ABSTRACT

Title of dissertation: NONLINEAR SAMPLING THEORY AND EFFICIENT SIGNAL RECOVERY

Kung-Ching Lin
Doctor of Philosophy, 2020

Dissertation directed by: Professor John Benedetto
Department of Mathematics

Sampling theory investigates signal recovery from its partial information, and one of the simplest and most well-known sampling schemes is uniform linear sampling, characterized by the celebrated classical sampling theorem. However, the requirements of uniform linear sampling may not always be satisfied, sparking the need for more general sampling theories.

In the thesis, we discuss the following three sampling scenarios: signal quantization, compressive sensing, and deep neural networks.

In signal quantization theory, the inability of digital devices to perfectly store analog samples leads to distortion when reconstructing the signal from its samples. Different quantization schemes are proposed so as to minimize such distortion. We adapt a quantization scheme used in analog-to-digital conversion called signal decimation to finite dimensional signals. In doing so, we are able to achieve theoretically optimal reconstruction error decay rate.

Compressive sensing investigates the possibility to recover high-dimensional signals from incomplete samples. It has been proven feasible as long as the signal

is sufficiently sparse. To this point, all of the most successful examples follow from random constructions rather than deterministic ones. Whereas the sparsity of the signal can be almost as large as the ambient dimension for random constructions, current deterministic constructions require the sparsity to be at most the square-root of the ambient dimension. This apparent barrier is the well-known *square-root bottleneck*. In this thesis, we propose a new explicit sampling scheme as a possible candidate for deterministic compressive sensing. We present a partial result, while the full generality is still work in progress.

For deep neural networks, one approximates signals with neural networks. To do so, many samples need to be drawn in order to find an optimal approximating neural network. A common approach is to employ stochastic gradient descent, but it is unclear if the resulting neural network is indeed optimal due to the non-convexity of the optimization scheme. We follow an alternative approach, utilizing the derivatives of the signal for stable reconstruction. In this thesis, we focus on non-smooth signals, and using weak differentiation, it is easy to obtain stable reconstruction for one-layer neural networks. We are currently working on the two-layer case, and our approach is outlined in this thesis.

NONLINEAR SAMPLING THEORY AND EFFICIENT SIGNAL RECOVERY

by

Kung-Ching Lin

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor John Benedetto, Chair/Advisor
Professor Radu Balan
Professor Wojciech Czaja
Professor Kasso Okoudjou
Professor Thomas Goldstein

Acknowledgments

I would like to thank my advisor, Prof. John Benedetto, for providing the constant help and guiding throughout my years in University of Maryland. His insights have led me to numerous interesting problems, all of which became parts of this thesis. He is also more than an advisor on research. He took me in when I was still new to Maryland, and he was a stabilizing force during my volatile first year.

I would also like to acknowledge the unwavering support of my family. My parents never questioned my decisions along the way, and my wife has accompanied me through all the hardship and struggling these years. It is comforting to know they have always been there for me.

Last be not least, I would like to acknowledge the support of ARO grant W911NF-17-1-0014 and NSF-DMS Grant 1814253. They helped smooth out the process of my research.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Overview

Harmonic analysis has always been one of the most applicable branches in mathematics, having strong and active interactions with the engineering communities. With the advancements of information technology, and more recently the uprise of machine learning and deep learning, this field is gaining even more attention, attracting the best mathematicians in the world, such as Daubechies, Mallat, Tao, Yau, Bourgain, and many more.

Harmonic analysis can be found in many places, both applied and pure. For direct applications, signal processing and data analysis is present in virtually all scientific fields, as long as there exist signals or data of any kind in the field. For instance, the introduction of compressive sensing has greatly reduced the cost and time needed for MRI scanning, while the research in machine learning makes many automated processes possible, from facial recognition to self-driving cars. On the other hand, harmonic analysis also has a profound impact in the development of mathematics itself. Fourier analysis, an integral part of harmonic analysis, has always been influential in partial differential equations (PDEs) and number theory, and in turn it also evolves along with every branch of mathematics to even greater

generalities. In short, harmonic analysis is both applied and abstract, and the feedback from both sides only accelerates its evolution.

Sampling theory in signal processing investigates the problem to retrieve full information of the objects in interest from only a fraction of it. Classical examples includes X-rays, MRIs, and Analog-to-digital (A/D) conversions. Originally, the sampling theory focused on the linear uniform sampling on $\mathbb{R}$, where the classical sampling theorem states that given a bandlimited function $f : \mathbb{R} \to \mathbb{R}$, i.e., an L^2 function whose Fourier transform is compactly supported, it is possible to reconstruct the function from its uniform discrete time samples $\{f(nT)\}_{n\in\mathbb{Z}}$, where $T > 0$ is sufficiently small. However, the limitation of available resources and physical constraints call for extension of the theory to non-uniform (balayage, short-time Fourier transform, etc.) and non-linear sampling (signal quantization, compressive sensing, etc.). This thesis contributes to nonlinear sampling theory in the form of signal quantization and compressive sensing.

