$ with the singular values of $K^h_{\boldsymbol{i},\boldsymbol{j}}$ so that $U^h_{\boldsymbol{i},\boldsymbol{j}}$ and $V^h_{\boldsymbol{j},\boldsymbol{i}}$ keep track of the importance of column and row spaces for further factorizations.

After computing the rank-r factorization in (11.19) for all $\boldsymbol{i}$ and $\boldsymbol{j}$, we assemble all left factors $U^h_{\boldsymbol{i},\boldsymbol{j}}$ into a matrix U^h, all middle factors into a matrix M^h, and all right factors into a matrix V^h such that

$$K \approx U^h M^h (V^h)^* \tag{11.20}$$

as visualized in Figure 11.3. Let us recall that $\boldsymbol{i}$ and $\boldsymbol{j}$ are vector indices. At the middle level, there are N different vector indices $\boldsymbol{i} = (i_1, i_2)$ with $0 \le i_1 \le 2^h - 1$ and $0 \le i_2 \le 2^h - 1$. Hence, $\boldsymbol{i}$ can be ordered as $\boldsymbol{i}_0, \ldots, \boldsymbol{i}_{N-1}$. Similarly, we have $\boldsymbol{j}_0, \ldots, \boldsymbol{j}_{N-1}$. With these notations ready, we can

explain the factors in (11.20). Here U^h is a block diagonal matrix of size $N^2 \times rN^2$ with N diagonal blocks $U^h_{\boldsymbol{i}}$ of size $N \times rN$. The diagonal block $U^h_{\boldsymbol{i}}$ is a stack of left factors $U^h_{\boldsymbol{i},\boldsymbol{j}}$ for all column indices $\boldsymbol{j}$ as follows:

$$U^h_{\boldsymbol{i}} = \begin{pmatrix} U^h_{\boldsymbol{i},\boldsymbol{j}_0} & U^h_{\boldsymbol{i},\boldsymbol{j}_1} & \cdots & U^h_{\boldsymbol{i},\boldsymbol{j}_{N-1}} \end{pmatrix} \in \mathbb{C}^{N \times rN}. \tag{11.21}$$

Similarly, V^h is a block diagonal matrix of size $N^2 \times rN^2$ with N diagonal blocks $V^h_{\boldsymbol{j}}$ of size $N \times rN$. The diagonal block $V^h_{\boldsymbol{j}}$ is a stack of left factors $V^h_{\boldsymbol{j},\boldsymbol{i}}$ for all column indices $\boldsymbol{i}$ as follows:

$$V^h_{\boldsymbol{j}} = \begin{pmatrix} V^h_{\boldsymbol{j},\boldsymbol{i}_0} & V^h_{\boldsymbol{j},\boldsymbol{i}_1} & \cdots & V^h_{\boldsymbol{j},\boldsymbol{i}_{N-1}} \end{pmatrix} \in \mathbb{C}^{N \times rN}. \tag{11.22}$$

The middle matrix $M^h \in \mathbb{C}^{rN^2 \times rN^2}$ serves as a weighted permutation matrix. Hence, it is a $N \times N$ block matrix with the $(\boldsymbol{i},\boldsymbol{j})$th block $M^h_{\boldsymbol{i},\boldsymbol{j}} \in \mathbb{C}^{rN \times rN}$ being an $N \times N$ block matrix again. Here $(\boldsymbol{i},\boldsymbol{j})$ means a 1×2 vector (k_1, k_2) where k_1 is the order of $\boldsymbol{i}$ and k_2 is the order of $\boldsymbol{j}$ among other vector indices on the same level. This ordering has been illustrated in Figure 11.2. The $(\boldsymbol{j},\boldsymbol{i})$th block of $M^h_{\boldsymbol{i},\boldsymbol{j}}$ is equal to $S^h_{\boldsymbol{i},\boldsymbol{j}}$ and the other blocks of $M^h_{\boldsymbol{i},\boldsymbol{j}}$ are zero. Hence, we have $K \approx U^h M^h (V^h)^*$ (see Figure 11.3 for an example of the middle level factorization when $N = 16$).

11.3.2 Recursive Factorization

In this section we will recursively factorize

$$U^\ell \approx U^{\ell+1} G^\ell \tag{11.23}$$

and

$$(V^\ell)^* \approx (H^\ell)^* V^{\ell+1})^* \tag{11.24}$$

for $\ell = h, h+1, \ldots, L-1$. After these recursive factorizations, we can construct the two-dimensional butterfly factorization

$$K \approx U^L G^{L-1} \cdots G^h M^h \left(H^h\right)^* \cdots \left(H^{L-1}\right)^* \left(V^L\right)^* \tag{11.25}$$

by substituting these recursive factorizations into (11.20).

Recursive factorization of U^h

Recall from the middle level factorization that we took advantage of the low-rank property of $K^h_{\boldsymbol{i},\boldsymbol{j}}$ to obtain $U^h_{\boldsymbol{i},\boldsymbol{j}}$ for $\boldsymbol{i} = (i_1, i_2)$ and $\boldsymbol{j} = (j_1, j_2)$, $i_1, i_2, j_1, j_2 = 0, 1, \ldots, 2^h - 1$. The matrix $K^h_{\boldsymbol{i},\boldsymbol{j}}$ represents the kernel function restricted in the domain $A^h_{\boldsymbol{i}} \times B^h_{\boldsymbol{j}} \in T_X \times T_\Omega$. Next, we are going to use the low-rank property of $K^{h+1}_{\boldsymbol{i},\boldsymbol{j}}$ that represents the kernel function restricted in $A^{h+1}_{\boldsymbol{i}} \times B^{h-1}_{\boldsymbol{j}} \in T_X \times T_\Omega$ for $\boldsymbol{i} = (i_1, i_2)$, $i_1, i_2 = 0, 1, \ldots, 2^{h+1} - 1$ and $\boldsymbol{j} = (j_1, j_2)$, $j_1, j_2 = 0, 1, \ldots, 2^{h-1} - 1$, to construct the factorization (11.23) for $\ell = h$. This process consists of three steps: a splitting step equivalent to

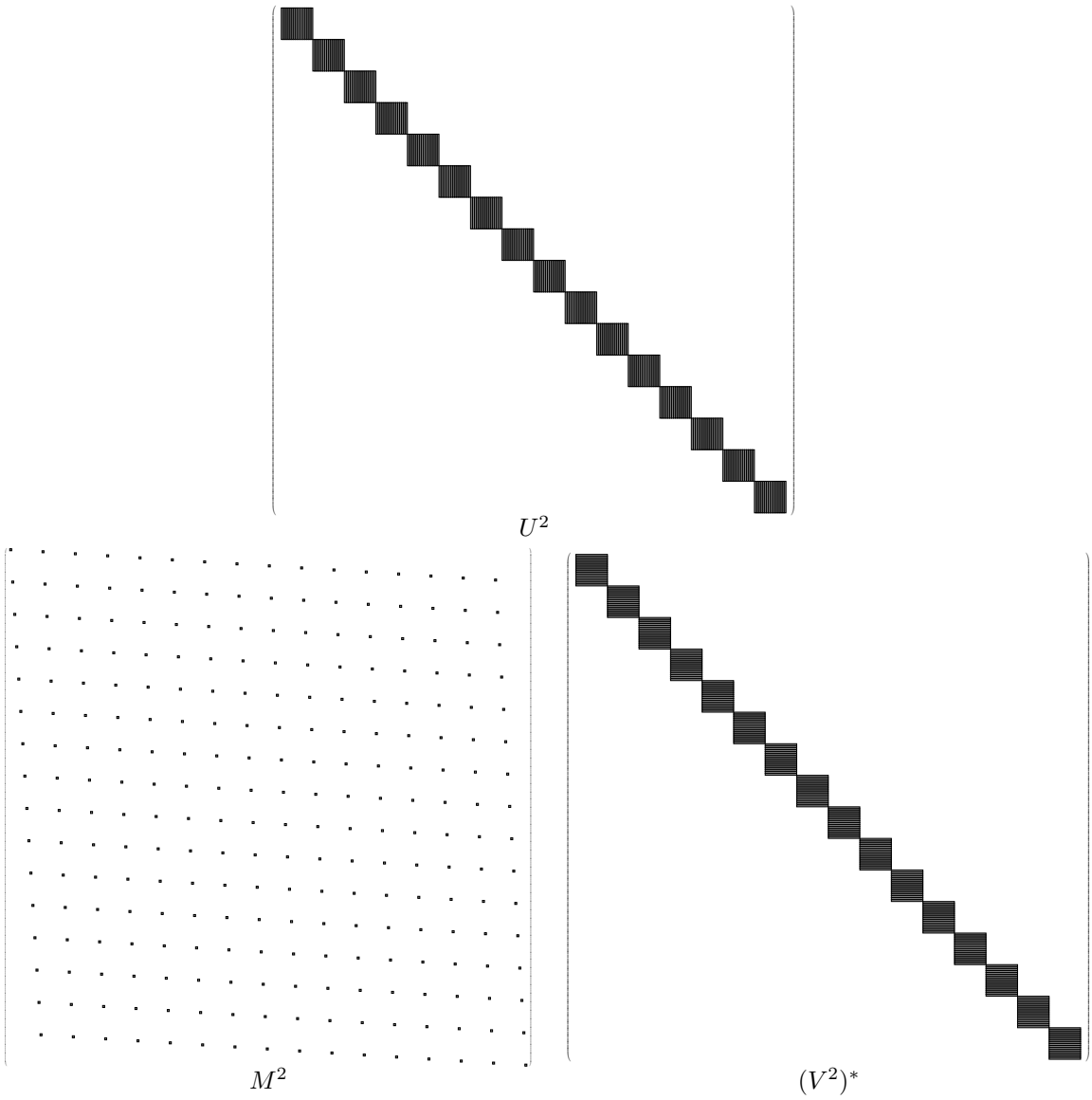


Figure 11.3: The middle level factorization of a $N^2 \times N^2 = 256 \times 256$ complementary low-rank matrix $U^2 M^2 (V^2)^*$ assuming $r = 1$. Grey blocks indicate nonzero blocks. U^2 and V^2 are block-diagonal matrices with 16 blocks. The diagonal blocks of U^2 and V^2 are assembled according to Equation (11.21) and (11.22) as indicated by black rectangles. M^2 is a 16×16 block matrix with each block $M^2_{\boldsymbol{i},\boldsymbol{j}}$ itself an 16×16 block matrix containing diagonal weight matrix on the $(\boldsymbol{j}, \boldsymbol{i})$ block.

moving from level h to level $h+1$ in T_X; a merging step equivalent to moving from level h to level $h-1$ in T_Ω; an assembling step for constructing (11.23) from small factorizations.

Let us start with the splitting step first. In the middle level factorization, we have constructed

$$U^h = \begin{pmatrix} U^h_{\boldsymbol{i}_0} & & & \\ & U^h_{\boldsymbol{i}_1} & & \\ & & \ddots & \\ & & & U^h_{\boldsymbol{i}_{4^h-1}} \end{pmatrix}$$

and

$$U^h_{\boldsymbol{i}} = \begin{pmatrix} U^h_{\boldsymbol{i},\boldsymbol{j}_0} & U^h_{\boldsymbol{i},\boldsymbol{j}_1} & \cdots & U^h_{\boldsymbol{i},\boldsymbol{j}_{4^h-1}} \end{pmatrix} \in \mathbb{C}^{N\times rN}$$

with each $U^h_{\boldsymbol{i},\boldsymbol{j}} \in \mathbb{C}^{N\times r}$. Each node $A^h_{\boldsymbol{i}}$ in the quadtree T_X on the level h has four child nodes on the level $h+1$. Recall that we have introduced four vectors $e_0 = (0,0)$, $e_1 = (0,1)$, $e_2 = (1,0)$ and $e_3 = (1,1)$. Hence, these child nodes can be denoted by $\{A^{h+1}_{2\boldsymbol{i}+e_t}\}_{t=0,1,2,3}$ using consistent vector notations. Correspondingly, if we quarter $U^h_{\boldsymbol{i}}$ and $U^h_{\boldsymbol{i},\boldsymbol{j}}$ into four parts by row, i.e.,

$$U^h_{\boldsymbol{i}} = \begin{pmatrix} U^{h,0}_{\boldsymbol{i}} \\ \hline U^{h,1}_{\boldsymbol{i}} \\ \hline U^{h,2}_{\boldsymbol{i}} \\ \hline U^{h,3}_{\boldsymbol{i}} \end{pmatrix} \text{ and } U^h_{\boldsymbol{i},\boldsymbol{j}} = \begin{pmatrix} U^{h,0}_{\boldsymbol{i},\boldsymbol{j}} \\ \hline U^{h,1}_{\boldsymbol{i},\boldsymbol{j}} \\ \hline U^{h,2}_{\boldsymbol{i},\boldsymbol{j}} \\ \hline U^{h,3}_{\boldsymbol{i},\boldsymbol{j}} \end{pmatrix},$$

then for each $t = 0, \ldots, 3$, we have the following relation:

$$U^{h,t}_{\boldsymbol{i}} \text{ corresponds to domain } A^{h+1}_{2\boldsymbol{i}+e_t} \times \Omega \tag{11.26}$$

$$U^{h,t}_{\boldsymbol{i},\boldsymbol{j}} \text{ corresponds to domain } A^{h+1}_{2\boldsymbol{i}+e_t} \times B^h_{\boldsymbol{j}}. \tag{11.27}$$

Combining the above split results, we have

$$U^h_{\boldsymbol{i}} = \begin{pmatrix} U^{h,0}_{\boldsymbol{i},\boldsymbol{j}_0} & U^{h,0}_{\boldsymbol{i},\boldsymbol{j}_1} & \ldots & U^{h,0}_{\boldsymbol{i},\boldsymbol{j}_{4^h-1}} \\ \hline U^{h,1}_{\boldsymbol{i},\boldsymbol{j}_0} & U^{h,1}_{\boldsymbol{i},\boldsymbol{j}_1} & \ldots & U^{h,1}_{\boldsymbol{i},\boldsymbol{j}_{4^h-1}} \\ \hline U^{h,2}_{\boldsymbol{i},\boldsymbol{j}_0} & U^{h,2}_{\boldsymbol{i},\boldsymbol{j}_1} & \ldots & U^{h,2}_{\boldsymbol{i},\boldsymbol{j}_{4^h-1}} \\ \hline U^{h,3}_{\boldsymbol{i},\boldsymbol{j}_0} & U^{h,3}_{\boldsymbol{i},\boldsymbol{j}_1} & \ldots & U^{h,3}_{\boldsymbol{i},\boldsymbol{j}_{4^h-1}} \end{pmatrix}. \tag{11.28}$$

This is the splitting step moving from level h to $h+1$ in T_X.

Next, the merging step merges adjacent matrices $U^{h,t}_{\boldsymbol{i},\boldsymbol{j}}$ to obtain low-rank matrices. Using the notation $\{e_t\}_{t=0,\ldots,3}$ again, we note that for each $\boldsymbol{i} = (i_1, i_2)$, $i_1, i_2 = 0, \ldots, 2^h - 1$, and $\boldsymbol{j} = (\boldsymbol{j}_1, \boldsymbol{j}_2)$,

$\boldsymbol{j}_1, \boldsymbol{j}_2 = 1, \ldots, 2^{h/2} - 1$, the column space of the submatrix

$$\begin{pmatrix} U^{h,t}_{\boldsymbol{i},2\boldsymbol{j}+e_0} & U^{h,t}_{\boldsymbol{i},2\boldsymbol{j}+e_1} & U^{h,t}_{\boldsymbol{i},2\boldsymbol{j}+e_2} & U^{h,t}_{\boldsymbol{i},2\boldsymbol{j}+e_3} \end{pmatrix} \in \mathbb{C}^{N/4 \times 2r} \tag{11.29}$$

in (11.28) are in the column space of $K^{h+1}_{2\boldsymbol{i}+e_t,\boldsymbol{j}}$ by the relation (11.27). By the complementary low-rank property of the matrix K, we know $K^{h+1}_{2\boldsymbol{i}+e_t,\boldsymbol{j}}$ corresponding to the kernel function restricted in $A^{h+1}_{2\boldsymbol{i}+e_t} \times B^{h-1}_{\boldsymbol{j}}$ is numerically low-rank. Hence, the matrix in (11.29) is a numerical low-rank matrix. By computing its rank-r approximation using the standard truncated SVD and (11.7), we have

$$\begin{pmatrix} U^{h,t}_{\boldsymbol{i},2\boldsymbol{j}+e_0} & U^{h,t}_{\boldsymbol{i},2\boldsymbol{j}+e_1} & U^{h,t}_{\boldsymbol{i},2\boldsymbol{j}+e_2} & U^{h,t}_{\boldsymbol{i},2\boldsymbol{j}+e_3} \end{pmatrix} \approx U^{h+1}_{2\boldsymbol{i}+e_t,\boldsymbol{j}} G^h_{2\boldsymbol{i}+e_t,\boldsymbol{j}} \tag{11.30}$$

for $\boldsymbol{i} = (i_1, i_2)$, $i_1, i_2 = 0, \ldots, 2^h - 1$, and $\boldsymbol{j} = (j_1, j_2)$, $j_1, j_2 = 1, \ldots, 2^{h-1} - 1$. This is the merging step equivalent to moving from level h to level $h-1$ in T_Ω.

In the assembling step, we construct the factorization in (11.23) for $\ell = h$ using the small factorizations in (11.30) as follows:

$$U^h \approx U^{h+1} G^h = \begin{pmatrix} U^{h+1}_{\boldsymbol{i}_0} & & & \\ & U^{h+1}_{\boldsymbol{i}_1} & & \\ & & \ddots & \\ & & & U^{h+1}_{\boldsymbol{i}_{4^{h+1}-1}} \end{pmatrix} \begin{pmatrix} G^h_{\boldsymbol{i}_0} & & & \\ & G^h_{\boldsymbol{i}_1} & & \\ & & \ddots & \\ & & & G^h_{\boldsymbol{i}_{4^h-1}} \end{pmatrix},$$

where

$$U^{h+1}_{\boldsymbol{i}} = \begin{pmatrix} U^{h+1}_{\boldsymbol{i},\boldsymbol{j}_0} & U^{h+1}_{\boldsymbol{i},\boldsymbol{j}_1} & \cdots & U^{h+1}_{\boldsymbol{i},\boldsymbol{j}_{4^{h-1}-1}} \end{pmatrix}$$

for $\boldsymbol{i} = \boldsymbol{i}_0, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_{4^{h+1}-1}$, and

$$G^h_{\boldsymbol{i}} = \begin{pmatrix} G^h_{2\boldsymbol{i}+e_0,\boldsymbol{j}_0} & & & \\ & G^h_{2\boldsymbol{i}+e_0,\boldsymbol{j}_1} & & \\ & & \ddots & \\ & & & G^h_{2\boldsymbol{i}+e_0,\boldsymbol{j}_{2^h-1}} \\ \hline G^h_{2\boldsymbol{i}+e_1,\boldsymbol{j}_0} & & & \\ & G^h_{2\boldsymbol{i}+e_1,\boldsymbol{j}_1} & & \\ & & \ddots & \\ & & & G^h_{2\boldsymbol{i}+e_1,\boldsymbol{j}_{2^h-1}} \\ \hline G^h_{2\boldsymbol{i}+e_2,\boldsymbol{j}_0} & & & \\ & G^h_{2\boldsymbol{i}+e_2,\boldsymbol{j}_1} & & \\ & & \ddots & \\ & & & G^h_{2\boldsymbol{i}+e_2,\boldsymbol{j}_{2^h-1}} \\ \hline G^h_{2\boldsymbol{i}+e_3,\boldsymbol{j}_0} & & & \\ & G^h_{2\boldsymbol{i}+e_3,\boldsymbol{j}_1} & & \\ & & \ddots & \\ & & & G^h_{2\boldsymbol{i}+e_3,\boldsymbol{j}_{2^h-1}} \end{pmatrix}$$

for $\boldsymbol{i} = \boldsymbol{i}_0, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_{4^h-1}$.

Since there are $O(1)$ nonzero entries in each $G^h_{\boldsymbol{i},\boldsymbol{j}}$ and there are $O(4^h * 4 * 4^{h-1}) = O(N^2)$ such matrices, there are only $O(N^2)$ nonzero entries in G^h.

In a similar way, we can factorize $U^\ell \approx U^{\ell+1} G^\ell$ for $h < \ell \leq L-1$, because the column space of

$$\begin{pmatrix} U^{\ell,t}_{\boldsymbol{i},2\boldsymbol{j}+e_0} & U^{\ell,t}_{\boldsymbol{i},2\boldsymbol{j}+e_1} & U^{\ell,t}_{\boldsymbol{i},2\boldsymbol{j}+e_2} & U^{\ell,t}_{\boldsymbol{i},2\boldsymbol{j}+e_3} \end{pmatrix} \tag{11.31}$$

is in the column space of $K^{\ell+1}_{2\boldsymbol{i}+e_t,\boldsymbol{j}}$. Computing the rank-$r$ approximations via the standard truncated SVD and (11.9) gives

$$\begin{pmatrix} U^{\ell,t}_{\boldsymbol{i},2\boldsymbol{j}+e_0} & U^{\ell,t}_{\boldsymbol{i},2\boldsymbol{j}+e_1} & U^{\ell,t}_{\boldsymbol{i},2\boldsymbol{j}+e_2} & U^{\ell,t}_{\boldsymbol{i},2\boldsymbol{j}+e_3} \end{pmatrix} \approx U^{\ell+1}_{2\boldsymbol{i}+e_t,\boldsymbol{j}} G^\ell_{2\boldsymbol{i}+e_t,\boldsymbol{j}} \tag{11.32}$$

for $\boldsymbol{i} = \boldsymbol{i}_0, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_{2^\ell-1}$ and $\boldsymbol{j} = \boldsymbol{j}_0, \boldsymbol{j}_1, \ldots, \boldsymbol{j}_{2^{L-\ell-1}-1}$. After assembling these factorizations together,

we obtain

$$U^\ell \approx U^{\ell+1}G^\ell = \begin{pmatrix} U^{\ell+1}_{\boldsymbol{i}_0} & & & \\ & U^{\ell+1}_{\boldsymbol{i}_1} & & \\ & & \ddots & \\ & & & U^{\ell+1}_{\boldsymbol{i}_{4^{\ell+1}-1}} \end{pmatrix} \begin{pmatrix} G^\ell_{\boldsymbol{i}_0} & & & \\ & G^\ell_{\boldsymbol{i}_1} & & \\ & & \ddots & \\ & & & G^\ell_{\boldsymbol{i}_{4^\ell-1}} \end{pmatrix},$$

where

$$U^{\ell+1}_{\boldsymbol{i}} = \begin{pmatrix} U^{\ell+1}_{\boldsymbol{i},\boldsymbol{j}_0} & U^{\ell+1}_{\boldsymbol{i},\boldsymbol{j}_1} & \cdots & U^{\ell+1}_{\boldsymbol{i},\boldsymbol{j}_{4^{L-\ell-1}-1}} \end{pmatrix}$$

for $\boldsymbol{i} = \boldsymbol{i}_0, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_{4^{\ell+1}-1}$, and

$$G^\ell_{\boldsymbol{i}} = \left(\begin{array}{cccc} G^\ell_{2\boldsymbol{i}+e_0,\boldsymbol{j}_0} & & & \\ & G^\ell_{2\boldsymbol{i}+e_0,\boldsymbol{j}_1} & & \\ & & \ddots & \\ & & & G^\ell_{2\boldsymbol{i}+e_0,\boldsymbol{j}_{2^{L-\ell-1}-1}} \\ \hline G^\ell_{2\boldsymbol{i}+e_1,\boldsymbol{j}_0} & & & \\ & G^\ell_{2\boldsymbol{i}+e_1,\boldsymbol{j}_1} & & \\ & & \ddots & \\ & & & G^\ell_{2\boldsymbol{i}+e_1,\boldsymbol{j}_{2^{L-\ell-1}-1}} \\ \hline G^\ell_{2\boldsymbol{i}+e_2,\boldsymbol{j}_0} & & & \\ & G^\ell_{2\boldsymbol{i}+e_2,\boldsymbol{j}_1} & & \\ & & \ddots & \\ & & & G^\ell_{2\boldsymbol{i}+e_2,\boldsymbol{j}_{2^{L-\ell-1}-1}} \\ \hline G^\ell_{2\boldsymbol{i}+e_3,\boldsymbol{j}_0} & & & \\ & G^\ell_{2\boldsymbol{i}+e_3,\boldsymbol{j}_1} & & \\ & & \ddots & \\ & & & G^\ell_{2\boldsymbol{i}+e_3,\boldsymbol{j}_{2^{L-\ell-1}-1}} \end{array}\right)$$

for $\boldsymbol{i} = \boldsymbol{i}_0, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_{4^\ell-1}$.

After $L-h$ steps of recursive factorizations

$$U^\ell \approx U^{\ell+1}G^\ell$$

for $\ell = h, h+1, \ldots, L-1$, we obtain the recursive factorization of U^h as

$$U^h \approx U^L G^{L-1} \cdots G^h. \tag{11.33}$$

Similar to the analysis of G^h, it is also easy to check that there are only $O(N^2)$ nonzero entries in each G^ℓ in (11.33). Since there are $O(N^2)$ diagonal blocks in U^L and each block contains $O(1)$ entries, there is $O(N^2)$ nonzero entries in U^L.

Recursive factorization of V^h

The recursive factorization of V^ℓ is similar to that of U^ℓ for $\ell = h, h+1, \ldots, L-1$. At each level ℓ, we benefit from the fact that $\begin{pmatrix} V^{\ell,t}_{\boldsymbol{j},2\boldsymbol{i}+e_0} & V^{\ell,t}_{\boldsymbol{j},2\boldsymbol{i}+e_1} & V^{\ell,t}_{\boldsymbol{j},2\boldsymbol{i}+e_2} & V^{\ell,t}_{\boldsymbol{j},2\boldsymbol{i}+e_3} \end{pmatrix}$ is in the row space of $K^{L-\ell-1}_{\boldsymbol{i},2\boldsymbol{j}+e_t}$ and hence is numerically low-rank. Applying the same procedure in Section 11.3.2 to V^h leads to

$$V^h \approx V^L H^{L-1}, \ldots, H^h. \tag{11.34}$$

Given the results of the middle level factorization in (11.20), recursive factorizations in (11.33) and (11.34), we arrive at the final butterfly factorization

$$K \approx U^L G^{L-1} \cdots G^h M^h \left(H^h\right)^* \cdots \left(H^{L-1}\right)^* \left(V^L\right)^*, \tag{11.35}$$

with all factors containing $O(N^2)$ non-zero entries.

11.3.3 Complexity Analysis

We split the complexity analysis of the construction of butterfly factorizations into two parts: the middle level factorization and the recursive factorization.

We have different complexity in the middle level factorization depending on the following conditions:

- In case (i), the dominant cost is to apply K and K^* to Gaussian random matrices of size $O(N^2 \times N)$. Assuming that the given blackbox routine for applying K and K^* once takes $O(C_K(N))$ operations, the total operation complexity is $O(C_K(N)N)$.
- In case (ii), we apply the randomized sampling algorithm to compute $O(N^2)$ submatrices of size $N \times N$. Since the operation complexity taken for each submatrix is $O(N)$, the overall complexity is $O(N^3)$.

In the recursive factorization, it takes the same operation complexity to factorize U^h and V^h. There are $O(\log(N))$ steps to factorize U^h. At the ℓ step, the matrix U^ℓ to be factorized consists of 4^ℓ diagonal blocks. There are $O(N^2/4^\ell)$ factorizations in every diagonal block and each factorization

takes $O(N^2/4^\ell)$ operations. Hence, the operation complexity to factorize U^ℓ is $O(N^4/4^\ell)$. Summing up all the operations in each step yields the overall operation complexity for recursively factorizing U^h:

$$\sum_{\ell=h}^{L-1} O(N^4/4^\ell) = O(N^3). \tag{11.36}$$

		Randomized SVD	Randomized sampling
Factorization Complexity	Middle level factorization	$O(C_K(N)N)$	$O(N^3)$
	Recursive factorization	$O(N^3)$	
	Total	$O(C_K(N)N)$	$O(N^3)$
Memory Complexity		$O(N^3)$	$O(N^2 \log N)$
Application Complexity		$O(N^2 \log N)$	

Table 11.1: Computational complexity and memory complexity of the two-dimensional butterfly factorization. $C_K(N)$ is the operation complexity of applying K and K^* once, e.g., $C_K(N) = O(N^2 \log N)$ for the butterfly algorithms in Section 11.2.

The memory peak of the butterfly factorization is due to the middle level factorization where we have to store the results of $O(N^2)$ factorizations of size $O(N)$. Hence, the memory complexity for the two-dimensional butterfly factorization is $O(N^3)$. By the same argument in [117] or in Section 10.3.3, we can lever the order of generation and recursive factorization of $U^h_{\boldsymbol{i},\boldsymbol{j}}$ and $V^h_{\boldsymbol{j},\boldsymbol{i}}$. If we factorize $U^h_{\boldsymbol{i},\boldsymbol{j}}$ and $V^h_{\boldsymbol{j},\boldsymbol{i}}$ individually instead of formulating (11.20), the memory complexity for case (ii) can be reduced to $O(N^2 \log N)$. Table 11.1 summarizes the complexity analysis for the two-dimensional butterfly factorization.

The storage and application complexity for the butterfly factorization is the number of nonzero entries in the final factorization, which is $O(N^2 \log N)$.

So far, we have introduced the two-dimensional butterfly factorization for a complementary low-rank kernel $K(x, \xi)$ in the whole domain $X \times \Omega$. Although we have assumed the uniform grid in (11.4) and (11.5), the butterfly factorization algorithm here does not rely on this grid. Actually, in the case when a quasi-uniform point set $X \times \Omega$ is given, we can still construct a butterfly factorization for $K(x, \xi)$ following the instruction in this section. On a fixed level ℓ, the number of points in each $A^\ell_{\boldsymbol{i}} \in T_X$ (or $B^\ell_{\boldsymbol{j}} \in T_\Omega$) might be different, but the numerical rank of $K^\ell_{\boldsymbol{i},\boldsymbol{j}}$ can still be bounded by a uniform constant r according to theorems in Section 11.2. Although we will encounter different

sizes of small factorization, the overall complexity analysis is still true for a quasi-uniform point set $X \times \Omega$.

11.4 Polar Butterfly Factorization

In Section 11.5, we have introduced a two-dimensional butterfly factorization for a complementary low-rank kernel $K(x, \xi)$ in the whole domain $X \times \Omega$. In this section, we will introduce a polar butterfly factorization method to deal with special kernel functions $K(x, \xi) = e^{2\pi\imath\Phi(x,\xi)}$ with irregularity at $\xi = 0$ encountered in FIOs.

The polar butterfly factorization refers to the idea of the polar butterfly algorithm in Section 11.2.2 that after the polar transformation from $\xi = (\xi_1, \xi_2)$ to $p = (p_1, p_2)$:

$$\xi = \frac{\sqrt{2}}{2} N p_1 e^{2\pi\imath p_2}, \quad e^{2\pi\imath p_2} = (\cos 2\pi p_2, \sin 2\pi p_2),$$

the new function

$$\Psi(x, p) := \frac{1}{N}\Phi(x, \xi) = \frac{\sqrt{2}}{2}\Phi(x, e^{2\pi\imath p_2})p_1,$$

is smooth in the whole domain $X \times P$, where $P = (0, 1)^2$. This leads to a complementary low-rank property of $e^{2\pi\imath N\Psi(x,p)}$ in the whole domain $X \times P$ as proved by Theorem 11.2.1. This inspires the polar butterfly factorization is as follows:

1. *Preliminary.* Reformulate the problem

$$u(x) = \sum_{\xi\in\Omega} e^{2\pi\imath\Phi(x,\xi)} g(\xi), \quad x \in X, \tag{11.37}$$

   into

$$u(x) = \sum_{p\in P} e^{2\pi\imath N\Psi(x,p)} g(p), \quad x \in X. \tag{11.38}$$

2. *Factorization.* Apply the two-dimensional butterfly factorization to the kernel $e^{2\pi\imath N\Psi(x,p)}$ defined on a nonuniform point set in $X \times P$ to compute the butterfly factorization

$$K \approx U^L G^{L-1} \cdots G^h M^h \left(H^h\right)^* \cdots \left(H^{L-1}\right)^* \left(V^L\right)^*, \tag{11.39}$$

3. *Application.* Transform the given data $g(\xi)$ into $g(p)$ on a polar grid. Multiply the factorization above to $g(p)$.

The polar butterfly factorization and the original butterfly factorization have the same complexity as summarized in Table 11.1.

11.5 Multiscale Butterfly Factorization

In this section, we will introduce another idea to address the point irregularity of $K(x,\xi) = e^{2\pi\imath\Phi(x,\xi)}$ at $\xi = 0$. Recall that the multiscale butterfly algorithm partitions the domain Ω into multiscale subdomains as follows:

$$u(x) = \sum_{\xi\in\Omega} e^{2\pi\imath\Phi(x,\xi)} g(\xi) = \sum_{\xi\in\Omega_d} e^{2\pi\imath\Phi(x,\xi)} g(\xi) + \sum_j \sum_{\xi\in\Omega_j} e^{2\pi\imath\Phi(x,\xi)} g(\xi), \tag{11.40}$$

where

$$\Omega_j = \{(\xi_1,\xi_2) : \frac{N}{2^{j+2}} < \max(|\xi_1|,|\xi_2|) \le \frac{N}{2^{j+1}}\} \cap \Omega, \tag{11.41}$$

for $j = 0, 1, \ldots, \log_2 N - s$, s is a small constant, and $\Omega_d = \Omega \setminus \cup_j \Omega_j$. Since the kernel $e^{2\pi\imath\Phi(x,\xi)}$ is complementary low-rank when it is restricted in $X \times \Omega_j$, we can apply the two-dimensional butterfly factorization in each subdomain. This motivates the multiscale butterfly factorization below.

1. *Preliminary.* Decompose domain Ω into subdomains as in (11.41). Reformulate the problem into a multiscale summation according to (11.40):

   $$u = Kg = K_{X,\Omega_d} g_{\Omega_d} + \sum_j K_{X,\Omega_j} g_{\Omega_j}, \tag{11.42}$$

   where subscripts denote the domains to which the submatrices and subvectors are corresponding.

2. *Factorization.* Let $L = \log_2 N$ and s be a small constant. For each $j = 0, 1, \ldots, L - s$, apply the two-dimensional butterfly factorization on $K(x,\xi) = e^{2\pi\imath\Phi(x,\xi)}$ restricted in $X \times \Omega_j$.