Another subject in applied harmonic analysis is machine learning, and in particular deep learning. Deep neural networks in deep learning have yielded excellent results in numerous tasks, but the theories to back up its success are relatively lacking. Current efforts from mathematicians on deep learning can be classified into two aspects: the approximation power of deep neural networks and the optimization problems on the parameters. As deep neural networks generally have millions of parameters, identification of the true parameters can be troublesome. A popular choice is to use stochastic gradient descent (SGD) to minimize certain loss functions. However, the optimization is often non-convex, so there is no guarantee for recovery

of the global minimum. Moreover, the optimization is often slow due to the high dimensionality.

An alternative is proposed to replace SGD, which utilizes the derivative of the neural network to recover the parameters via tensor decomposition. Posed as a signal recovery problem from its samples, I studied the recovery of ReLU neural networks with two layers.

In terms of connections, signal quantization and compressive sensing, and to some extent, weight identification for deep learning, belong in nonlinear sampling theory. Compressive sensing focuses on designing sampling schemes for data acquisition, while signal quantization and weight identification are concerned more on the post-processing and signal recovery. In Section 1.2, we shall give a brief outline on the problems solved in the thesis.

1.2 Description of the Problems and Contribution

1.2.1 Signal Quantization

Signal quantization theory, which leverages between resources spent to record signal samples and the reconstruction accuracy from the quantized samples, originally comes from analog-to-digital (A/D) conversion. As mentioned above, a bandlimited function can be recovered from its discrete time samples, but due to the discrete nature of digital storing devices, one is only able to store quantized samples, whose values come from a finite alphabet. It is natural to take the direct round-offs of the samples and store them as quantized samples. However, this approach

has been shown [26] to be vulnerable to hardware imperfection. A more robust scheme called $\Sigma\Delta$ quantization was introduced [41] in 1963 and became popular in the engineering community, with the caveat of its slow linear reconstruction error decay rate. Higher order $\Sigma\Delta$ quantization yields polynomial error decay rate, and subsequent developments gave birth to schemes with exponential error decay rate.

In Chapter 2, we introduce two novel quantization schemes for signal quantization on finite frames called *alternative decimation* and *adapted decimation*, respectively. Signal decimation was introduced and studied [16] as post-processing on $\Sigma\Delta$ quantization for A/D conversion. It was hypothesized in that paper that signal decimation coupled with $\Sigma\Delta$ quantization yields exponential error decay rate with respect to the bit usage, and it was proven [25] later that it is indeed the case. We adapted signal decimation to finite frames, first in the form of alternative decimation. It has a obvious connection to signal decimation, but it is only applicable up to second order $\Sigma\Delta$ quantization. Further generalization requires factorization and re-arrangement of the decimation operator which is adapted decimation.

One close competitor of decimation is distributed noise shaping, or more specifically, the beta dual [17, 18]. However, we argue that the beta dual requires multiplication of some carefully tailored $\beta > 1$ on the quantized samples, while adaptive decimation only involves summation, which makes adaptive decimation more viable for simple hardware.

1.2.2 Deterministic Compressive Sensing

Compressive sensing investigates the feasibility to recover sparse signals from highly incomplete measurements. Tao et al. [12, 13, 14, 15] and Donoho [29] introduced the concept of restricted isometry property (RIP) to guarantee the recovery of sparse signals. At the time, random matrices were immediately proven to satisfy RIP for rather dense signals where the support of the signals can almost be as large as the ambient space. This fact has led to improvements to medical imaging such as MRI.

On the other hand, deterministic construction of such matrices saw much less success. It has been known that given a matrix, low coherence of its columns (the largest absolute value among the inner products of columns) guarantees the RIP for sparse signals. However, as the Welch bound gives a lower bound of coherence, one can only produce matrices with RIP for signals with sparsity the square-root of the ambient dimension. Such an obstacle is aptly named *the square-root bottleneck*. Numerous efforts utilized number theory for the construction, but for a long time the bottleneck seemed unbreakable. To this date, the bottleneck was only barely eclipsed once by Bourgain et al.[9] in 2011. Using techniques in additive combinatorics, they were able to construct RIP matrices with sparsity up to $n^{1/2+\epsilon}$, where n is the ambient dimension, and $\epsilon \sim 10^{-16}$.

Motivated by the conjectures in the properties of Legendre symbols, we investigated in matrices constructed by the Gabor system of such symbols. Following the work by Bourgain et al., the key estimate was on the sum of inner products between

columns. In particular, the problem of deterministic compressive sensing can be reformulated as the uniform estimate over multidimensional character sums. Interestingly, the traditional Weil estimate or Polya-Vinogradov inequality both yield a bottleneck estimate, so an even finer estimate tailored to our construction is needed. In Chapter 3, we shall prove that for subsets consisting of consecutive indices [48], the estimate is better than the bottleneck bound. We are continuing our effort with the aim to tackle the full problem.

1.2.3 Deep Neural Networks

Even though the concept was already known in the 20th century, deep learning only became feasible in recent years thanks to significant leaps in computing power. Deep learning has been shown to be extremely useful in many tasks, but theoretical results on it are still relatively lacking. With its promise in many fields, mathematical aspects of deep learning have become an extremely popular topic among mathematicians.