   Note that each domain Ω_j contains an empty block in the middle. Factorizing the kernel $e^{2\pi\imath\Phi(x,\xi)}$ restricted in $X \times \Omega_j$ is equivalent to factorizing a new kernel

   $$\tilde{K}(x,\xi) = \begin{cases} e^{2\pi\imath\Phi(x,\xi)}, & \text{for } \xi \in \Omega_j; \\ 0, & \text{for } \xi \in \cup_{k>j}\Omega_k \end{cases} \tag{11.43}$$

   restricted in $X \times \cup_{k\ge j}\Omega_k$. Let $\tilde{\Omega}_j = \cup_{k\ge j}\Omega_k$. We construct two quadtrees T_X and $T_{\tilde{\Omega}_j}$ by hierarchically dyadic partition. Let X and $\tilde{\Omega}_j$ be the roots of T_X and $T_{\tilde{\Omega}_j}$ (corresponding to level 0 in the quadtrees). Let $L_j = 2\lfloor (L-j)/2 \rfloor$, where $\lfloor\cdot\rfloor$ is the largest integer leas than or equal to a given number. To make sure that T_X and $T_{\tilde{\Omega}_j}$ have the same depth, we will partition X and $\tilde{\Omega}_j$ until we reach the L_jth level.

Applying the two-dimensional butterfly factorization using the quadtrees T_X and $T_{\tilde{\Omega}_j}$ constructed above gives the jth butterfly factorization below

$$K_{X,\Omega_j} \approx U^{j,L_j} G^{j,L_j-1} \cdots G^{j,\frac{L_j}{2}} M^{j,\frac{L_j}{2}} \left(H^{j,\frac{L_j}{2}}\right)^* \cdots \left(H^{j,L_j-1}\right)^* \left(V^{j,L_j}\right)^*.$$

Note that $\tilde{K}(x,\xi)$ is zero when $\xi \in \cup_{k>j}\Omega_k$. There are $\frac{1}{4}$ small matrices to be factorized in the process of butterfly factorization are zero matrices. We can simply ignore the computation for these matrices.

Once we have computed all butterfly factorizations, then the multiscale summation in (11.42) becomes

$$u = K_{X,\Omega_d} g_{\Omega_d} + \sum_j U^{j,L_j} G^{j,L_j-1} \cdots M^{j,\frac{L_j}{2}} \cdots \left(H^{j,L_j-1}\right)^* \left(V^{j,L_j}\right)^* g_{\Omega_j} \tag{11.44}$$

3. *Application* In practice, after computing the butterfly factorization for K_{X,Ω_j}, we only store the nonzero submatrices of its factors. When an input vector g is given, we divide g into g_{Ω_d}, g_{Ω_j}, $j = 0, \ldots, L-s$, and evaluate (11.44) for u.

The overall factorization and application complexity of the multiscale butterfly factorization is the same as that of the regular butterfly factorization as summarized in Table 11.1.

In the case (i) when we have a blackbox routine for apply K and its adjoint, the middle level factorizations of all butterfly factorization on different scale are conducted simultaneously to maintain $C_K(N)N$ operation complexity. In case (ii) when we can evaluate individual entries of K, randomized sampling evaluates $O(N^3)$ entries. In the middle level factorization on the jth scale, since the number of matrices to be factorized is $O(N^2/4^j)$ and the complexity to factorize one matrix is $O(N2^j)$, the operation complexity for the middle level factorization after multiplying Gaussian random matrices or randomized sampling is $O(N^3/2^j)$. Hence, the overall operational complexity of the middle level factorization for the multiscale butterfly factorization is equal to the one for the regular butterfly factorization.

In the recursive factorization on the jth scale, the one-step factorization $U^{j,\frac{L_j}{2}} \approx U^{j,\frac{L_j}{2}+1} G^{j,\frac{L_j}{2}+1}$ takes $O(N^3/2^j)$ operations and hence the recursive factorization $U^{j,\frac{L_j}{2}+1}$ takes $O(N^3/2^j)$ operations. The recursive factorization of $V^{j,\frac{L_j}{2}}$ is cheaper since the matrix is smaller. Hence, it takes $O(N^3/2^j)$ operations to factorize K_{X,Ω_j} resulting in an $O(N^3)$ operation complexity for the overall recursive factorization of all K_{X,Ω_j}.

Since there are only $O(N^2 \log N)$ nonzero entries in (11.44), the application complexity of the multiscale butterfly factorization is $O(N^2 \log N)$.

The memory peak of the multiscale butterfly factorization is also in the middle level factorization. Similar analysis shows the same memory complexity as that of the regular butterfly factorization.

11.6 Conclusion

This chapter introduces a multi-dimensional butterfly factorization method to provide a data-sparse approximation of multi-dimensional complementary low-rank matrices. We also provide examples and methods to deal with some locally multi-dimensional complementary low-rank matrices (e.g., multi-dimensional FIO kernels). The butterfly factorization gives rise to highly efficient matrix-vector multiplications with $O(N^d \log N)$ operation and memory complexity, where d is the number of dimensions.

Chapter 12

Conclusions of Part II

12.1 Summary

The second part of this thesis has introduced several fast algorithms for oscillatory integral operators in computational harmonic analysis. They can be applied to accelerate the application of Fourier integral operators (including pseudo differential operators, the generalized Radon transform, the nonuniform Fourier transform, etc.) and special function transforms (including the Fourier-Bessel transform, the spherical harmonic transform, etc.). Since these operators and transforms have been playing an important role in many engineering problems, these fast algorithms are of great interest especially in large-scale numerical simulations.

The proposed fast algorithms fall into two types. The first type of algorithms is useful when an oscillatory integral operator with a complementary low-rank kernel $K(x, \xi)$ is applied for only a few times. In this case, we can apply the multiscale butterfly algorithm. This algorithm requires linear memory complexity and quasilinear operation complexity without precomputation to apply $K(x, \xi)$. The second type of algorithms is useful when the integral operator is repeatedly applied. This type of algorithms represents the discrete analogue of K as a product of a few data-sparse matrices with nearly optimal number of nonzero entries. Although constructing these matrix factors is more expensive than applying the butterfly algorithm once, storing and applying K becomes optimally fast with a small prefactor after the butterfly factorization.

12.2 Future Work

Although the application complexity of the butterfly factorization is optimal, the construction complexity of the current method is far from optimal. It is interesting to see whether the construction complexity can be reduced under a reasonably stronger condition than what we have assumed in this thesis. Instead of designing a universal tool, one could construct a butterfly factorization for

specific kernel functions with optimal complexity.

Another important direction is to compute the inverse of a complementary low-rank matrix or its butterfly factorization efficiently. The butterfly factorization in this thesis has degenerate factors and hence we cannot invert the original matrix K via the butterfly factorization. It is interesting to see whether we could construct invertible butterfly factorizations. This is of great significance because many transforms have both forward and inverse transforms via a complementary low-rank matrix.

While the numerical results of this thesis include a couple of examples, it is important to prove the complementary low-rank property with rigorous mathematical analysis. It is also natural to consider other important class of transforms. A software library containing basic routines for evaluating various transforms via the proposed algorithms is of great practical use.

Appendix A

A Long Proof of the Robustness

A.1 Proofs for the Theorems in Section 3.3

Proof of Theorem 3.3.1

Proof. We only sketch out the proof of this theorem, because its proof is similar to the proof in Theorem 3.2.1. By the definition of 2D wave packet transform and Lemma 2.2.8 and 2.2.9, we obtain the following two estimates:

$$|W_e(a,b)| \lesssim \sqrt{\epsilon_1}|a|^{-s}, \tag{A.1}$$

and

$$|\nabla_b W_e(a,b)| \lesssim \sqrt{\epsilon}_1 \left(1+|a|^{1-s}\right). \tag{A.2}$$

If $(a,b) \in R_\delta$, then $W_g(a,b)| \geq |a|^{-s}\delta$ and Equation (A.1) imply

$$|W_f(a,b)| \geq |a|^{-s}\sqrt{\epsilon}. \tag{A.3}$$

Hence, $S_\delta \subset R_\delta \subset R_\epsilon$, where R_ϵ is defined in Theorem 2.2.7 and is a subset of $\bigcup_{1\leq k\leq K} Z_k$. So, (i) is true by Theorem 2.2.7. As for (ii), since $R_\delta \subset R_\epsilon$, then $(a,b) \in R_\delta \cap Z_k$ implies $(a,b) \in R_\epsilon \cap Z_k$ and $|a| \simeq N_k$. By Theorem 2.2.7, we have

$$\frac{|v_f(a,b) - N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|} \lesssim \sqrt{\epsilon},$$

when $N > N_0$. Hence,

$$\begin{aligned}
& \frac{|v_g(a,b) - N_k \nabla \phi_k(b)|}{|N_k \nabla \phi_k(b)|} \\
\leq & \frac{|v_f(a,b) - N_k \nabla \phi_k(b)|}{|N_k \nabla \phi_k(b)|} + \frac{|v_f(a,b) - v_g(a,b)|}{|N_k \nabla \phi_k(b)|} \\
\lesssim & \sqrt{\epsilon} + \frac{|W_e(a,b)|}{|W_g(a,b)|} + \frac{|\nabla_b W_e(a,b)|}{N_k |W_g(a,b)|} \\
\lesssim & \sqrt{\epsilon} + \frac{\sqrt{\epsilon}_1}{\delta} \\
\lesssim & \sqrt{\epsilon} + \epsilon_1^p,
\end{aligned}$$

when $N > N_0$. With a similar argument, when $(a,b) \in S_\delta \cap Z_k$, we can show that

$$\begin{aligned}
& \frac{|v_g(a,b) - N_k \nabla \phi_k(b)|}{|N_k \nabla \phi_k(b)|} \\
\lesssim & \frac{\sqrt{\epsilon}}{N_k^s} + \frac{|W_e(a,b)|}{|W_g(a,b)|} + \frac{|\nabla_b W_e(a,b)|}{N_k |W_g(a,b)|} \\
\lesssim & \frac{\sqrt{\epsilon}}{N_k^s} + \frac{\sqrt{\epsilon}_1}{a^s \delta} \\
\lesssim & \frac{\sqrt{\epsilon} + \epsilon_1^p}{N_k^s},
\end{aligned}$$

when $N > N_0$. □

Proof of Theorem 3.3.2

Proof. The sketch of the proof of this theorem is similar to the proof of Theorem 3.2.2 and 3.2.3, but much trickier. Since $w_{ab} \in L^1 \cap C^m$ and $\nabla_b w_{ab} \in L^1 \cap C^{m-1} \times L^1 \cap C^{m-1}$, we know $W_g(a,b) = W_f(a,b) + W_e(a,b)$ and $\nabla_b W_g(a,b) = \nabla_b W_f(a,b) + \nabla_b W_e(a,b)$ are Gaussian random variables. By the properties of zero mean stationary Gaussian processes and the geometric supports of wave packets in the frequency domain, we can still check that $W_e(a,b)$ and $(W_e(a,b), \partial_{b_1} W_e(a,b), \partial_{b_2} W_e(a,b))$ have nearly zero pseudo-covariance matrices. Hence, they are nearly circularly symmetric. We also divide the proof into two steps.

Step 1: We prove the case when the mother wave packet is of type $(0,m)$.

In this case, $W_e(a,b)$ and $(W_e(a,b), \partial_{b_1} W_e(a,b), \partial_{b_2} W_e(a,b))$ are circularly symmetric. The variance of $W_e(a,b)$ is $\int_{\mathbb{R}^2} |\widehat{w}(\xi)|^2 \widehat{e}\,(|a|^s \xi + a)\, d\xi$, which is denoted by σ^2. Suppose that $\Xi = (\Xi_1, \Xi_2)^T$ is a real random vector with a joint probability density function $h(\xi) = \sigma^{-2} |\widehat{w}(\xi)|^2 \widehat{e}\,(|a|^s \xi + a)$, then

the covariance matrix of $(W_e(a,b), \partial_{b_1} W_e(a,b), \partial_{b_2} W_e(a,b))$ is $\sigma^2 V$, where V is the matrix below:

$$\begin{pmatrix} 1 & -2\pi i\mathbb{E}\left[|a|^s\Xi_1 + a_1\right] & -2\pi i\mathbb{E}\left[|a|^s\Xi_2 + a_2\right] \\ 2\pi i\mathbb{E}\left[|a|^s\Xi_1 + a_1\right] & 4\pi^2\mathbb{E}\left[\left(|a|^s\Xi_1 + a_1\right)^2\right] & 4\pi^2\mathbb{E}\left[\left(|a|^s\Xi_1 + a_1\right)\left(|a|^s\Xi_2 + a_2\right)\right] \\ 2\pi i\mathbb{E}\left[|a|^s\Xi_2 + a_2\right] & 4\pi^2\mathbb{E}\left[\left(|a|^s\Xi_1 + a_1\right)\left(|a|^s\Xi_2 + a_2\right)\right] & 4\pi^2\mathbb{E}\left[\left(|a|^s\Xi_2 + a_2\right)^2\right] \end{pmatrix}.$$

The distributions of $W_e(a,b)$ and $(W_e(a,b), \partial_{b_1} W_e(a,b), \partial_{b_2} W_e(a,b))$ are described by the probability density functions

$$\frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^{-2}}$$

and

$$\frac{e^{-\sigma^{-2}z^* V^{-1} z}}{\pi^3\sigma^6 \det V},$$

where $z = (z_1, z_2, z_3)^T$. Part (i) is true by previous theorems. To prove Part (ii) to (v), we need to define the following events

$$G_1 = \left\{|W_e(a,b)| < |a|^{-s} M_a^{1/(2+2q)}\right\},$$

$$G_2 = \left\{|W_e(a,b)| < M_a^{1/(2+2q)}\right\},$$

$$G_3 = \left\{|\nabla_b W_e(a,b)| < M_a^{1/(2+2q)}\left(1 + |a|^{1-s}\right)\right\},$$

$$H_k = \left\{\frac{|v_g(a,b) - N_k \nabla_b \phi_k(b)|}{|N_k \phi_k'(b)|} \lesssim \sqrt{\epsilon} + M_a^{p/(1+q)}\right\},$$

and

$$J_k = \left\{\frac{|v_g(a,b) - N_k \nabla_b \phi_k(b)|}{|N_k \phi_k'(b)|} \lesssim N_k^{-s}\left(\sqrt{\epsilon} + M_a^{p/(1+q)}\right)\right\},$$

for $1 \le k \le K$. Next, we are going to estimate the probability $P(G_1)$, $P(G_2)$, $P(G_1 \cap G_3)$, $P(G_2 \cap G_3)$, $P(H_k)$ and $P(J_k)$. By the calculations above, we have

$$P(G_1) = \int_{|z_1| < |a|^{-s} M_a^{1/(2+2q)}} \frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^{-2}} dz_1 = 1 - e^{-|a|^{-2s}\sigma^{-2} M_a^{1/(1+q)}} \ge 1 - e^{-|a|^{-2s} M_a^{-q/(1+q)}},$$

and similarly

$$P(G_2) = \int_{|z_1| < M_a^{1/(2+2q)}} \frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^{-2}} dz_1 \ge 1 - e^{-M_a^{-q/(1+q)}}.$$

We are now ready to conclude (ii) and (iii). If $(a,b) \in R_{\delta_a}$, then

$$|W_e(a,b) + W_f(a,b)| \ge |a|^{-s}\left(M_a^{(1/2-p)/(1+q)} + \sqrt{\epsilon}\right). \tag{A.4}$$

If $(a,b) \notin \bigcup_{1\leq k\leq K} Z_k$, then by Lemma 2.2.9,

$$|W_f(a,b)| \leq |a|^{-s}\epsilon. \tag{A.5}$$

$|W_e(a,b)| \geq |a|^{-s} M_a^{1/(2+2q)}$ follows from Equation (A.4) and (A.5). Hence,

$$P\left((a,b) \notin \bigcup_{1\leq k\leq K} Z_k\right) \leq P\left(|W_e(a,b)| \geq |a|^{-s} M_a^{1/(2+2q)}\right) = 1 - P\left(G_1\right).$$

This means that if $(a,b) \in R_{\delta_a}$, then $(a,b) \in \bigcup_{1\leq k\leq K} Z_k$ with a probability at least $P\left(G_1\right) \geq 1 - e^{-|a|^{-2s} M_a^{-q/(1+q)}} = 1 - e^{-O\left(N_k^{-2s} M_a^{-q/(1+q)}\right)}$, since $|a| \simeq N_k$ if $(a,b) \in Z_k$. So, (ii) is true. A similar argument applied to $(a,b) \in S_{\delta_a}$ shows that $(a,b) \in \bigcup_{1\leq k\leq K} Z_k$ with a probability at least $P\left(G_2\right) = 1 - e^{-O\left(M_a^{-q/(1+q)}\right)}$. Hence, (iii) is proved.

Because V is invertible and self-adjoint, there exist a unitary matrix U and a diagonal matrix D such that $V^{-1} = U^* D U$. For $z \in \mathbb{C}^3$, let $z' = Uz$. Introduce notations $\delta_1 = |a|^{-s} M_a^{1/(2+3q)}$, $\delta_2 = M_a^{1/(2+3q)}$, $\delta_3 = \left(1 + |a|^{1-s}\right) M_a^{1/(2+3q)}$, $d_1 = \min\{\frac{\delta_1}{\sqrt{2}}, \frac{\delta_3}{2}\}$, and $d_2 = \min\{\frac{\delta_2}{\sqrt{2}}, \frac{\delta_3}{2}\}$. Similar to the proof in Theorem 3.2.2 and 3.2.3, by a simple property of high dimensional polydisk, we have

$$\begin{aligned}
P\left(G_1 \cap G_3\right) &= \int_{\{|z_1|<\delta_1, |z_2|^2+|z_3|^2<\delta_3^2\}} \frac{e^{-\sigma^{-2} z^* V^{-1} z}}{\pi^3 \sigma^6 \det V} dz_1 dz_2 dz_3 \\
&\geq \int_{\{|z_1|<\delta_1, |z_2|<\frac{\delta_3}{\sqrt{2}}, |z_3|<\frac{\delta_3}{\sqrt{2}}\}} \frac{e^{-\sigma^{-2} z^* V^{-1} z}}{\pi^3 \sigma^6 \det V} dz_1 dz_2 dz_3 \\
&= \int_{\{\{|z_1|<\delta_1, |z_2|<\frac{\delta_3}{\sqrt{2}}, |z_3|<\frac{\delta_3}{\sqrt{2}}\}} \frac{e^{-M_a^{-1}\left(D_{11}|z_1'|^2 + D_{22}|z_2'|^2 + D_{33}|z_3'|^2\right)}}{\pi^3 \sigma^6 \det V} dz_1' dz_2' dz_3' \\
&\geq \int_{\{|z_1'|<d_1, |z_2'|<d_1, |z_3'|<d_1\}} \frac{e^{-M_a^{-1}\left(D_{11}|z_1'|^2 + D_{22}|z_2'|^2 + D_{33}|z_3'|^2\right)}}{\pi^3 \sigma^6 \det V} dz_1' dz_2' dz_3' \\
&= \left(1 - e^{-\frac{D_{11} d_1^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22} d_1^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{33} d_1^2}{\sigma^2}}\right),
\end{aligned}$$

and similarly

$$\begin{aligned}
P\left(G_2 \cap G_3\right) &= \int_{\{|z_1|<\delta_2, |z_2|^2+|z_3|^2<\delta_2^2\}} \frac{e^{-\sigma^{-2} z^* V^{-1} z}}{\pi^3 \sigma^6 \det V} dz_1 dz_2 dz_3 \\
&\geq \left(1 - e^{-\frac{D_{11} d_2^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22} d_2^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{33} d_2^2}{\sigma^2}}\right).
\end{aligned}$$

Next, we are going to estimate the asymptotic behavior of D_{11}, D_{22} and D_{33} as $|a|$ increases. This relies on the estimates of $D_{11}D_{22}D_{33}$, $D_{11} + D_{22} + D_{33}$ and $D_{11}^{-1} + D_{22}^{-1} + D_{33}^{-1}$ as follows. Careful

algebraic calculation shows that

$$\begin{aligned}
\frac{\det V}{16\pi^4} &= \mathbb{E}\left[(|a|^s\Xi_1+a_1)^2\right]\mathbb{E}\left[(|a|^s\Xi_2+a_2)^2\right]-\mathbb{E}^2\left[(|a|^s\Xi_1+a_1)(|a|^s\Xi_2+a_2)\right]\\
&\quad +2\mathbb{E}\left[|a|^s\Xi_1+a_1\right]\mathbb{E}\left[|a|^s\Xi_2+a_2\right]\mathbb{E}\left[(|a|^s\Xi_1+a_1)(|a|^s\Xi_2+a_2)\right]\\
&\quad -\mathbb{E}^2\left[|a|^s\Xi_2+a_2\right]\mathbb{E}\left[(|a|^s\Xi_1+a_1)^2\right]-\mathbb{E}^2\left[|a|^s\Xi_1+a_1\right]\mathbb{E}\left[(|a|^s\Xi_2+a_2)^2\right]\\
&= \mathbb{E}\left[(|a|^s\Xi_1)^2\right]\mathbb{E}\left[(|a|^s\Xi_2)^2\right]-\mathbb{E}^2\left[(|a|^s\Xi_1)(|a|^s\Xi_2)\right]\\
&\quad +2\mathbb{E}\left[|a|^s\Xi_1\right]\mathbb{E}\left[|a|^s\Xi_2\right]\mathbb{E}\left[(|a|^s\Xi_1)(|a|^s\Xi_2)\right]\\
&\quad -\mathbb{E}^2\left[|a|^s\Xi_2\right]\mathbb{E}\left[(|a|^s\Xi_1)^2\right]-\mathbb{E}^2\left[|a|^s\Xi_1\right]\mathbb{E}\left[(|a|^s\Xi_2)^2\right]\\
&= |a|^{4s}\big(\mathbb{E}\left[\Xi_1^2\right]\mathbb{E}\left[\Xi_2^2\right]-\mathbb{E}^2\left[\Xi_1\Xi_2\right]+2\mathbb{E}\left[\Xi_1\right]\mathbb{E}\left[\Xi_2\right]\mathbb{E}\left[\Xi_1\Xi_2\right]\\
&\quad -\mathbb{E}^2\left[\Xi_2\right]\mathbb{E}\left[\Xi_1^2\right]-\mathbb{E}^2\left[\Xi_1\right]\mathbb{E}\left[\Xi_2^2\right]\big)\\
&\lesssim |a|^{4s}.
\end{aligned}$$

Hence,

$$D_{11}D_{22}D_{33}=\det\left(V^{-1}\right)\gtrsim |a|^{-4s}. \tag{A.6}$$

Similarly, we know

$$\begin{aligned}
\operatorname{trace}\left(V^{-1}\right) &= \frac{1}{\det V}\Big(16\pi^4\mathbb{E}\left[(|a|^s\Xi_1+a_1)^2\right]\mathbb{E}\left[(|a|^s\Xi_2+a_2)^2\right]\\
&\quad -16\pi^4\mathbb{E}^2\left[(|a|^s\Xi_1+a_1)(|a|^s\Xi_2+a_2)\right]+4\pi^2\mathbb{E}\left[(|a|^s\Xi_2+a_2)^2\right]\\
&\quad -4\pi^2\mathbb{E}^2\left[(|a|^s\Xi_1+a_1)\right]+4\pi^2\mathbb{E}\left[(|a|^s\Xi_1+a_1)^2\right]-4\pi^2\mathbb{E}^2\left[(|a|^s\Xi_1+a_1)\right]\Big)\\
&\simeq \frac{1}{\det V}\left(\mathbb{E}\left[(|a|^s\Xi_1+a_1)^2\right]\mathbb{E}\left[(|a|^s\Xi_2+a_2)^2\right]-\mathbb{E}^2\left[(|a|^s\Xi_1+a_1)(|a|^s\Xi_2+a_2)\right]\right)\\
&\simeq \frac{|a|^{2+2s}}{\det V}.
\end{aligned}$$

Therefore,

$$D_{11}+D_{22}+D_{33}=\operatorname{trace}(V^{-1})\simeq\frac{|a|^{2+2s}}{\det V}. \tag{A.7}$$

Note that

$$\operatorname{trace}(V)=1+4\pi^2\mathbb{E}\left[(|a|^s\Xi_1+a_1)^2\right]+4\pi^2\mathbb{E}\left[(|a|^s\Xi_2+a_2)^2\right]\simeq|a|^2,$$

then

$$D_{11}^{-1}+D_{22}^{-1}+D_{33}^{-1}\simeq|a|^2. \tag{A.8}$$

Equation (A.6), (A.7) and (A.8) imply $D_{11} \gtrsim |a|^{2-2s}$, $D_{22} \simeq |a|^{-2s}$ and $D_{33} \simeq |a|^{-2}$. Therefore,

$$\begin{aligned} P(G_1 \cap G_3) &\geq \left(1 - e^{-\frac{D_{11}d_1^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22}d_1^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{33}d_1^2}{\sigma^2}}\right) \\ &\simeq \left(1 - e^{-O\left(|a|^{2-4s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(|a|^{-4s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(|a|^{-2s-2}M_a^{-q/(1+q)}\right)}\right). \end{aligned}$$

A similar argument leads to

$$\begin{aligned} P(G_2 \cap G_3) &\geq \left(1 - e^{-\frac{D_{11}d_2^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22}d_2^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{33}d_2^2}{\sigma^2}}\right) \\ &\simeq \left(1 - e^{-O\left(|a|^{2-2s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(|a|^{-2s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(|a|^{-2}M_a^{-q/(1+q)}\right)}\right). \end{aligned}$$

By Theorem 3.3.1, if $(a,b) \in R_{\delta_a} \cap Z_k$ for some k, then

$$P(H_k) \geq P(H_k|G_1 \cap G_3)\, P(G_1 \cap G_3) = P(G_1 \cap G_3).$$

Note that $|a| \simeq N_k$ when $(a,b) \in Z_k$, then

$$P(H_k) \geq \left(1 - e^{-O\left(N_k^{2-4s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-4s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-2s-2}M_a^{-q/(1+q)}\right)}\right).$$

Similarly, if $(a,b) \in S_\delta \cap Z_k$ for some k, then

$$P(J_k) \geq \left(1 - e^{-O\left(N_k^{2-2s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-2s}M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-2}M_a^{-q/(1+q)}\right)}\right).$$

These arguments prove (iv) and (v).

Step 2: we prove the case for a mother wave packet of type (ϵ, m) such that

$$m \geq \max\left\{\frac{2(1+s)}{1-s}, \frac{2}{1-s} + 4\right\}.$$

Larger m keeps our approximation errors sufficiently small.

Now $W_e(a,b)$ and $(W_e(a,b), \partial_{b_1}W_e(a,b), \partial_{b_2}W_e(a,b))$ have nearly zero Pseudo-covariance matrices and they are nearly circularly symmetric. Suppose they have covariance matrices C_1 and C_2, pseudo-covariance matrices P_1 and P_2, respectively. We can still check that they have zero mean and $C_1 = \sigma^2$ and $C_2 = \sigma^2 V$, where V is defined in the first step. By the definition of the 2D mother wave packet of type (ϵ, m) and the same process in the proof of Theorem 3.2.2 and 3.2.3, we can obtain a almost similar result:

1. The covariance matrix of $(W_e(a,b), W_e^*(a,b))$ is

$$V_1 = \begin{pmatrix} C_1 & P_1 \\ P_1^* & C_1^* \end{pmatrix},$$

and the distribution of $W_e(a,b)$ is described by

$$\frac{e^{-\frac{1}{2}(z_1^*,z_1)V_1^{-1}(z_1,z_1^*)^T}}{\pi\sqrt{\det V_1}},$$

which is well approximated by

$$\frac{e^{-\frac{C_1|z_1|^2-\Re\epsilon\left(P_1^*z_1^2\right)}{C_1^2-P_1P_1^*}}}{\pi\sqrt{C_1^2-P_1P_1^*}}=\frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^2}\left(1+O\left(\frac{\epsilon|z_1|^2}{\sigma^2|a|^{m(1-s)}}\right)\right).$$

2. The covariance matrix of $(W_e(a,b),\partial_bW_e(a,b),W_e^*(a,b),\partial_bW_e^*(a,b))$ is

$$V_2=\begin{pmatrix}C_2 & P_2\\ P_2^* & C_2^*\end{pmatrix},$$

and its distribution is described by the joint probability density

$$\frac{e^{-\frac{1}{2}(z_1^*,z_2^*,z_3^*,z_1,z_2,z_3)V_2^{-1}(z_1,z_2,z_3,z_1^*,z_2^*,z_3^*)^T}}{\pi^3\sqrt{\det V_2}},$$

which is well approximated by

$$\frac{e^{-\sigma^{-2}z^*V^{-1}z}e^{-\frac{1}{2}(z_1^*,z_2^*,z_3^*,z_1,z_2,z_3)P_\epsilon(z_1,z_2,z_3,z_1^*,z_2^*,z_3^*)^T}}{\pi^3\sigma^6\sqrt{(\det V)^2+O\left(\frac{\epsilon}{|a|^{m-2-(m+6)s}}\right)}},$$

where P_ϵ is a matrix with 2-norm bounded by $O\left(\frac{\epsilon}{\sigma^2|a|^{(m-4)(1-s)}}\right)$.

The only different result is the determinant error bound $O\left(\frac{\epsilon}{|a|^{m-2-(m+6)s)}}\right)$. Since the matrix V here has positive eigenvalues bounded above by $O(|a|^2)$, $O(|a|^{2s})$ and $O(|a|^{2(s-1)})$, then C_2 has positive eigenvalues bounded above by $O(\sigma^2|a|^2)$, $O(\sigma^2|a|^{2s})$ and $O(\sigma^2|a|^{2(s-1)})$. Because every entry in P_2 is bounded by $O\left(\frac{\sigma^2\epsilon}{|a|^{m(1-s)}}\right)$, then determinant error bound comes from

$$O\left(|a|^2|a|^2|a|^{2s}|a|^{2s}|a|^{2(s-1)}\frac{\epsilon}{|a|^{(m-4)(1-s)}}\right)=O\left(\frac{\epsilon}{|a|^{m-2-(m+6)s}}\right).$$

By the same argument in the first step, we can show that there exist a diagonal matrix $D=\mathrm{diag}\{D_{11},D_{22},D_{33}\}$ and a unitary matrix U such that $V^{-1}=U^*DU$. Furthermore, $D_{11}\gtrsim|a|^{2(1-s)}$, $D_{22}\simeq|a|^{-2s}$, and $D_{33}\simeq|a|^{-2}$. Part (i) is still true by previous theorems. To conclude Part (ii) to (v), we still need to estimate the probability of those events defined in the first step, i.e., $P(G_1)$,

$P(G_2)$, $P(G_1 \cap G_3)$, $P(G_2 \cap G_3)$, $P(H_k)$ and $P(J_k)$. By the calculations above, we have

$$\begin{aligned} P(G_1) &= \int_{|z_1|<|a|^{-s}M_a^{1/(2+2q)}} \frac{e^{-\frac{1}{2}(z_1^*,z_1)V_1^{-1}(z_1,z_1^*)^T}}{\pi\sqrt{\det V_1}} dz_1 \\ &= \int_{|z_1|<|a|^{-s}M_a^{1/(2+2q)}} \frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^2}\left(1+O\left(\frac{\epsilon|z_1|^2}{\sigma^2 a^{m(1-s)}}\right)\right) dz_1 \\ &= 1-e^{-|a|^{-2s}M_a^{-q/(1+q)}} + O\left(\frac{\epsilon}{|a|^{m(1-s)}}\right), \end{aligned}$$

and similarly

$$P(G_2) = 1-e^{-M_a^{-q/(1+q)}} + O\left(\frac{\epsilon}{|a|^{m(1-s)}}\right).$$

Hence, we can conclude (ii) and (iii) follows the same proof in the first step. Next, we look at the last two parts of this theorem.