Besides the studies on the approximation power of deep neural networks, it is also important to be able to determine the optimal parameters efficiently. As deep neural networks generally have millions of parameters, the identification of the true parameters can be troublesome. A popular choice is to use stochastic gradient descent (SGD) to minimize certain loss functions. However, the optimization is often non-convex, so there is no guarantee for recovery of the global minimum. Moreover, the optimization is often slow due to the high dimensionality.

Fornasier et al.[32, 33, 34] proposed an alternative to SGD, using derivatives and tensor decomposition to recover the true parameters reliably for up to two hidden layers. However, their results only work for smooth activation functions, therefore excluding the rectified linear unit (ReLU) which is a popular activation function. In Chapter 4, we relax their assumption to include ReLU activation functions. As ReLU functions are not smooth, distributional derivative is necessary.

Our result for the two layer case creates a different outlook than Fornasier's result. In their case, there would be *entangled weights* that vary for different sampling points. As for ours, the number of entangled weights grows exponentially with the dimension of the first layer. Our current problem is also related to the work of Yau et al.[45] where one constructs a convex polytope with prescribed face volumes and normal directions. In particular, the distribution of face volumes of the convex polytope will play an important role in our algorithm. Shen et al.[53, 54] examined the approximation power of ReLU neural networks, and their method provides a hint on reducing the input dimension to simplify the problem further. In view of the works above, I have proposed a technique called *net-spreading*, performing weak differentiation locally for ReLU neural networks. Leveraging the properties of ReLU, a multi-scale version of the original algorithm used by Fornasier et al. can be employed. Favorable properties of ReLU also allows for reliable function recovery, which remains open in the work of Fornasier et al..

1.3 Results

1.3.1 Signal Quantization

We formulate and prove Theorem 2.3.5 and Theorem 2.3.8, which is an extension of Theorem 2.1.4 to finite frames. In particular, using our notion of alternative decimation, which will be defined in Section 2.3, we shall prove exponential error decay with respect to the total number of bits used.

In Section 2.3, we define alternative decimation and state our main results. Theorem 2.3.4 is a special case of Theorem 2.3.5, where we restrict ourselves to *finite harmonic frames*, a subclass of unitarily generated frames. The same result for unitarily generated frames satisfying certain mild conditions is proven in Theorem 2.3.5, and it is further extended to the second order in Theorem 2.3.8. The multiplicative structure of decimation is proven in Theorem 2.3.7, and this enables us to perform decimation iteratively.

We prove Theorems 2.3.4, 2.3.5, 2.3.7, and 2.3.8 in Sections 2.3.1, 2.3.2, 2.3.3, and 2.3.4, respectively. Generalization to orders greater than two is done by means of adaptive decimation, defined in Section 2.4 and Theorem 2.4.3. Its proof is given in Section 2.4.6.

1.3.2 Compressive Sensing

Explained in [9], an equivalent property to RIP is the cancellation between sums of inner products over arbitrary subsets of pre-determined size. As the first

step to breaking the square-root bottleneck, we prove the cancellation when the summands consist of consecutive indices, which is given in Corollary 3.3.2, derived easily from Theorem 3.3.1. The proof of Theorem 3.3.1 is given in Section 3.4.

We also include the motivation and possible guideline of estimates over arbitrary summands in Section 3.8. In particular, we proposed power methods and exact counting using properties of Legendre symbols. The problem is still open, and the work for general cases is ongoing.

1.3.3 Deep Learning

In Section 4.2, we show the feasibility to use the same method as [33] for shallow ReLU neural networks. We first compute the weak differentiation of such networks in Section 4.2.2. As it yields similar forms as networks with smooth activation functions, it is possible to follow the same methods almost verbatim, as is described in the following sections.

In Section 4.3, we show that for two-layer networks, the weak derivative of ReLU neural networks is different from the derivative of those with smooth activation functions. This leads to an incompatibility between the methods in [34] and our setting. We describe our new algorithm in Section 4.3.2 and provide partial

theoretical guarantees.

Chapter 2: Signal Decimation

In this chapter, we propose two quantization schemes for finite frames that satisfy theoretically optimal error decay rates. First, we give a brief exposition on the recent development of signal quantization theorey in Section 2.1. Then, we formulate and prove the properties of alternative decimation and adapted decimation in Sections 2.3 and 2.4 respectively.

2.1 Preliminaries

2.1.1 Classical Sampling Theorem and Analog-to-Digital Conversion

Analog-to-digital (A/D) conversion is a process where bandlimited signals, e.g., audio signals, are digitized for storage and transmission, which is feasible thanks to the classical sampling theorem. In particular, the theorem indicates that discrete sampling is sufficient to capture all features of a given bandlimited signal, provided that the sampling rate is higher than the Nyquist rate.

Given a function $f \in L^1(\mathbb{R})$, its Fourier transform $\hat{f}$ is defined as

$$\hat{f}(\gamma) = \int_{-\infty}^{\infty} f(t) e^{-2\pi \imath t \gamma} \, dt.$$

The Fourier transform can also be uniquely extended to $L^2(\mathbb{R})$ as a unitary transformation.