Recall that we have defined a transform $z' = Uz$ and introduced notations $\delta_1 = |a|^{-s}M_a^{1/(2+3q)}$, $\delta_2 = M_a^{1/(2+3q)}$, $\delta_3 = \left(1+|a|^{1-s}\right)M_a^{1/(2+3q)}$, $d_1 = \min\{\frac{\delta_1}{\sqrt{2}}, \frac{\delta_3}{2}\}$, and $d_2 = \min\{\frac{\delta_2}{\sqrt{2}}, \frac{\delta_3}{2}\}$ in the first step. Let

$$g(z) = -\frac{1}{2}(z_1^*, z_2^*, z_3^*, z_1, z_2, z_3)P_\epsilon(z_1, z_2, z_3, z_1^*, z_2^*, z_3^*)^T,$$

and

$$\tilde{g}(z') = g(U^*z').$$

Using the same notations and a similar argument, we have

$$
\begin{aligned}
& P\left(G_1 \cap G_3\right) \\
= & \int_{\{|z_1|<\delta_1,|z_2|^2+|z_3|^2<\delta_3^2\}} \frac{e^{-\frac{1}{2}(z_1^*,z_2^*,z_3^*,z_1,z_2,z_3)V_2^{-1}(z_1,z_2,z_3,z_1^*,z_2^*,z_3^*)^T}}{\pi^3\sqrt{\det V_2}} dz_1dz_2dz_3 \\
\geq & \int_{\{|z_1|<\delta_1,|z_2|<\frac{\delta_3}{\sqrt{2}},|z_3|<\frac{\delta_3}{\sqrt{2}}\}} \frac{e^{-\sigma^{-2}z^*V^{-1}z}e^{g(z)}}{\pi^3\sigma^6\sqrt{(\det V)^2+O\left(\frac{\epsilon}{|a|^{m-2-(m+6)s}}\right)}} dz_1dz_2dz_3 \\
= & \int_{\{\{|z_1|<\delta_1,|z_2|<\frac{\delta_3}{\sqrt{2}},|z_3|<\frac{\delta_3}{\sqrt{2}}\}} \frac{e^{-M_a^{-1}\left(D_{11}|z_1'|^2+D_{22}|z_2'|^2+D_{33}|z_3'|^2\right)}e^{\tilde{g}(z')}}{\pi^3\sigma^6\det V} dz_1'dz_2'dz_3' + O\left(\frac{\epsilon}{|a|^{m-2-(m+2)s)}}\right) \\
\geq & \int_{\{|z_1'|<d_1,|z_2'|<d_1,|z_3'|<d_1\}} \frac{e^{-M_a^{-1}\left(D_{11}|z_1'|^2+D_{22}|z_2'|^2+D_{33}|z_3'|^2\right)}e^{\tilde{g}(z')}}{\pi^3\sigma^6\det V} dz_1'dz_2'dz_3' + O\left(\frac{\epsilon}{|a|^{m-2-(m+2)s)}}\right) \\
= & \int_{\{|z_1'|<d_1,|z_2'|<d_1,|z_3'|<d_1\}} \frac{e^{-M_a^{-1}\left(D_{11}|z_1'|^2+D_{22}|z_2'|^2+D_{33}|z_3'|^2\right)}\left(e^{\tilde{g}(z')}-1\right)}{\pi^3\sigma^6\det V} dz_1'dz_2'dz_3' \\
& + \int_{\{|z_1'|<d_1,|z_2'|<d_1,|z_3'|<d_1\}} \frac{e^{-M_a^{-1}\left(D_{11}|z_1'|^2+D_{22}|z_2'|^2+D_{33}|z_3'|^2\right)}}{\pi^3\sigma^6\det V} dz_1'dz_2'dz_3' + O\left(\frac{\epsilon}{|a|^{m-2-(m+2)s)}}\right) \\
= & \left(1-e^{-\frac{D_{11}d_1^2}{\sigma^2}}\right)\left(1-e^{-\frac{D_{22}d_1^2}{\sigma^2}}\right)\left(1-e^{-\frac{D_{33}d_1^2}{\sigma^2}}\right) + O\left(\frac{\epsilon}{|a|^{(m-4)(1-s)-2}}\right) + O\left(\frac{\epsilon}{|a|^{m-2-(m+2)s)}}\right),
\end{aligned}
$$

and similarly

$$
\begin{aligned}
& P\left(G_2 \cap G_3\right) \\
= & \int_{\{|z_1|<\delta_2,|z_2|^2+|z_3|^2<\delta_3^2\}} \frac{e^{-\frac{1}{2}(z_1^*,z_2^*,z_3^*,z_1,z_2,z_3)V_2^{-1}(z_1,z_2,z_3,z_1^*,z_2^*,z_3^*)^T}}{\pi^3\sqrt{\det V_2}} dz_1dz_2dz_3 \\
\geq & \left(1-e^{-\frac{D_{11}d_2^2}{\sigma^2}}\right)\left(1-e^{-\frac{D_{22}d_2^2}{\sigma^2}}\right)\left(1-e^{-\frac{D_{33}d_2^2}{\sigma^2}}\right) + O\left(\frac{\epsilon}{|a|^{(m-4)(1-s)-2}}\right) + O\left(\frac{\epsilon}{|a|^{m-2-(m+2)s)}}\right).
\end{aligned}
$$

The rest of the proof is exactly the same as the one in the first step and consequently we know this theorem is also true for a mother wave packets of type (ϵ, m) with $m \geq \max\left\{\frac{2(1+s)}{1-s}, \frac{2}{1-s}+4\right\}$. □

A.2 Proofs for the Theorems in Section 3.4

Proof of Theorem 3.4.3

Proof. The proof of this theorem is nearly identical to the proof of Theorem 3.3.1. By the definition of 2D generalized curvelet transform and Lemma 2.3.8 and 2.3.9, we know the following two estimates:

$$
|W_e(a,\theta,b)| \lesssim \sqrt{\epsilon_1}a^{-\frac{s+t}{2}}, \tag{A.9}
$$

and

$$|\nabla_b W_e(a,\theta,b)| \lesssim \sqrt{\epsilon}_1 \left(a^{\frac{t-s}{2}} + a^{1-\frac{s+t}{2}} \right). \tag{A.10}$$

If $(a,\theta,b) \in R_\delta$, then $W_g(a,\theta,b)| \geq a^{-\frac{s+t}{2}}\delta$. Together with Equation (A.1), we have

$$|W_f(a,\theta,b)| \geq a^{-\frac{s+t}{2}}\sqrt{\epsilon}. \tag{A.11}$$

Hence, $S_\delta \subset R_\delta \subset R_\epsilon$, where R_ϵ is defined in Theorem 2.3.7 and is a subset of $\bigcup_{1\leq k\leq K} Z_k$. So, (i) is true by Theorem 2.3.7. As for (ii), since $R_\delta \subset R_\epsilon$, then $(a,\theta,b) \in R_\delta \cap Z_k$ implies $(a,\theta,b) \in R_\epsilon \cap Z_k$ and $a \simeq N_k$. By Theorem 2.3.7, we have

$$\frac{|v_f(a,\theta,b) - N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|} \lesssim \sqrt{\epsilon},$$

when $N > N_0$. Hence,

$$\begin{aligned}
& \frac{|v_g(a,\theta,b) - N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|} \\
\leq\ & \frac{|v_f(a,\theta,b) - N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|} + \frac{|v_f(a,\theta,b) - v_g(a,\theta,b)|}{|N_k\nabla\phi_k(b)|} \\
\lesssim\ & \sqrt{\epsilon} + \frac{|W_e(a,\theta,b)|}{|W_g(a,\theta,b)|} + \frac{|\nabla_b W_e(a,\theta,b)|}{N_k|W_g(a,\theta,b)|} \\
\lesssim\ & \sqrt{\epsilon} + \frac{\sqrt{\epsilon}_1}{\delta} + \frac{\sqrt{\epsilon}_1\left(a^{-(s+t)/2} + a^{1-(s+t)/2}\right)}{\delta N_k a^{-(s+t)/2}} \\
\lesssim\ & \sqrt{\epsilon} + \epsilon_1^p,
\end{aligned}$$

when $N > N_0$. With a similar argument, when $(a,\theta,b) \in S_\delta \cap Z_k$, we can show that

$$\begin{aligned}
& \frac{|v_g(a,\theta,b) - N_k\nabla\phi_k(b)|}{|N_k\nabla\phi_k(b)|} \\
\lesssim\ & \frac{\sqrt{\epsilon}}{N_k^{(s+t)/2}} + \frac{|W_e(a,\theta,b)|}{|W_g(a,\theta,b)|} + \frac{|\nabla_b W_e(a,\theta,b)|}{N_k|W_g(a,\theta,b)|} \\
\lesssim\ & \frac{\sqrt{\epsilon}}{N_k^{(s+t)/2}} + \frac{\sqrt{\epsilon}_1}{a^{(s+t)/2}\delta} + \frac{\sqrt{\epsilon}_1\left(a^{-(s+t)/2} + a^{1-(s+t)/2}\right)}{\delta N_k} \\
\lesssim\ & \frac{\sqrt{\epsilon} + \epsilon_1^p}{N_k^{(s+t)/2}},
\end{aligned}$$

when $N > N_0$. Hence, (iii) is true. □

Proof of Theorem 3.4.4

Proof. The proof of this theorem is similar to the proof of Theorem 3.3.2, but notations are much heavier. Some new notations are introduced to simplify the statement:

1. Let $T_{a\theta j}$ denote the jth element of $R_\theta A_a \xi + a \cdot u_\theta$.
2. Let $n = (n_1, n_2)^T = (a\cos\theta, a\sin\theta)^T$.
3. Let $\sigma^2 = \int_{\mathbb{R}^2} |\widehat{w}(\xi)|^2 \widehat{e}\,(R_\theta A_a \xi + a \cdot u_\theta)\, d\xi$ and $\Xi = (\Xi_1, \Xi_2)^T$ be a random vector with a joint probability density function $\sigma^{-2}|\widehat{w}(\xi)|^2 \widehat{e}\,(R_\theta A_a \xi + a \cdot u_\theta)$.
4. Let
$$g_1(\Xi, n) = a^{t-1}\Xi_1 n_1 - a^{s-1}\Xi_2 n_2,$$
$$g_2(\Xi, n) = a^{t-1}\Xi_1 n_2 - a^{s-1}\Xi_2 n_1,$$
and
$$g(\Xi, n) = \left(g_1\left(\Xi, n\right), g_2\left(\Xi, n\right)\right)^T.$$
5. Let $\tilde{g}_1(\Xi) = a^t \Xi_1$, $\tilde{g}_2(\Xi) = a^s \Xi_2$, and $\tilde{g}(\Xi) = (\tilde{g}_1(\Xi), \tilde{g}_2(\Xi))^T$.

We would also prove the case for a mother curvelet of type $(0, m)$ first. The proof for a mother curvelet is of type (ϵ, m) in the second step is similar, but need to deal with nearly circularly symmetric Gaussian variable. The main difficulty is to estimate the asymptotic behavior of the eigenvalues of the covariance matrices, which will be address in Step 1. The trick to deal with nearly circularly symmetric Gaussian variable is exactly the same as used in Theorem 3.3.2. Hence, we only sketch out Step 2.

Step 1:

Since $w_{a\theta b} \in L^1 \cap C^m$ and $\nabla_b w_{a\theta b} \in L^1 \cap C^{m-1} \times L^1 \cap C^{m-1}$, we know $W_g(a, \theta, b) = W_f(a, \theta, b) + W_e(a, \theta, b)$ and $\nabla_b W_g(a, \theta, b) = \nabla_b W_f(a, \theta, b) + \nabla_b W_e(a, \theta, b)$ are Gaussian random variables. By the properties of zero mean stationary Gaussian processes and the geometric supports of curvelets in the frequency domain, we know $W_e(a, \theta, b)$ and $(W_e(a, \theta, b), \partial_{b_1} W_e(a, \theta, b), \partial_{b_2} W_e(a, \theta, b))$ have zero pseudo-covariance matrices. Furthermore, the variance of $W_e(a, \theta, b)$ is σ^2 and the covariance matrix of
$$(W_e(a, \theta, b), \partial_{b_1} W_e(a, \theta, b), \partial_{b_2} W_e(a, \theta, b))$$
is $\sigma^2 V$, where V is an invertible and self-adjoint matrix given below:

$$\begin{pmatrix} 1 & -2\pi i \mathbb{E}\left[g_1\left(\Xi, n\right) + n_1\right] & -2\pi i \mathbb{E}\left[g_2\left(\Xi, n\right) + n_2\right] \\ 2\pi i \mathbb{E}\left[g_1\left(\Xi, n\right) + n_1\right] & 4\pi^2 \mathbb{E}\left[\left(g_1\left(\Xi, n\right) + n_1\right)^2\right] & 4\pi^2 \mathbb{E}\left[\left(g_1\left(\Xi, n\right) + n_1\right)\left(g_2\left(\Xi, n\right) + n_2\right)\right] \\ 2\pi i \mathbb{E}\left[g_2\left(\Xi, n\right) + n_2\right] & 4\pi^2 \mathbb{E}\left[\left(g_1\left(\Xi, n\right) + n_1\right)\left(g_2\left(\Xi, n\right) + n_2\right)\right] & 4\pi^2 \mathbb{E}\left[\left(g_2\left(\Xi, n\right) + n_2\right)^2\right] \end{pmatrix}.$$

Hence, $W_e(a,\theta,b)$ and $(W_e(a,\theta,b), \partial_{b_1} W_e(a,\theta,b), \partial_{b_2} W_e(a,\theta,b))$ are circularly symmetric and their distributions are described by the probability density functions

$$\frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^{-2}}$$

and

$$\frac{e^{-\sigma^{-2}z^* V^{-1} z}}{\pi^3\sigma^6 \det V},$$

where $z = (z_1, z_2, z_3)^T$. Part (i) is true by previous theorems. To prove Part (ii) to (v), we need to define the following events

$$G_1 = \{|W_e(a,\theta,b)| < a^{-(s+t)/2} M_a^{1/(2+2q)},$$

$$G_2 = \{|W_e(a,\theta,b)| < M_a^{1/(2+2q)},$$

$$G_3 = \{|\nabla_b W_e(a,\theta,b)| < M_a^{1/(2+2q)} \left(a^{(t-s)/2} + a^{1-(s+t)/2}\right)\},$$

$$H_k = \left\{ \frac{|v_g(a,\theta,b) - N_k \nabla_b \phi_k(b)|}{|N_k \nabla_b \phi_k(b)|} \lesssim \sqrt{\epsilon} + M_a^{p/(1+q)} \right\},$$

and

$$J_k = \left\{ \frac{|v_g(a,\theta,b) - N_k \nabla_b \phi_k(b)|}{|N_k \nabla_b \phi_k(b)|} \lesssim N_k^{-(s+t)/2} \left(\sqrt{\epsilon} + M_a^{p/(1+q)}\right) \right\},$$

for $1 \le k \le K$. Next, we are going to estimate the probability $P(G_1)$, $P(G_2)$, $P(G_1 \cap G_3)$, $P(G_2 \cap G_3)$, $P(H_k)$ and $P(J_k)$. By the calculations above, we have

$$P(G_1) = \int_{|z_1| < a^{-(s+t)/2} M_a^{1/(2+2q)}} \frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^{-2}} dz_1 \ge 1 - e^{-a^{-(s+t)} M_a^{-q/(1+q)}},$$

and similarly

$$P(G_2) = \int_{|z_1| < M_a^{1/(2+2q)}} \frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^{-2}} dz_1 \ge 1 - e^{-M_a^{-q/(1+q)}}.$$

We are now ready to conclude (ii) and (iii). If $(a,\theta,b) \in R_{\delta_a}$, then

$$|W_e(a,\theta,b) + W_f(a,\theta,b)| \ge a^{-(s+t)/2} \left(M_a^{(1/2-p)/(1+q)} + \sqrt{\epsilon}\right). \tag{A.12}$$

If $(a,\theta,b) \notin \bigcup_{1\le k\le K} Z_k$, then by Lemma 2.3.9,

$$|W_f(a,\theta,b)| \le a^{-(s+t)/2} \epsilon. \tag{A.13}$$

Equation (A.12) and (A.13) lead to $|W_e(a,\theta,b)| \geq a^{-(s+t)/2} M_a^{1/(2+2q)}$. Hence,

$$P\left((a,\theta,b) \notin \bigcup_{1\leq k\leq K} Z_k\right) \leq P\left(|W_e(a,\theta,b)| \geq a^{-(s+t)/2} M_a^{1/(2+2q)}\right) = 1 - P(G_1).$$

This means that if $(a,\theta,b) \in R_{\delta_a}$, then $(a,\theta,b) \in \bigcup_{1\leq k\leq K} Z_k$ with a probability at least $P(G_1) \geq 1 - e^{-a^{-(s+t)} M_a^{-q/(1+q)}} = 1 - e^{-O\left(N_k^{-(s+t)} M_a^{-q/(1+q)}\right)}$, since $a \simeq N_k$ if $(a,\theta,b) \in Z_k$. So, (ii) is true. A similar argument applied to $(a,\theta,b) \in S_{\delta_a}$ shows that $(a,\theta,b) \in \bigcup_{1\leq k\leq K} Z_k$ with a probability at least $P(G_2) \geq 1 - e^{-M_a^{-q/(1+q)}}$. Hence, (iii) is proved.

Because V is invertible and self-adjoint, there exist a unitary matrix U and a diagonal matrix D such that $V^{-1} = U^* D U$. For $z \in \mathbb{C}^3$, let $z' = Uz$. Introduce notations $\delta_1 = a^{-(s+t)/2} M_a^{1/(2+2q)}$, $\delta_2 = M_a^{1/(2+2q)}$, $\delta_3 = \left(a^{(t-s)/2} + a^{1-(s+t)/2}\right) M_a^{1/(2+2q)}$, $d_1 = \min\{\frac{\delta_1}{\sqrt{2}}, \frac{\delta_3}{2}\}$, and $d_2 = \min\{\frac{\delta_2}{\sqrt{2}}, \frac{\delta_3}{2}\}$. Similar to the proof in Theorem 3.3.2, by a simple property of high dimensional polydisk, we have

$$\begin{aligned} P(G_1 \cap G_3) &= \int_{\{|z_1|<\delta_1, |z_2|^2+|z_3|^2<\delta_3^2\}} \frac{e^{-\sigma^{-2} z^* V^{-1} z}}{\pi^3 \sigma^6 \det V} dz_1 dz_2 dz_3 \\ &\geq \left(1 - e^{-\frac{D_{11} d_1^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22} d_1^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{33} d_1^2}{\sigma^2}}\right), \end{aligned}$$

and similarly

$$\begin{aligned} P(G_2 \cap G_3) &= \int_{\{|z_1|<\delta_2, |z_2|^2+|z_3|^2<\delta_3^2\}} \frac{e^{-\sigma^{-2} z^* V^{-1} z}}{\pi^3 \sigma^6 \det V} dz_1 dz_2 dz_3 \\ &\geq \left(1 - e^{-\frac{D_{11} d_2^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22} d_2^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{33} d_2^2}{\sigma^2}}\right). \end{aligned}$$

Next, we are going to estimate the asymptotic behavior of D_{11}, D_{22} and D_{33} as a increases. This relies on the estimates of $D_{11}D_{22}D_{33}$, $D_{11} + D_{22} + D_{33}$ and $D_{11}^{-1} + D_{22}^{-1} + D_{33}^{-1}$ as follows. Careful

algebraic calculation shows that

$$\begin{aligned}
\frac{\det V}{16\pi^4} &= \mathbb{E}\left[(g_1(\Xi,n)+n_1)^2\right]\mathbb{E}\left[(g_2(\Xi,n)+n_2)^2\right]-\mathbb{E}^2\left[(g_1(\Xi,n)+n_1)(g_2(\Xi,n)+n_2)\right]\\
&\quad +2\mathbb{E}\left[g_1(\Xi,n)+n_1\right]\mathbb{E}\left[g_2(\Xi,n)+n_2\right]\mathbb{E}\left[(g_1(\Xi,n)+n_1)(g_2(\Xi,n)+n_2)\right]\\
&\quad -\mathbb{E}^2\left[g_2(\Xi,n)+n_2\right]\mathbb{E}\left[(g_1(\Xi,n)+n_1)^2\right]-\mathbb{E}^2\left[g_1(\Xi,n)+n_1\right]\mathbb{E}\left[(g_2(\Xi,n)+n_2)^2\right]\\
&= \mathbb{E}\left[(g_1(\Xi,n))^2\right]\mathbb{E}\left[(g_2(\Xi,n))^2\right]-\mathbb{E}^2\left[(g_1(\Xi,n))(g_2(\Xi,n))\right]\\
&\quad +2\mathbb{E}\left[g_1(\Xi,n)\right]\mathbb{E}\left[g_2(\Xi,n)\right]\mathbb{E}\left[(g_1(\Xi,n))(g_2(\Xi,n))\right]\\
&\quad -\mathbb{E}^2\left[g_2(\Xi,n)\right]\mathbb{E}\left[(g_1(\Xi,n))^2\right]-\mathbb{E}^2\left[g_1(\Xi,n)\right]\mathbb{E}\left[(g_2(\Xi,n))^2\right]\\
&= \mathbb{E}\left[(\tilde{g}_1(\Xi))^2\right]\mathbb{E}\left[(\tilde{g}_2(\Xi))^2\right]-\mathbb{E}^2\left[(\tilde{g}_1(\Xi))(\tilde{g}_2(\Xi))\right]\\
&\quad +2\mathbb{E}\left[\tilde{g}_1(\Xi)\right]\mathbb{E}\left[\tilde{g}_2(\Xi)\right]\mathbb{E}\left[(\tilde{g}_1(\Xi))(\tilde{g}_2(\Xi))\right]\\
&\quad -\mathbb{E}^2\left[\tilde{g}_2(\Xi)\right]\mathbb{E}\left[(\tilde{g}_1(\Xi))^2\right]-\mathbb{E}^2\left[\tilde{g}_1(\Xi)\right]\mathbb{E}\left[(\tilde{g}_2(\Xi))^2\right]\\
&= a^{2(s+t)}\big(\mathbb{E}\left[\Xi_1^2\right]\mathbb{E}\left[\Xi_2^2\right]-\mathbb{E}^2\left[\Xi_1\Xi_2\right]+2\mathbb{E}\left[\Xi_1\right]\mathbb{E}\left[\Xi_2\right]\mathbb{E}\left[\Xi_1\Xi_2\right]\\
&\quad -\mathbb{E}^2\left[\Xi_2\right]\mathbb{E}\left[\Xi_1^2\right]-\mathbb{E}^2\left[\Xi_1\right]\mathbb{E}\left[\Xi_2^2\right]\big)\\
&\lesssim a^{2(s+t)}
\end{aligned}$$

Hence,

$$D_{11}D_{22}D_{33}=\det\left(V^{-1}\right)\gtrsim a^{-2(s+t)}. \tag{A.14}$$

Similarly, we know

$$\begin{aligned}
&\operatorname{trace}\left(V^{-1}\right)\\
=\ &\frac{1}{\det V}\Big(16\pi^4\mathbb{E}\left[(g_1(\Xi,n)+n_1)^2\right]\mathbb{E}\left[(g_2(\Xi,n)+n_2)^2\right]\\
&-16\pi^4\mathbb{E}^2\left[(g_1(\Xi,n)+n_1)(g_2(\Xi,n)+n_2)\right]+4\pi^2\mathbb{E}\left[(g_2(\Xi,n)+n_2)^2\right]\\
&-4\pi^2\mathbb{E}^2\left[(g_1(\Xi,n)+n_1)\right]+4\pi^2\mathbb{E}\left[(g_1(\Xi,n)+n_1)^2\right]-4\pi^2\mathbb{E}^2\left[(g_1(\Xi,n)+n_1)\right]\Big)\\
\simeq\ &\frac{1}{\det V}\left(\mathbb{E}\left[(g_1(\Xi,n)+n_1)^2\right]\mathbb{E}\left[(g_2(\Xi,n)+n_2)^2\right]-\mathbb{E}^2\left[(g_1(\Xi,n)+n_1)(g_2(\Xi,n)+n_2)\right]\right)\\
\simeq\ &\frac{a^{2+2s}}{\det V}.
\end{aligned}$$

Therefore,

$$D_{11}+D_{22}+D_{33}=\operatorname{trace}(V^{-1})\simeq\frac{a^{2+2s}}{\det V}. \tag{A.15}$$

Note that

$$\operatorname{trace}(V)=1+4\pi^2\mathbb{E}\left[(g_1(\Xi,n)+n_1)^2\right]+4\pi^2\mathbb{E}\left[(g_2(\Xi,n)+n_2)^2\right]\simeq a^2,$$

then

$$D_{11}^{-1} + D_{22}^{-1} + D_{33}^{-1} \simeq a^2. \tag{A.16}$$

Equation (A.14), (A.15) and (A.16) imply $D_{11} \gtrsim a^{2-2t}$, $D_{22} \simeq a^{-2s}$ and $D_{33} \simeq a^{-2}$. Therefore,

$$\begin{aligned} & P(G_1 \cap G_3) \\ \geq & \left(1 - e^{-\frac{D_{11} d_1^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22} d_1^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{33} d_1^2}{\sigma^2}}\right) \\ = & \left(1 - e^{-O\left(a^{2-s-3t} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(a^{-3s-t} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(a^{-s-t-2} M_a^{-q/(1+q)}\right)}\right). \end{aligned}$$

A similar argument leads to

$$\begin{aligned} & P(G_2 \cap G_3) \\ \geq & \left(1 - e^{-\frac{D_{11} d_2^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{22} d_2^2}{\sigma^2}}\right)\left(1 - e^{-\frac{D_{33} d_2^2}{\sigma^2}}\right) \\ = & \left(1 - e^{-O\left(a^{2-2t} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(a^{-2s} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(a^{-2} M_a^{-q/(1+q)}\right)}\right). \end{aligned}$$

By Theorem 3.4.3, if $(a, \theta, b) \in R_{\delta_a} \cap Z_k$ for some k, then

$$P(H_k) \geq P(H_k | G_1 \cap G_3) P(G_1 \cap G_3) = P(G_1 \cap G_3).$$

Note that $a \simeq N_k$ when $(a, \theta, b) \in Z_k$, then

$$\begin{aligned} P(H_k) & \geq \left(1 - e^{-O\left(a^{2-s-3t} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(a^{-3s-t} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(a^{-s-t-2} M_a^{-q/(1+q)}\right)}\right) \\ & = \left(1 - e^{-O\left(N_k^{2-s-3t} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-3s-t} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-s-t-2} M_a^{-q/(1+q)}\right)}\right). \end{aligned}$$

Similarly, if $(a, \theta, b) \in S_{\delta_a} \cap Z_k$ for some k, then

$$P(J_k) \geq P(G_2 \cap G_3) \geq \left(1 - e^{-O\left(N_k^{2-2t} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-2s} M_a^{-q/(1+q)}\right)}\right)\left(1 - e^{-O\left(N_k^{-2} M_a^{-q/(1+q)}\right)}\right).$$

These arguments prove (iv) and (v).

Step 2: we now prove the case for a mother curvelet of type (ϵ, m) such that

$$m \geq \max\left\{\frac{2(1+s)}{1-t}, \frac{2}{1-t} + 4\right\}.$$

Sufficiently large m keeps our approximation errors small enough.

Now $W_e(a,\theta,b)$ and $(W_e(a,\theta,b),\partial_{b_1}W_e(a,\theta,b),\partial_{b_2}W_e(a,\theta,b))$ have nearly zero Pseudo-covariance matrices and are nearly circularly symmetric. Assume $W_e(a,\theta,b)$ and

$$(W_e(a,\theta,b),\partial_{b_1}W_e(a,\theta,b),\partial_{b_2}W_e(a,\theta,b))$$

have covariance matrices C_1 and C_2, pseudo-covariance matrices P_1 and P_2, respectively. We can still check that they have zero mean and $C_1=\sigma^2$ and $C_2=\sigma^2 V$, where V is defined in the first step. By the definition of the 2D mother wave packet of type (ϵ,m) and the same process in the proof of Theorem 3.2.2 and 3.2.3, we can obtain a almost similar result:

1. The covariance matrix of $(W_e(a,\theta,b),W_e^*(a,\theta,b))$ is

$$V_1=\begin{pmatrix} C_1 & P_1\\ P_1^* & C_1^*\end{pmatrix},$$

and the distribution of $W_e(a,\theta,b)$ is described by

$$\frac{e^{-\frac{1}{2}(z_1^*,z_1)V_1^{-1}(z_1,z_1^*)^T}}{\pi\sqrt{\det V_1}},$$

which is well approximated by

$$\frac{e^{-\frac{C_1|z_1|^2-\Re\epsilon\left(P_1^*z_1^2\right)}{C_1^2-P_1P_1^*}}}{\pi\sqrt{C_1^2-P_1P_1^*}}=\frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^2}\left(1+O\left(\frac{\epsilon|z_1|^2}{\sigma^2a^{m(1-t)}}\right)\right).$$

2. The covariance matrix of $(W_e(a,\theta,b),\partial_bW_e(a,\theta,b),W_e^*(a,\theta,b),\partial_bW_e^*(a,\theta,b))$ is

$$V_2=\begin{pmatrix} C_2 & P_2\\ P_2^* & C_2^*\end{pmatrix},$$

and its distribution is described by the joint probability density

$$\frac{e^{-\frac{1}{2}(z_1^*,z_2^*,z_3^*,z_1,z_2,z_3)V_2^{-1}(z_1,z_2,z_3,z_1^*,z_2^*,z_3^*)^T}}{\pi^3\sqrt{\det V_2}},$$

which is well approximated by

$$\frac{e^{-\sigma^{-2}z^*V^{-1}z}e^{-\frac{1}{2}(z_1^*,z_2^*,z_3^*,z_1,z_2,z_3)P_\epsilon(z_1,z_2,z_3,z_1^*,z_2^*,z_3^*)^T}}{\pi^3\sigma^6\sqrt{(\det V)^2+O\left(\frac{\epsilon}{a^{m-2-(m+2)t-4s}}\right)}},$$

where P_ϵ is a matrix with 2-norm bounded by $O\left(\frac{\epsilon}{\sigma^2a^{(m-4)(1-t)}}\right)$.

Different to Theorem 3.3.2, here we have two scaling parameters t and s with $s < t$. To understand those error bounds above intuitively, we could say that s is shrinking the support of a wave packet in the frequency domain in the angular direction to make it a curvelet and hence to increase the probability of a good estimate, as we have seen that smaller parameters resulting in better robustness. Hence, Most of the error bound above is determined by t, the larger one.

Since the matrix V here has positive eigenvalues bounded above by $O(a^2)$, $O(a^{2s})$ and $O(a^{2(t-1)})$, then C_2 has positive eigenvalues bounded above by $O(\sigma^2 a^2)$, $O(\sigma^2 a^{2s})$ and $O(\sigma^2 a^{2(t-1)})$. Because every entry in P_2 is bounded by $O\left(\frac{\sigma^2\epsilon}{a^{m(1-t)}}\right)$, then determinant error bound comes from

$$O\left(a^2a^2a^{2s}a^{2s}a^{2(t-1)}\frac{\epsilon}{a^{(m-4)(1-t)}}\right) = O\left(\frac{\epsilon}{a^{m-2-(m+2)4-4s}}\right).$$

By the same argument in the first step, we can show that there exist a diagonal matrix $D = \mathrm{diag}\{D_{11}, D_{22}, D_{33}\}$ and a unitary matrix U such that $V^{-1} = U^*DU$. Furthermore, $D_{11} \gtrsim a^{2(1-t)}$, $D_{22} \simeq a^{-2s}$, $D_{33} \simeq a^{-2}$. Part (i) is still true by previous theorems. To conclude Part (ii) to (v), we still need to estimate the probability of those events defined in the first step, i.e., $P(G_1)$, $P(G_2)$, $P(G_1 \cap G_3)$, $P(G_2 \cap G_3)$, $P(H_k)$ and $P(J_k)$. By the calculations above, we have

$$\begin{aligned} P(G_1) &= \int_{|z_1|<a^{-(s+t)/2}M_a^{1/(2+2q)}} \frac{e^{-\frac{1}{2}(z_1^*,z_1)V_1^{-1}(z_1,z_1^*)^T}}{\pi\sqrt{\det V_1}}dz_1 \\ &= \int_{|z_1|<a^{-(s+t)/2}M_a^{1/(2+2q)}} \frac{e^{-\sigma^{-2}|z_1|^2}}{\pi\sigma^2}\left(1+O\left(\frac{\epsilon|z_1|^2}{\sigma^2a^{m(1-t)}}\right)\right)dz_1 \\ &= 1-e^{-a^{-(s+t)}M_a^{-q/(1+q)}} + O\left(\frac{\epsilon}{a^{m(1-t)}}\right), \end{aligned}$$

and similarly

$$P(G_2) = 1-e^{-M_a^{-q/(1+q)}} + O\left(\frac{\epsilon}{a^{m(1-t)}}\right).$$

Hence, we can conclude (ii) and (iii) follows the same proof in the first step. Next, we look at the last two parts of this theorem.

Recall that we have defined a transform $z' = Uz$ and introduced notations $\delta_1 = a^{-(s+t)/2}M_a^{1/(2+2q)}$, $\delta_2 = M_a^{1/(2+2q)}$, $\delta_3 = \left(a^{(t-s)/2} + a^{1-(s+t)/2}\right)M_a^{1/(2+2q)}$, $d_1 = \min\{\frac{\delta_1}{\sqrt{2}}, \frac{\delta_3}{2}\}$, and $d_2 = \min\{\frac{\delta_2}{\sqrt{2}}, \frac{\delta_3}{2}\}$ in the first step. Let

$$g(z) = -\frac{1}{2}(z_1^*, z_2^*, z_3^*, z_1, z_2, z_3)P_\epsilon(z_1, z_2, z_3, z_1^*, z_2^*, z_3^*)^T,$$

and

$$\tilde{g}(z') = g(U^*z').$$

Using the same notations and a similar argument, we have

$$
\begin{aligned}
& P\left(G_1 \cap G_3\right) \\
= & \int_{\{|z_1|<\delta_1,|z_2|^2+|z_3|^2<\delta_3^2\}} \frac{e^{-\frac{1}{2}(z_1^*,z_2^*,z_3^*,z_1,z_2,z_3)V_2^{-1}(z_1,z_2,z_3,z_1^*,z_2^*,z_3^*)^T}}{\pi^3\sqrt{\det V_2}} dz_1dz_2dz_3 \\
\geq & \int_{\{|z_1|<\delta_1,|z_2|<\frac{\delta_3}{\sqrt{2}},|z_3|<\frac{\delta_3}{\sqrt{2}}\}} \frac{e^{-\sigma^{-2}z^*V^{-1}z}e^{g(z)}}{\pi^3\sigma^6\sqrt{(\det V)^2+O\left(\frac{\epsilon}{a^{m-2-(m+2)t-4s}}\right)}} dz_1dz_2dz_3 \\
= & \int_{\{\{|z_1|<\delta_1,|z_2|<\frac{\delta_3}{\sqrt{2}},|z_3|<\frac{\delta_3}{\sqrt{2}}\}} \frac{e^{-M_a^{-1}\left(D_{11}|z_1'|^2+D_{22}|z_2'|^2+D_{33}|z_3'|^2\right)}e^{\tilde{g}(z')}}{\pi^3\sigma^6\det V} dz_1'dz_2'dz_3' + O\left(\frac{\epsilon}{a^{m-2-mt-2s}}\right) \\
\geq & \int_{\{|z_1'|<d_1,|z_2'|<d_1,|z_3'|<d_1\}} \frac{e^{-M_a^{-1}\left(D_{11}|z_1'|^2+D_{22}|z_2'|^2+D_{33}|z_3'|^2\right)}e^{\tilde{g}(z')}}{\pi^3\sigma^6\det V} dz_1'dz_2'dz_3' + O\left(\frac{\epsilon}{a^{m-2-mt-2s}}\right) \\
= & \int_{\{|z_1'|<d_1,|z_2'|<d_1,|z_3'|<d_1\}} \frac{e^{-M_a^{-1}\left(D_{11}|z_1'|^2+D_{22}|z_2'|^2+D_{33}|z_3'|^2\right)}\left(e^{\tilde{g}(z')}-1\right)}{\pi^3\sigma^6\det V} dz_1'dz_2'dz_3' \\
& + \int_{\{|z_1'|<d_1,|z_2'|<d_1,|z_3'|<d_1\}} \frac{e^{-M_a^{-1}\left(D_{11}|z_1'|^2+D_{22}|z_2'|^2+D_{33}|z_3'|^2\right)}}{\pi^3\sigma^6\det V} dz_1'dz_2'dz_3' + O\left(\frac{\epsilon}{a^{m-2-mt-2s}}\right) \\
= & \left(1-e^{-\frac{D_{11}d_1^2}{\sigma^2}}\right)\left(1-e^{-\frac{D_{22}d_1^2}{\sigma^2}}\right)\left(1-e^{-\frac{D_{33}d_1^2}{\sigma^2}}\right) + O\left(\frac{\epsilon}{a^{(m-4)(1-t)-2}}\right) + O\left(\frac{\epsilon}{a^{m-2-mt-2s}}\right),
\end{aligned}
$$

and similarly

$$
\begin{aligned}
& P\left(G_2 \cap G_3\right) \\
= & \int_{\{|z_1|<\delta_2,|z_2|^2+|z_3|^2<\delta_3^2\}} \frac{e^{-\frac{1}{2}(z_1^*,z_2^*,z_3^*,z_1,z_2,z_3)V_2^{-1}(z_1,z_2,z_3,z_1^*,z_2^*,z_3^*)^T}}{\pi^3\sqrt{\det V_2}} dz_1dz_2dz_3 \\
\geq & \left(1-e^{-\frac{D_{11}d_2^2}{\sigma^2}}\right)\left(1-e^{-\frac{D_{22}d_2^2}{\sigma^2}}\right)\left(1-e^{-\frac{D_{33}d_2^2}{\sigma^2}}\right) + O\left(\frac{\epsilon}{a^{(m-4)(1-t)-2}}\right) + O\left(\frac{\epsilon}{a^{m-2-mt-2s}}\right).
\end{aligned}
$$

The rest of the proof is exactly the same as the one in the first step and consequently we know this theorem is also true for a mother wave packets of type (ϵ, m) with m larger than $\max\left\{\frac{2(1+s)}{1-t}, \frac{2}{1-t}+4\right\}$. □

Bibliography

[1] `http://users.ece.utexas.edu/~bevans/projects/rfi/software/`.