Definition 2.1.1. Given $f \in L^2(\mathbb{R})$, $f \in PW_\Omega$ if its Fourier transform $\hat{f} \in L^2(\mathbb{R})$ is supported in $[-\Omega, \Omega]$.

An important component of A/D conversion is the following theorem:

Theorem 2.1.2 (Classical Sampling Theorem)**.** *Given $f \in PW_{1/2}$, for any $g \in L^2(\mathbb{R})$ satisfying*

- $\hat{g}(\omega) = 1$ *on* $[-1/2, 1/2]$
- $\hat{g}(\omega) = 0$ *for* $|\omega| \geq 1/2 + \epsilon$,

with $\epsilon > 0$ and $T \in (0, 1 - 2\epsilon)$, $t \in \mathbb{R}$, one has

$$f(t) = T \sum_{n \in \mathbb{Z}} f(nT) g(t - nT),$$

where the convergence is both uniform on compact sets of $\mathbb{R}$ and in $L^2(\mathbb{R})$.

As an extreme case, for $g(t) = \sin(\pi t)/(\pi t)$ and $T = 1$, the following identity holds in $L^2(\mathbb{R})$:

$$f(t) = \sum_{n \in \mathbb{Z}} f(n) \frac{\sin(\pi(t - n))}{\pi(t - n)}.$$

2.1.2 Signal Quantization

However, the discrete nature of digital data storage makes it impossible to store exactly the samples $\{f(nT)\}_{n \in \mathbb{Z}}$. Instead, the quantized samples $\{q_n\}_{n \in \mathbb{Z}}$ chosen

from a pre-determined finite alphabet $\mathscr{A}$ are stored. This results in the following reconstructed signal

$$\tilde{f}(t) = T \sum q_n g(t - nT).$$

As for the choice of the quantized samples $\{q_n\}_n$, we shall discuss the following two schemes

- Pulse Code Modulation (PCM):

  Quantized samples are taken as the direct-roundoff of the current sample, i.e.,

$$q_n = Q_0(f(nT)) := \arg\min_{q \in \mathscr{A}} |q - f(nT)|. \tag{2.1}$$

- $\Sigma\Delta$ Quantization:

  A sequence of auxiliary variables $\{u_n\}_{n\in\mathbb{Z}}$ is introduced for this scheme. $\{q_n\}_{n\in\mathbb{Z}}$ is defined recursively as

$$q_n = Q_0(u_{n-1} + f(nT)),$$
$$u_n = u_{n-1} + f(nT) - q_n.$$

$\Sigma\Delta$ quantization was introduced [41] in 1963, and it is still widely used due to some of its advantages over PCM. Specifically, $\Sigma\Delta$ quantization is robust against hardware imperfection [26], a decisive weakness for PCM. For $\Sigma\Delta$ quantization, and the more general noise shaping schemes to be explained below, the boundedness of $\{u_n\}_{n\in\mathbb{Z}}$ turns out to be essential. Quantization schemes with $\|u\|_\infty < \infty$ are said to be *stable*.

Despite its merits over PCM, $\Sigma\Delta$ quantization merely yields linear error decay with respect to the bit-rate as opposed to exponential error decay by its counterpart PCM. Thus, it is desirable to generalize $\Sigma\Delta$ quantization for better error decay rates.

As a direct generalization, given $r \in \mathbb{N}$, one can consider an *r-th order $\Sigma\Delta$ quantization scheme*:

Theorem 2.1.3 (Higher Order $\Sigma\Delta$ Quantization, [24]). *Given $f \in PW_{1/2}$ and $T < 1$, consider the following stable quantization scheme*

$$f(nT) - q_n = (\Delta^r u) := \sum_{l=0}^{r} (-1)^l \binom{r}{l} u_{n-l},$$

where $\{q_n\}$ and $\{u_n\}$ are the quantized samples and auxiliary variables, respectively. Then, for all $t \in \mathbb{R}$,

$$|f(t) - T \sum_{n \in \mathbb{Z}} q_n g(t - nT)| \leq T^r \|u\|_\infty \left\| \frac{d^r g}{dt^r} \right\|_1.$$

2.1.3 Signal Decimation

Higher order $\Sigma\Delta$ quantization has been known for a long time [20, 31], and the r-th order $\Sigma\Delta$ quantization improves the error decay rate from linear to polynomial degree r while preserving the advantages of a first order $\Sigma\Delta$ quantization scheme.

From here, a natural question arises: is it possible to generalize $\Sigma\Delta$ quantization further so that the reconstruction error decay matches the exponential decay of PCM? Two solutions have been proposed for this question. The first one is to adopt

different quantization schemes. Many of the proposed schemes, including higher order $\Sigma\Delta$ quantization, can be categorized as noise shaping quantization schemes, and a brief summary of such schemes will be provided in Section 2.1.6.