[2] F. F. Abraham, R. Walkup, H. Gao, M. Duchaineau, T. Diz De La Rubia, and M. Seager. Simulating materials failure by using up to one billion atoms and the world's fastest computer: Work-hardening. *Proc. Nat. Acad. Sci. USA*, 99:5783–5787, 2002.

[3] C. Anderson and M. D. Dahleh. Rapid computation of the discrete Fourier transform. *SIAM J. Sci. Comput.*, 17(4):913–919, July 1996.

[4] F. Andersson, M. V. De Hoop, and H. Wendt. Multiscale discrete approximation of Fourier integral operators. *Multiscale Model. Simul.*, 10(1):111–145, 2012.

[5] M. Aoi, K. Lepage, Y. Lim, U. Eden, and T. Gardner. An approach to time-frequency analysis with ridges of the continuous chirplet transform. *IEEE Trans. Signal Proc.*, 63(3):699–710, Feb 2015.

[6] F. Auger, E. Chassande-Mottin, and P. Flandrin. Making reassignment adjustable: The levenberg-marquardt approach. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 3889–3892, March 2012.

[7] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans. Signal Proc.*, 43(5):1068 –1089, 1995.

[8] F. Auger, P. Flandrin, Y.-T. Lin, S. McLaughlin, S. Meignen, T. Oberlin, and H.-T. Wu. Time-frequency reassignment and synchrosqueezing: An overview. *IEEE Signal Processing Magazine*, 30(6):32–41, Nov 2013.

[9] J.-F. Aujol and T. F. Chan. Combining geometrical and textured information to perform image classification. *J. Vis. Commun. Image. R.*, 17:1004–1023, 2006.

[10] A. Averbuch, E. Braverman, R. Coifman, M. Israeli, and A. Sidi. Efficient computation of oscillatory integrals via adaptive multiscale local Fourier bases. *Appl. Comput. Harmon. Anal.*, 9(1):19 – 53, 2000.

[11] G. Bao and W. Symes. Computation of pseudo-differential operators. *SIAM J. Sci. Comput.*, 17(2):416–429, 1996.

[12] B. Berkels, A. Rätz, M. Rumpf, and A. Voigt. Extracting grain boundaries and macroscopic deformations from images on atomic scale. *J. Sci. Comput.*, 35:1–23, 2008.

[13] B. Boashash. Estimating and interpreting the instantaneous frequency of a signal. In *Proceedings of the IEEE*, pages 520–538, 1992.

[14] M. Boerdgen, B. Berkels, M. Rumpf, and D. Cremers. Convex relaxation for grain segmentation at atomic scale. In *Vision, Modeling, and Visualization*, pages 179–186. Eurographics Association, 2010.

[15] B. Bradie, R. Coifman, and A. Grossmann. Fast numerical computations of oscillatory integrals related to acoustic scattering, I. *Appl. Comput. Harmon. Anal.*, 1(1):94 – 99, 1993.

[16] P. J. Brockwell and R. A. Davis. *Introduction to time series and forecasting*. Springer, 2nd edition, Mar. 2002.

[17] M. Brown and R. Clapp. (t, x) *domain, patternbased ground roll removal.*

[18] X. Cai, R. Chan, and T. Zeng. A two-stage image segmentation method using a convex variant of the Mumford–Shah model and thresholding. *SIAM J. Imaging Sci.*, 6:368–390, 2013.

[19] E. Candès, L. Demanet, D. Donoho, and L. Ying. Fast discrete curvelet transforms. *Multiscale Model. Simul.*, 5(3):861–899, 2006.

[20] E. J. Candès, P. R. Charlton, and H. Helgason. Detecting highly oscillatory signals by chirplet path pursuit. *Appl. Comput. Harmon. Anal.*, 24(1):14 – 40, 2008.

[21] E. J. Candès, L. Demanet, and L. Ying. Fast computation of Fourier integral operators. *SIAM J. Sci. Comput.*, 29(6):2464–2493, 2007.

[22] E. J. Candès, L. Demanet, and L. Ying. A fast butterfly algorithm for the computation of Fourier integral operators. *Multiscale Model. Simul.*, 7(4):1727–1750, 2009.

[23] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Comm. Pure Appl. Math.*, 57:219–266, 2004.

[24] E. J. Candès and D. L. Donoho. Continuous curvelet transform. I. Resolution of the wavefront set. *Appl. Comput. Harmon. Anal.*, 19(2):162–197, 2005.

[25] E. J. Candès and D. L. Donoho. Continuous curvelet transform. II. Discretization and frames. *Appl. Comput. Harmon. Anal.*, 19(2):198–222, 2005.

[26] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Trans. Image Proc.*, 10:266–277, 2001.

[27] E. Chassande-Mottin, F. Auger, and P. Flandrin. Time-frequency/time-scale reassignment. In *Wavelets and signal processing*, Appl. Numer. Harmon. Anal., pages 233–267. Birkhäuser Boston, Boston, MA, 2003.

[28] E. Chassande-Mottin, I. Daubechies, F. Auger, and P. Flandrin. Differential reassignment. *IEEE Signal Processing Letters*, 4(10):293 –294, 1997.

[29] E. Chassande-Mottin and P. Flandrin. On the time-frequency detection of chirps1. *Appl. Comput. Harmon. Anal.*, 6(2):252 – 281, 1999.

[30] E. Chassande-Mottin and A. Pai. Best chirplet chain: Near-optimal detection of gravitational wave chirps. *Phys. Rev. D*, 73(4), 2006.

[31] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, Dec. 1998.

[32] Y.-C. Chen, M.-Y. Cheng, and H.-T. Wu. Non-parametric and adaptive modelling of dynamic periodicity and trend with heteroscedastic and dependent errors. *J. R. Statist. Soc. B*, 76(3):651–682, 2014.

[33] H. Cheng, Z. Gimbutas, P. Martinsson, and V. Rokhlin. On the compression of low rank matrices. *SIAM J. Sci. Comput.*, 26(4):1389–1404, 2005.

[34] C. K. Chui, Y.-T. Lin, and H.-T. Wu. Real-time dynamics acquisition from irregular samples – with application to anesthesia evaluation. *arXiv:1406.1276 [math.NA]*, 2014.

[35] C. K. Chui and H. Mhaskar. Signal decomposition and analysis via extraction of frequencies. *Appl. Comput. Harmon. Anal.*, (0):–, 2015.

[36] A. Cicone, J. Liu, and H. Zhou. Adaptive local iterative filtering for signal decomposition and instantaneous frequency analysis, 2014. preprint.

[37] M. Clausel, T. Oberlin, and V. Perrier. The monogenic synchrosqueezed wavelet transform: a tool for the decomposition/demodulation of AMFM images . *Appl. Comput. Harmon. Anal.*, (0):–, 2014.

[38] L. Cohen. Generalized phase-space distribution functions. *J. of Math. Phys.*, 7(5):781–786, 1966.

[39] L. Cohen. Time-frequency distributions-a review. *Proceedings of the IEEE*, 77(7):941–981, Jul 1989.

[40] E. Cordero, F. Nicola, and L. Rodino. Sparsity of Gabor representation of Schrödinger propagators. *Appl. Comput. Harmon. Anal.*, 26(3):357 – 370, 2009.

[41] B. Cornelis, A. Dooms, A. Munteanu, J. Cornelis, and P. Schelkens. Experimental study of canvas characterization for paintings. *Proc. SPIE*, 7531:753103, 2010.

[42] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

[43] I. Daubechies, J. Lu, and H.-T. Wu. Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool. *Appl. Comput. Harmon. Anal.*, 30(2):243–261, 2011.

[44] I. Daubechies and S. Maes. A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models. In *Wavelets in Medicine and Biology*, pages 527–546. CRC Press, 1996.

[45] M. de Hoop, G. Uhlmann, A. Vasy, and H. Wendt. Multiscale discrete approximations of Fourier integral operators associated with canonical transformations and caustics. *Multiscale Model. Simul.*, 11(2):566–585, 2013.

[46] A. M. De Livera and R. J. Hyndman. Forecasting time series with complex seasonal patterns using exponential smoothing. Monash Econometrics and Business Statistics Working Papers 15/09, Monash University, Department of Econometrics and Business Statistics, 2009.

[47] E. Delèchelle, J. Lemoine, and O. Niang. Empirical mode decomposition: an analytical approach for sifting process. *IEEE Signal Processing Letters*, pages 764–767, 2005.

[48] N. Delprat, B. Escudié, P. Guillemain, R. Kronland-Martinet, P. Tchamitchian, and B. Torrésani. Asymptotic wavelet and Gabor analysis: Extraction of instantaneous frequencies. *IEEE Trans. Inform. Theory*, 38(2):644–664, 1992.

[49] L. Demanet and L. Ying. Wave atoms and sparsity of oscillatory patterns. *Appl. Comput. Harmon. Anal.*, 23(3):368–387, 2007.

[50] L. Demanet and L. Ying. Scattering in flatland: Efficient representations via wave atoms. *Found. Comput. Math.*, 10(5):569–613, Oct. 2010.

[51] L. Demanet and L. Ying. Discrete symbol calculus. *SIAM Rev.*, 53(1):71–104, 2011.

[52] L. Demanet and L. Ying. Fast wave computation via Fourier integral operators. *Math. Comput.*, 81(279), 2012.

[53] L. Demanet and L. Ying. Fast wave computation via Fourier integral operators. *Mathematics of Computation*, 81:1455–1486, 2012.

[54] E. H. S. Diop, R. Alexandre, and A. Boudraa. Analysis of intrinsic mode functions: A PDE approach. *IEEE Signal Processing Letters*, 17(4):398–401, April 2010.

[55] D. L. Donoho. De-noising by soft-thresholding. *IEEE Trans. Inform. Theory*, 41(3):613–627, may 1995.

[56] D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[57] P. P. D'Ors, C. R. Johnson Jr., and D. H. Johnson. Velazquez in Fraga: a new hypothesis about the portraits of El Primo and Philip IV. *The Burlington Magazine*, CLIV:620–625, 2012.

[58] K. Dragomiretskiy and D. Zosso. Variational mode decomposition. *IEEE Trans. Signal Proc.*, 62(3):531–544, Feb 2014.

[59] K. Dragomiretskiy and D. Zosso. Two-dimensional variational mode decomposition. 8932:197–208, 2015.

[60] A. Dutt and V. Rokhlin. Fast Fourier transforms for nonequispaced data. *SIAM J. Sci. Comput.*, 14(6):1368–1393, 1993.

[61] K. R. Elder and M. Grant. Modeling elastic and plastic deformations in nonequilibrium processing using phase field crystals. *Phys. Rev. E*, 70:051605, 2004.

[62] M. Elsey and B. Wirth. Segmentation of crystal defects via local analysis of crystal distortion. In *Proceedings of the 8th International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 294–302. 2013.

[63] M. Elsey and B. Wirth. A simple and efficient scheme for phase field crystal simulation. *ESAIM: Math. Mod. Num. Anal.*, 47:1413–1432, 2013.

[64] M. Elsey and B. Wirth. Fast automated detection of crystal distortion and crystal defects in polycrystal images. *Multi. Model. Simul.*, 12:1–24, 2014.

[65] M. Elsey and B. Wirth. Redistancing dynamics for vector-valued multilabel segmentation with costly fidelity: grain identification in polycrystal images. *J. Sci. Comp.*, pages 1–28, 2014.

[66] B. Engquist and L. Ying. A fast directional algorithm for high frequency acoustic scattering in two dimensions. *Communications in Mathematical Sciences*, 7(2):327–345, 06 2009.

[67] B. Engquist and L. Ying. A fast directional algorithm for high frequency acoustic scattering in two dimensions. *Commun. Math. Sci.*, 7(2):327–345, 2009.

[68] R. Erdmann, C. R. Johnson Jr., M. Schafer, and J. Twilley. Reuniting Poussin's Bacchanals painted for Cardinal Richelieu through quantitative canvas weave analysis. In *41st Annual Meeting of American Institute for Conservation of Historic and Artistic Works*, Indianapolis, IN, May 2013.

[69] P. Flandrin. *Time-Frequency / Time-Scale Analysis*. Academic Press, London, 1998.

[70] P. Flandrin and P. Goncalvès. Emprical mode decompositions as data-driven wavelet-like expansions. *International Journal of Wavelets, Multiresolution and Information Processing*, 02(04):477–496, 2004.

[71] P. Flandrin, G. Rilling, and P. Goncalves. Empirical mode decomposition as a filter bank. *IEEE Signal Processing Letters*, 11(2):112–114, Feb 2004.

[72] S. Fomel. Appliations of plane-wave destruction filters. *Geophysics*, 2002.

[73] S. Fomel. Seismic data decomposition into spectral components using regularized nonstationary autoregression. *Geophysics*, 78(6):O69–O76, 2013.

[74] S. A. Fulop and K. Fitz. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *J. Acoust. Soc. Am.*, 119(1):360–371, 2006.

[75] D. Gabor. Theory of communication. part 1: The analysis of information. *Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of*, 93(26):429–441, November 1946.

[76] R. G. Gallager. Circularly-symmetric gaussian random vectors. *preprint*, 2008.

[77] J. Gilles. Empirical wavelet transform. *IEEE Trans. Signal Proc.*, 61(16):3999–4010, 2013.

[78] J. Gilles, G. Tran, and S. Osher. 2D empirical transforms. wavelets, ridgelets, and curvelets revisited. *SIAM J. Imaging Sci.*, 7(1):157–186, 2014.

[79] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215 PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[80] M. Gu and S. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM J. Sci. Comput.*, 17(4):848–869, 1996.

[81] P. Guillemain and R. Kronland-Martinet. Characterization of acoustic signals through continuous linear time-frequency representations. *Proceedings of the IEEE*, 84(4):561–585, Apr 1996.

[82] W. Hackbusch and S. Börm. Data-sparse approximation by adaptive $\mathcal{H}^2$-matrices. *Computing*, 69(1):1–35, 2002.

[83] N. Halko, P. Martinsson, and J. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, 2011.

[84] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM rev.*, 53(2):217–288, 2011.

[85] E. Hendriks, L. Jansen, J. Salvant, E. Ravaud, M. Eveno, M. Menu, I. Fielder, M. Geldof, L. Megens, M. van Bommel, C. R. Johnson Jr., and D. H. Johnson. A comparative study of Vincent van Gogh's Bedroom Series. In M. Spring, editor, *Studying Old Master Paintings: Technology and Practice - The National Gallery Technical Bulletin 30th Anniversary Conference Postprints*, pages 237–243. Archetype Publications, 2011.

[86] R. Herrera, J. Han, and M. van der Baan. Applications of the synchrosqueezing transform in seismic time-frequency analysis. *Geophysics*, 79(3):V55–V64, 2014.

[87] F. Hlawatsch and G. Boudreaux-Bartels. Linear and quadratic time-frequency signal representations. *IEEE Signal Processing Magazine*, 9(2):21–67, April 1992.

[88] L. Hörmander. Fourier integral operators. I. *Acta Math.*, 127(1):79–183, 1971.

[89] T. Hou and Z. Shi. Sparse time-frequency decomposition by adaptive basis pursuit. *arXiv:1311.1163 [cs.IT]*, 2013.

[90] T. Hou, Z. Shi, and P. Tavallali. Convergence of a data-driven time-frequency analysis method. *arXiv:1303.7048 [math.NA]*, 2013.

[91] T. Hou, Z. Shi, and P. Tavallali. Sparse time frequency representations and dynamical systems. *arXiv:1312.0202*, 2013.

[92] T. Y. Hou and Z. Shi. Adaptive data analysis via sparse time-frequency representation. *Adv. Adapt. Data Anal.*, 3(1-2):1–28, 2011.

[93] T. Y. Hou and Z. Shi. Data-driven time-frequency analysis. *Appl. Comput. Harmon. Anal.*, 35(2):284 – 308, 2013.