The other possibility is to enhance data storage efficiency while maintaining the same level of reconstruction accuracy, and *signal decimation* belongs in this category. Signal decimation is implemented as follows: given an r-th order $\Sigma\Delta$ quantization scheme, there exists $\{q_n^T\}, \{u_n\}$ such that

$$f_n^{(T)} - q_n^T = f(nT) - q_n^T = (\Delta^r u)_n, \tag{2.2}$$

where $\|u\|_\infty < \infty$, and $\{f_n^{(T)}\}_n = \{f(nT)\}_n$. Then, consider

$$\tilde{q}_n^{T_0} := (S_\rho^r q^T)_{(2\rho+1)n},$$

a sub-sampled sequence of $S_\rho^r q^T$, where $(S_\rho h)_n := \frac{1}{2\rho+1}\sum_{m=-\rho}^{\rho} h_{n+m}$. Signal decimation is the process with which one converts the quantized samples $\{q_n^T\}$ to $\{\tilde{q}_n^{T_0}\}$. See Figure 2.1 for an illustration.

Decimation has been known in the engineering community [16], and it was observed that decimation results in exponential error decay with respect to the bit-rate, even though the observation remained a conjecture until 2015 [25], when Daubechies and Saab proved the following theorem:

Theorem 2.1.4 (Signal Decimation for Bandlimited Functions, [25]). *Given $f \in PW_{1/2}$, $T < 1$, and $T_0 = (2\rho+1)T < 1$, there exists a function $\tilde{g}$ such that*

$$f(t) = T_0 \sum [S_\rho^r f^{(T)}]_{(2\rho+1)n} \tilde{g}(t - nT_0),$$

$$|f(t) - T_0 \sum \tilde{q}_n^{T_0} \tilde{g}(t - nT_0)| \leq C\|u\|_\infty \left(\frac{T}{T_0}\right)^r =: \mathcal{D}, \tag{2.3}$$

where $\{f_n^{(T)}\}_{n\in\mathbb{Z}}$ *is defined in* (2.2)*, and* C *is a constant such that* $T_0^r \tilde{g}_1^{(r)} \leq C$*. Moreover, the number of bits needed for each unit interval is*

$$\frac{1}{T_0}\log_2((2\rho+1)^r + 1) \leq \frac{1}{T_0}\log_2\left(2\left(\frac{T_0}{T}\right)^r\right) =: \mathcal{R}. \tag{2.4}$$

Consequently,

$$\mathcal{D}(\mathcal{R}) = 2\|u\|_\infty C 2^{-T_0\mathcal{R}}.$$

From (2.3) and (2.4), we can see that the reconstruction error after decimation still decays polynomially with respect to the sampling rate. As for the data storage, the number of bits needed changes from $O(T^{-1})$ to $O(\log(1/T))$. Thus, the reconstruction error decays exponentially with respect to the bits used.

2.1.4 Unitarily Generated Frames

A unitarily generated frame T_u is generated by a cyclic group: given a unit base vector $\phi_0 \in \mathbb{C}^k$ and a Hermitian matrix $\Omega \in \mathbb{C}^{k\times k}$, the frame elements of T_u are defined as

$$\phi_j^{(m)} = U_{j/m}\phi_0, \quad U_t := e^{2\pi\imath\Omega t}.$$

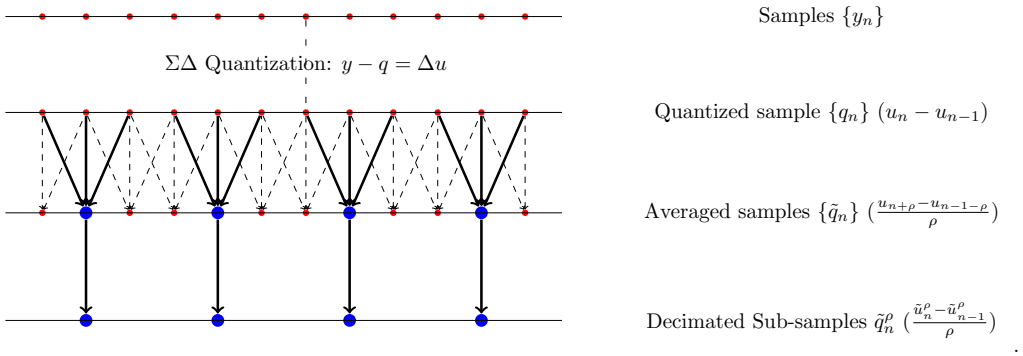


Figure 2.1: Illustration of the first order decimation scheme. After obtaining the quantized samples $\{q_n\}_n$ in the first step, decimation takes the average of quantized samples within disjoint blocks in the second step. The outputs are used as the decimated sub-samples $\{\tilde{q}^\rho_n\}$ in the third step. The effect on the reconstruction (replacing q_n with $y_n - q_n$) is illustrated in parentheses.

The analysis operator Φ of T_u has $\{\phi_j^*\}_j$ as its rows.

As symmetry occurs naturally in many applications, it is not surprising that unitarily generated frames receive serious attention, and their applications in signal processing abound, [17, 18, 30, 35].