[94] T. Y. Hou, M. P. Yan, and Z. Wu. A variant of the EMD method for multi-scale data. *Adv. Adapt. Data Anal.*, 1(4):483–516, 2009.

[95] J. Hu, S. Fomel, L. Demanet, and L. Ying. A fast butterfly algorithm for generalized Radon transforms. *Geophysics*, 78(4):U41–U51, June 2013.

[96] N. E. Huang. Computer implemented empirical mode decomposition apparatus, method and article of manufacture for two-dimensional signals. *US Patent 6,311,130 B1, Granted Oct. 30, 2001.*

[97] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.*, 454(1971):903–995, 1998.

[98] N. E. Huang, Z. Wu, S. R. Long, K. C. Arnold, X. Chen, and K. Blank. On instantaneous frequency. *Adv. Adapt. Data Anal.*, 1(2):177–229, 2009.

[99] D. Huybrechs and S. Vandewalle. A two-dimensional wavelet-packet transform for matrix compression of integral equations with highly oscillatory kernel. *J. Comput. Appl. Math.*, 197(1):218 – 232, 2006.

[100] D. Johnson, C. Johnson Jr., A. Klein, W. Sethares, H. Lee, and E. Hendriks. A thread counting algorithm for art forensics. In *Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. DSP/SPE 2009. IEEE 13th*, pages 679–684, Jan 2009.

[101] D. H. Johnson, R. G. Erdmann, and C. R. Johnson Jr. Whole-painting canvas analysis using high- and low-level features. In *Proc. 36th Int. Conf. on Acoustics, Speech and Signal Processing*, pages 969–972, 2011.

[102] D. H. Johnson, E. Hendriks, M. Geldof, and C. R. Johnson Jr. Do weave matches imply canvas roll matches? In *38th Annual Meeting of American Institute for Conservation of Historic and Artistic Works*, Milwaukee, WI, May 2010.

[103] D. H. Johnson, E. Hendriks, and C. R. Johnson Jr. Interpreting canvas weave matches. *Art Matters*, 5:53–61, 2013.

[104] D. H. Johnson, C. R. Johnson Jr., and R. G. Erdmann. Weave analysis of paintings on canvas from radiographs. *Signal Process.*, 93:527–540, 2013.

[105] C. R. Johnson Jr., E. Hendriks, P. Noble, and M. Franken. Advances in computer-assisted canvas examination: Thread counting algorithms. In *37th Annual Meeting of American Institute for Conservation of Historic and Artistic Works*, Los Angeles, CA, May 2009.

[106] C. R. Johnson Jr., D. H. Johnson, N. Hamashima, H. S. Yang, and E. Hendriks. On the utility of spectral-maximum-based automated thread counting from X-rays of paintings on canvas. *Studies in Conservation*, 56:104–114, 2011.

[107] C. R. Johnson Jr., D. H. Johnson, I. Verslype, R. Lugtigheid, and R. G. Erdmann. Detecting weft snakes. *Art Matters*, 5:48–52, 2013.

[108] M. Jung, G. Peyré, and L. Cohen. Nonlocal active contours. *SIAM J. Imaging Sci.*, 5:1022–1054, 2012.

[109] W. E. King, G. H. Campbell, S. M. Foiles, D. Cohen, and K. M. Hanson. Quantitative HREM observation of the $\sum 11(113)/[\bar{1}10]$ grain-boundary structure in aluminium and comparison with atomistic simulation. *J. Microsc.*, pages 131–143, 1998.

[110] A. Kirsh and R. Levenson. *Seeing through Paintings: Physical Examination in Art Historical Studies*. Yale University Press, 2000.

[111] C. Kittel. *Introduction to Solid State Physics (7th Ed.)*. Wiley, 1995.

[112] K. Kodera, R. Gendrin, and C. Villedary. Analysis of time-varying signals with small bt values. *IEEE Trans. Acoustics, Speech and Signal Proc.*, 26(1):64–76, Feb 1978.

[113] K. Kodera, C. Villedary, and R. Gendrin. Analysis of time-varying signals with small bt values. *Phys. Earth Planet. Interiors*, (12):142–150, 1976.

[114] H.-H. Kuo. *White noise distribution theory*. Probability and stochastics series. CRC Press, Boca Raton (USA), 1996.

[115] C. Li and M. Liang. A generalized synchrosqueezing transform for enhancing signal timefrequency representation. *Signal Processing*, 92(9):2264 – 2274, 2012.

[116] Y. Li and L. Demanet. Phase and amplitude tracking for seismic event separation. *Submitted*.

[117] Y. Li, H. Yang, E. Martin, K. Ho, and L. Ying. Butterfly factorization. *Multiscale Model. Simul.*, 2015.

[118] Y. Li, H. Yang, and L. Ying. Multi-dimensional butterfly factorization. *Preprint*, 2015.

[119] Y. Li, H. Yang, and L. Ying. A multiscale butterfly aglorithm for Fourier integral operatos. *Multiscale Model. Simul.*, 2015.

[120] E. Liberty, F. Woolfe, P.-G. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proc. Natl. Acad. Sci. USA*, 104(51):20167–20172, 2007.

[121] W. Liedtke, C. R. Johnson Jr., and D. H. Johnson. Canvas matches in Vermeer: A case study in the computer analysis of fabric supports. *Metropolitan Museum Journal*, 47:99–106, 2012.

[122] L. Lin, J. Lu, and L. Ying. Fast construction of hierarchical matrix representation from matrix-vector multiplication. *J. Comput. Phys.*, 230(10):4071–4087, 2011.

[123] A. Linderhed. Variable sampling of the empirical mode decomposition of twodimensional signals. *Int. J. Wavelets Multresolution Inform. Process*, 2005.

[124] A. Linderhed. Image empirical mode decomposition: A new tool for image processing. *Adv. Adapt. Data Anal*, 2009.

[125] K. Lister, C. Peres, and I. Fiedler. Appendix: tracing an interaction: supporting evidence, experimental grounds. In D. Druick and P. Zegers, editors, *Van Gogh and Gauguin: The Studio of the South*, pages 354–369. Thames & Hudson, 2001.

[126] J. Lu, B. Wirth, and H. Yang. Combining 2D synchrosqueezed wave packet transform with optimization for crystal image analysis. *arXiv:1501.06254v1*, Submitted on 26 Jan 2015.

[127] M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. USA*, 106(3):697–702, 2009.

[128] S. Mallat. *A wavelet tour of signal processing.* Elsevier/Academic Press, Amsterdam, third edition, 2009.

[129] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Proc.*, 41(12):3397–3415, Dec 1993.

[130] P. G. Martinsson. A fast randomized algorithm for computing a hierarchically semiseparable representation of a matrix. *SIAM J. Matrix Anal. Appl.*, 32(4):1251–1274, 2011.

[131] Y. Meyer. *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations: The Fifteenth Dean Jacqueline B. Lewis Memorial Lectures.* American Mathematical Society, Boston, MA, USA, 2001.

[132] E. Michielssen and A. Boag. A multilevel matrix decomposition algorithm for analyzing scattering from large structures. *IEEE Trans. Antennas Propagat.*, 44(8):1086–1093, Aug 1996.

[133] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42:577–685, 1989.

[134] F. Neeser and J. Massey. Proper complex random processes with applications to information theory. *IEEE Trans. Inform. Theory*, 39(4):1293–1302, Jul 1993.

[135] J. C. Nunes, Y. Bouaoune, E. Delechelle, O. Niang, and P. Bunel. Image analysis by bidimensional empirical mode decomposition. *Image Vision Comput.*, 2003.

[136] J. C. Nunes, O. Niang, Y. Bouaoune, E. Delechelle, and P. Bunel. Bidimensional empirical mode decomposition modified for texture analysis. *Image Anal. Proc.*, 2003.

[137] T. Oberlin, S. Meignen, and V. Perrier. Second-order synchrosqueezing transform or invertible reassignment? towards ideal time-frequency representations. *Signal Processing*, 63(5):1335 – 1344, 2015.

[138] M. O'Neil, F. Woolfe, and V. Rokhlin. An algorithm for the rapid evaluation of special function transforms. *Appl. Comput. Harmon. Anal.*, 28(2):203–226, 2010.

[139] B. Picinbono. On instantaneous amplitude and phase of signals. *IEEE Trans. Signal Proc.*, pages 552–560, 1997.

[140] D. Potts, G. Steidl, and M. Tasche. Fast Fourier transforms for nonequispaced data: A tutorial, 2001.

[141] J. Poulson, N. Demanet, L. Maxwell, and L. Ying. A parallel butterfly algorithm. *SIAM J. Sci. Comput.*, to appear.

[142] J. Qian and L. Ying. Fast multiscale Gaussian wavepacket transforms and multiscale Gaussian beams for the wave equation. *Multiscale Model. Simul.*, 8(5):1803–1837, 2010.

[143] S. Richwalski, K. Roy-Chowdhury, and J. C. Mondt. Multi-component wavefield separation applied to high-resolution surface seismic data. *J. Appl. Geophys.*, 46(2):101–114, 2001.

[144] RKD (Netherlands Institute for Art History). Image sharing.

[145] B. Sandberg, T. F. Chan, and L. A. Vese. A level-set and Gabor-based active contour algorithm for segmenting textured images. Technical report, UCLA Department of Mathematics CAM report, 2002.

[146] D. S. Seljebotn. Wavemoth-fast spherical harmonic transforms by butterfly matrix compression. *The Astrophysical Journal Supplement Series*, 199(1):5, 2012.

[147] H. M. Singer and I. Singer. Analysis and visualization of multiply oriented lattice structures by a two-dimensional continuous wavelet transform. *Phys. Rev. E*, 74:031103, 2006.

[148] J.-L. Starck, E. J. Candès, and D. L. Donoho. The curvelet transform for image denoising. *IEEE Trans. Image Proc.*, 11(6):670–684, 2002.

[149] D. Steel. Catologue entry: 1. Giotto and his workshop, 'Christ Blessing with Saint John the Evangelist, the Virgin Mary, Saint John the Baptist, and Saint Francis (Peruzzi Altarpiece)'. In C. Sciacca, editor, *Florence at the Dawn of the Renaissance, Painting and Illumination: 1300-1350*, pages 24–28. The J. Paul Getty Museum, 2012.

[150] E. Strekalovskiy and D. Cremers. Total variation for cyclic structures: Convex relaxation and efficient minimization. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1905–1911, 2011.

[151] A. Stukowski and K. Albe. Dislocation detection algorithm for atomistic simulations. *Model. Simul. Mater. Sci. Eng.*, 18:025016, 2010.

[152] A. Stukowski and K. Albe. Extracting dislocations and non-dislocation crystal defects from atomistic simulation data. *Model. Simul. Mater. Sci. Eng.*, 18:085001, 2010.

[153] J. B. Tary, R. H. Herrera, J. Han, and M. van der Baan. Spectral estimation–what is new? what is next? *Rev. Geophys.*, pages n/a–n/a, 2014.

[154] M. Taylor. Random fields: stationarity, ergodicity, and spectral behavior, `http://www.unc.edu/math/Faculty/met/rndfcn.pdf`.

[155] G. Thakur, E. Brevdo, N. S. Fučkar, and H.-T. Wu. The synchrosqueezing algorithm for time-varying spectral analysis: robustness properties and new paleoclimate applications. *Signal Processing*, 93(5):1079–1094, 2013.

[156] G. Thakur and H.-T. Wu. Synchrosqueezing-based recovery of instantaneous frequency from nonuniform samples. *SIAM J. Math. Analysis*, 43(5):2078–2095, 2011.

[157] D. O. Trad, T. J. Ulrych, and M. D. Sacchi. Accurate interpolation with high-resolution time-variant radon transforms. *Geophysics*, 67(2):644–656, 2002.

[158] M. Tygert. Fast algorithms for spherical harmonic expansions, III. *Journal of Computational Physics*, 229(18):6181 – 6192, 2010.

[159] M. Unser. Texture classification and segmentation using wavelet frames. *IEEE Trans. Image Proc.*, 4:1549–1560, 1995.

[160] E. van de Wetering. *Rembrandt: The Painter at Work*. Amsterdam University Press, Amsterdam, 1997.

[161] A. Van Den Bos. The multivariate complex normal distribution – a generalization. *IEEE Trans. Inform. Theory*, 41(2):537–539, Mar 1995.

[162] M. van der Baan. `http://www.ualberta.ca/~vanderba/ftp/benchmark_signals.zip`.

[163] M. van der Baan. PP/PS wavefield separation by independent component analysis. *Geophys. J. Int.*, (166):339–348, 2006.

[164] L. van Tilborgh, T. Meedendorp, E. Hendriks, D. H. Johnson, C. R. Johnson Jr., and R. G. Erdmann. Weave matching and dating of van Gogh's paintings: An interdisciplinary approach. *The Burlington Magazine*, CLIV:112–122, 2012.

[165] L. A. Vese and T. F. Chan. A multiphase level set framework for image segmentation using the Mumford and Shah model. *Int. J. Comput. Vision*, 50:271–293, 2002.

[166] L. A. Vese and S. J. Osher. Modeling textures with total variation minimization and oscillating patterns in image processing. *J. Sci. Comput.*, 19:553–572, 2003.

[167] J. Ville. Theorie et Applications de la Notion de Signal Analytique. *Cables et Transmission*, 1:61–74, 1948.

[168] E. Wigner. On the quantum correction for thermodynamic equilibrium. *Phys. Rev.*, 40:749–759, Jun 1932.

[169] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Appl. Comput. Harmon. Anal.*, 25(3):335 – 366, 2008.

[170] H.-T. Wu. Instantaneous frequency and wave shape functions (I). *Appl. Comput. Harmon. Anal.*, 35(2):181 – 199, 2013.

[171] H.-T. Wu, Y.-H. Chan, Y.-T. Lin, and Y.-H. Yeh. Using synchrosqueezing transform to discover breathing dynamics from ECG signals. *Appl. Comput. Harmon. Anal.*, 36(2):354 – 359, 2014.

[172] H.-T. Wu, P. Flandrin, and I. Daubechies. One or two frequencies? The synchrosqueezing answers. *Adv. Adapt. Data Anal.*, 3(1-2):29–39, 2011.

[173] H.-T. Wu, S.-S. Hseu, M.-Y. Bien, Y. R. Kou, and I. Daubechies. Evaluating physiological dynamics via synchrosqueezing: Prediction of ventilator weaning. *IEEE Trans. Biomedical Engineering*, 61(3):736–744, March 2014.

[174] Z. Wu and N. E. Huang. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv. Adapt. Data Anal.*, 01(01):1, 2009.

[175] Z. Wu, N. E. Huang, and X. Chen. The multi-dimensional ensemble empirical mode decomposition method. *Adv. Adapt. Data Anal.*, 1(3):339–372, 2009.

[176] Z. Wu, N. E. Huang, and X. Chen. Some considerations on physical analysis of data. *Adv. Adapt. Data Anal.*, 3(1-2):95–113, 2011.

[177] J. Xia, S. Chandrasekaran, M. Gu, and X. S. Li. Fast algorithms for hierarchically semiseparable matrices. *Numerical Linear Algebra with Applications*, 17(6):953–976, 2010.

[178] H. Yang. Synchrosqueezed wave packet transforms and diffeomorphism based spectral analysis for 1D general mode decompositions. *Appl. Comput. Harmon. Anal.*, 39(1):33 – 66, 2015.

[179] H. Yang, J. Lu, W. Brown, I. Daubechies, and L. Ying. Quantitative canvas weave analysis using 2D synchrosqueezed transforms. *IEEE Signal Processing Magazine, Special Issue on Art Investigations*, 2015.

[180] H. Yang, J. Lu, and L. Ying. Crystal image analysis using 2D synchrosqueezed transforms. *arXiv:1402.1262 [math.NA]*, Submitted on 6 Feb 2014.

[181] H. Yang and L. Ying. A fast algorithm for multilinear operators. *Appl. Comput. Harmon. Anal.*, 33(1):148 – 158, 2012.

[182] H. Yang and L. Ying. Synchrosqueezed wave packet transform for 2D mode decomposition. *SIAM J. Imaging Sci.*, 6(4):1979–2009, 2013.

[183] H. Yang and L. Ying. Robustness analysis of synchrosqueezed transforms, 2014. preprint.

[184] H. Yang and L. Ying. Synchrosqueezed curvelet transform for two-dimensional mode decomposition. *SIAM J. Math. Anal*, 46(3):2052–2083, 2014.

[185] C. Yarham, U. Boeniger, and F. Herrmann. *Curvelet based ground roll removal*.

[186] B. Yazici, L. Wang, and K. Duman. Synthetic aperture inversion with sparsity constraints. In *Electromagnetics in Advanced Applications (ICEAA), 2011 International Conference on*, pages 1404–1407, Sept 2011.

[187] R. Yin, D. Dunson, B. Cornelis, W. Brown, N. Ocon, and I. Daubechies. Digital cradle removal in x-ray images of art paintings. In *IEEE International Conference on Image Processing*, Paris, Oct 2014.

[188] L. Ying. Sparse Fourier transform via butterfly algorithm. *SIAM J. Sci. Comput.*, 31(3):1678–1694, Feb. 2009.

[189] X. Zeng, B. Gipson, Z. Y. Zheng, L. Renault, and H. Stahlberg. Automatic lattice determination for two-dimensional crystal images. *J. Struct. Biol.*, 160:353–361, 2007.