One particular application comes from dynamical sampling, which records the spatiotemporal samples of a signal in interest. Mathematically speaking, one tries to recover a signal f on a domain D from the samples $\{f(X), f_{t_1}(X), \ldots, f_{t_N}(X)\}$ where $X \subset D$, and $f_{t_j} = A^{t_j} f$ denotes the evolved signal. Equivalently, one recovers f from $\{<A^{t_j} f, e_i>\}_{i,j} = \{<f, (A^{t_j})^* e_i>\}_{i,j}$, which aligns with the frame reconstruction problems, [1, 2]. In particular, Lu and Vetterli [49, 50] investigated the reconstruction from spatiotemporal samples for a diffusion process. They noted that one can compensate under-sampled spatial information with sufficiently over-sampled temporal data. Unitarily generated frames represent the cases when the evolution process is unitary and the spatial information is one-dimensional.

It should be noted that unitarily generated frames are group frames with the generator $G = U_{1/m}$ provided that $U_1 = G^m = I_k$, while harmonic frames are tight unitarily generated frames. Here, a frame $T = \{e_j\}_j \subset \mathcal{H}$ is tight if for all $v \in \mathcal{H}$, there exists a constant $A > 0$ such that $\sum_j |<v, e_j>|^2 = A\|v\|^2$.

A special class of harmonic frames that we shall discuss is the exponential frame with generator Ω as a diagonal matrix with integer entries and the base vector $\phi_0 = (1, \ldots, 1)^t/\sqrt{k}$.

2.1.5 $\Sigma\Delta$ Quantization on Finite Frames

Signal quantization theory on finite frames is well motivated from the need to deal with data corruption or erasure [37, 38]. The authors considered the PCM quantization scheme described above and modeled the quantization error as random noise. In [6], deterministic analysis on $\Sigma\Delta$ quantization for finite frames showed that a linear error decay rate is obtained with respect to the oversampling ratio. Moreover, if the frame satisfies certain smoothness conditions, the decay rate can be super-linear for first order $\Sigma\Delta$ quantization. Noise shaping schemes for finite frames have also been investigated, some of which yield exponential error decay rate [17, 18, 19].

Fix a separable Hilbert space $\mathcal{H}$ along with a set of vectors $T = \{e_j\}_{j\in\mathbb{Z}} \subset \mathcal{H}$. The collection of vectors T forms a frame for $\mathcal{H}$ if there exist $A, B > 0$ such that for

any $v \in \mathcal{H}$, the following inequality holds:

$$A\|v\|_{\mathcal{H}}^2 \leq \sum_{j\in\mathbb{Z}} |<v, e_j>|^2 \leq B\|v\|_{\mathcal{H}}^2.$$

The concept of frames is a generalization of orthonormal bases in a vector space. Different from bases, frames are usually over-complete: the vectors form a linearly dependent spanning set. Over-completeness of frames is particularly useful for noise reduction, and consequently frames are more robust against data corruption than orthonormal bases.

Let us restrict ourselves to the case when $\mathcal{H} = \mathbb{C}^k$ is a finite dimensional Euclidean space, and the frame consists of a finite number of vectors. Given a finite frame $T = \{e_j\}_{j=1}^m$, the linear operator $E : \mathbb{C}^k \to \mathbb{C}^m$ satisfying $Ev = \{<v, e_j>\}_{j=1}^m$ is called the *analysis operator*. Its adjoint operator $E^* : \mathbb{C}^m \to \mathbb{C}^k$ satisfies $E^*c = \sum_{j=1}^m c_j e_j$ and is called the *synthesis operator*. The *frame operator* $\mathcal{S}$ is defined by $\mathcal{S} = E^*E : \mathbb{C}^k \to \mathbb{C}^k$.

Remark 2.1.5. Note that since $\mathcal{S}$ is Hermitian,

$$\|\mathcal{S}\|_2 = \max_{v:\|v\|_2=1} |v^T \mathcal{S} v| = \max_{v:\|v\|_2=1} \sum_{j=1}^m |<v, e_j>|^2 \leq B.$$

Similarly, $\|\mathcal{S}^{-1}\|_2 \leq A^{-1}$. In particular, the 2-norm of $\mathcal{S}$ is directly tied to the lower frame bound of T.

Under this framework, one considers the quantized samples q of $y = Ex$ and reconstructs $\tilde{x} = \mathcal{S}^{-1}E^*q$, where $\mathcal{S} = E^*E$. The frame-theoretic greedy $\Sigma\Delta$ quanti-

zation is defined as follows: given a finite alphabet $\mathscr{A} \subset \mathbb{C}$, consider the auxiliary variable $\{u_n\}_{n=0}^m$, where we shall set $u_0 = 0$. For $n = 1, \ldots, m$, we calculate $\{q_n\}_n$ and $\{u_n\}_n$ as follows:

$$\begin{aligned} q_n &= Q_0(u_{n-1} + y_n) \\ u_n &= u_{n-1} + y_n - q_n, \end{aligned} \tag{2.5}$$

where Q_0 is defined in (2.1). In the matrix form, we have

$$y - q = \Delta u, \tag{2.6}$$

where $\Delta \in \mathbb{Z}^{m \times m}$ is the backward difference matrix, i.e., $\Delta_{i,i} = 1$ for all $1 \leq i \leq m$, and $\Delta_{i,i-1} = -1$ for $2 \leq i \leq m$. For an r-th order $\Sigma\Delta$ quantization, we have instead $y - q = \Delta^r u$.

In practice, the quantization alphabet $\mathscr{A}$ is often chosen to be $\mathscr{A}_0$ which is uniformly spaced and symmetric around the origin: given $\delta > 0$, we define a mid-rise uniform quantizer $\mathscr{A}_0$ of length $2L$ to be $\mathscr{A}_0 = \{(2j+1)\delta/2 : -L \leq j \leq L-1\}$.

For complex Euclidean spaces, we define $\mathscr{A} = \mathscr{A}_0 + \imath\mathscr{A}_0$. In both cases, $\mathscr{A}$ is called a mid-rise uniform quantizer. Throughout this paper we shall always be using $\mathscr{A}$ as our quantization alphabet.

2.1.6 Noise Shaping Schemes and the Choice of Dual Frames

$\Sigma\Delta$ quantization is a subclass of the more general noise shaping quantization, where the quantization scheme is designed such that the reconstruction error is

easily separated from the true signal in the frequency domain. For instance, it is pointed out in [18] that the reconstruction error of $\Sigma\Delta$ quantization for bandlimited functions is concentrated in high frequency ranges. Since audio signals have finite bandwidth, it is then possible to separate the signal from the error using low-pass filters.

Noise shaping quantization has been well established for A/D conversion since the mid 20th century [55], and in terms of finite frames, noise shaping schemes generalize the $\Sigma\Delta$ scheme in the following way:

$$y - q = Hu,$$

where y, q, and u are the samples, quantized samples, and the auxiliary variable, respectively, while the transfer matrix H is lower-triangular. Now, given an analysis operator E, a transfer matrix H, and a dual F to E, i.e. , $FE = I_k$, the reconstruction error in this setting is

$$\|x - Fq\|_2 = \|F(Ex - q)\|_2 = \|FHu\|_2 \leq \|FH\|_{\infty,2}\|u\|_\infty,$$

where $\|\cdot\|_{\infty,2}$ is the operator norm between ℓ^∞ and ℓ^2, i.e.,

$$\|T\|_{\infty,2} := \sup_{\|x\|_\infty = 1} \|Tx\|_2.$$

The choice of the dual frame F plays a role in the reconstruction error. For instance, [8] proved that $\arg\min_{FE=I_k} \|FH\|_2 = (H^{-1}E)^\dagger H^{-1}$, where given any

matrix A, $A^\dagger$ is defined as the canonical dual $(A^*A)^{-1}A^*$. More generally, one can consider a V-dual, namely $(VE)^\dagger V$, provided that VE is still a frame. With this terminology, decimation can be viewed as a special case of V-duals, and conversely every V-dual can be associated with corresponding post-processing on the quantized sample q.

2.1.7 Perspective and Prior Works

- Quantization for Bandlimited Functions:

  Despite its simple form and robustness, $\Sigma\Delta$ quantization only results in linear error decay with respect to the sampling period T as $T \to 0$. It was shown [20, 24, 31] that a generalization of $\Sigma\Delta$ quantization, namely the r-th order $\Sigma\Delta$ quantization, has error decay rate of polynomial order r. Leveraging the different constants for this family of quantization schemes, sub-exponential decay can also be achieved. A different family of quantization schemes was proven [39] to yield exponential error decay with a small exponent ($c \approx 0.07$.) In [27], the exponent was improved to $c \approx 0.102$.

- Finite Frames:

  $\Sigma\Delta$ quantization can also be applied to finite frames. It was proven [6] that for any family of finite frames with bounded frame variation, the reconstruction error decays linearly with respect to the oversampling ratio m/k, where the corresponding analysis operator E is an $m \times k$ matrix. With different choices of dual frames, [8] proved that the so-called Sobolev dual achieves minimum

induced matrix 2-norm for reconstructions. By carefully matching between the dual frame and the quantization scheme, [18] proved that using the β-dual for random frames results in exponential error decay of near-optimal exponent with high probability.

- Decimation:

  In [16], using the assumption that the noise in $\Sigma\Delta$ quantization is random along with numerical experiments, it was asserted that decimation greatly reduces the number of bits needed while maintaining the reconstruction accuracy. In [25], a rigorous proof was given to show that such an assertion is indeed valid, and the reduction of bits used turns the linear decay into exponential decay with respect to the bit-rate.

  Adapting decimation to finite frames is by no means a new idea. Iwen and Saab [42] used probabilistic arguments and the property of efficient storage to construct random quantization schemes with exponential error decay rate with respect to the bit usage. In [40], similar ideas are used on $\Sigma\Delta$. Moreover, the connection between decimation and distributed noise shaping can be seen in it.

  [40, 42] both use probabilistic arguments that only ensure success with some probability instead of deterministic guarantee. Different from their work, we shall propose two deterministic quantization schemes in Chapter 2.

- Beta Dual of Distributed Noise Shaping:

Chou and Günturk [17, 18] proposed a distributed noise shaping quantization scheme with beta duals as an example. The definition of a beta dual is as follows:

Definition 2.1.6 (Beta Dual)**.** Let $E \in \mathbb{R}^{m\times k}$ be an analysis operator and $k \mid m$. Recall that $F_V \in \mathbb{R}^{k\times m}$ is a V-dual of E if

$$F_V = (VE)^\dagger V,$$

where $V \in \mathbb{R}^{p\times m}$ such that VE is still a frame.

Given $\beta > 1$, the β-dual $F_V = (VE)^\dagger V$ has $V = V_{\beta,m}$, a k-by-m block matrix such that each block is $v = [\beta^{-1}, \beta^{-2}, \ldots, \beta^{-m/k}] \in \mathbb{R}^{1\times m/k}$.

In this case, the transfer matrix H is an m-by-m block matrix where each block h is an m/k-by-m/k matrix with unit diagonal entries and $-\beta$ as sub-diagonal entries. Under this setting, it is proven that the reconstruction error decays exponentially.

One may notice the similarity between the beta dual and decimation. Indeed, if one chooses $\beta = 1$ and normalizes V by $\frac{k}{m}$, the same result as decimation can be obtained, achieving linear error decay with respect to the oversampling ratio and exponential decay with respect to the bit usage. Nonetheless, its generalization to higher order error decay with respect to the oversampling ratio is lacking, whereas the alternative decimation we propose can be extended to the second order. In particular, the raw performance of the second order

decimation is superior to the *1-dual* under the same oversampling ratio.

2.2 Notation

The following notation is used in this chapter:

- $x \in \mathbb{C}^k$: the signal of interest.
- $E \in \mathbb{C}^{m\times k}$: a fixed frame.
- $y = Ex \in \mathbb{C}^m$: the sample.
- $\rho \in \mathbb{N}$: the block size of the decimation.
- $\eta = \lfloor m/\rho \rfloor \in \mathbb{N}$: the greatest integer smaller than the ratio m/ρ.
- $\mathscr{A} = \mathscr{A}_0 + \imath\mathscr{A}_0 \subset \mathbb{C}$: the quantization alphabet. $\mathscr{A}$ is said to have length $2L$ with gap δ if $\mathscr{A}_0 = \{(2j+1)\delta/2 : -L \le j \le L-1\}$ for some $\delta > 0$.
- $q \in \mathbb{C}^m$: the quantized sample obtained from the greedy $\Sigma\Delta$ quantization defined in (2.5).
- $u \in \mathbb{C}^m$: the auxiliary variable of $\Sigma\Delta$ quantization.
- $F \in \mathbb{C}^{k\times m}$: a dual to the analysis operator E, i.e. $FE = I_k$.
- $\mathscr{E}$: the reconstruction error $\mathscr{E} = \|x - Fq\|_2$.
- $\mathscr{R}$: total number of bits used to record the quantized sample.
- $\Omega \in \mathbb{C}^{k\times k}$: a Hermitian matrix with eigenvalues $\{\lambda_j\}_{j=1}^k \subset \mathbb{R}$ and corresponding orthonormal eigenvectors $\{v_j\}_{j=1}^k$.

- $\Phi \in \mathbb{C}^{m\times k}$: the analysis operator of the unitarily generated frame (UGF) with the generator Ω and the base vector $\phi_0 \in \mathbb{C}^k$.

- $U_t \in \mathbb{C}^{k\times k}$: the unitary matrix defined as $U_t = e^{2\pi\imath\Omega t}$ for any $t \in \mathbb{R}$.

- $B = B_\Phi \in \mathbb{C}^{k\times k}$: a unitary matrix that simultaneously diagonalizes U_t and Ω. In particular, $\Omega = B\Lambda B^*$ and $U_t = Be^{2\pi\imath\Lambda t}B^*$, where $\Lambda = diag(\lambda_1, \ldots, \lambda_k)$.

- $\|\cdot\|_{p,q}$: the p-to-q norm. For any matrix M, $\|M\|_{p,q} := \sup_{v:\|v\|_p=1} \|Mv\|_q$. For simplicity, we denote $\|\cdot\|_2 := \|\cdot\|_{2\to 2}$ for matrices.

- $\delta : \mathbb{Z} \to \{0, 1\}$: the Kronecker delta. $\delta(k) = 1$ if $k = 0$, and 0 otherwise. With some abuse of notation, we may also view δ as a function on the cyclic group $\mathbb{Z}/\ell\mathbb{Z}$ for any $\ell \in \mathbb{N}$.

2.3 Alternative Decimation

For the rest of the chapter, we shall also assume that our $\Sigma\Delta$ quantization scheme is stable, i.e. , $\|u\|_\infty$ remains bounded as the dimension $m \to \infty$. Recall the definition of a unitarily generated frame in 2.1.4.

It will be shown that, for unitarily generated frames Φ satisfying conditions specified in Theorem 2.3.5, $\Sigma\Delta$ quantization coupled with alternative decimation still has linear reconstruction error decay rate with respect to the oversampling ratio ρ. As for the data storage, decimation allows for highly efficient storage, and the error decays exponentially with respect to the number of bits used.

Definition 2.3.1 (Alternative Decimation)**.** Given fixed $m, \rho \in \mathbb{N}$, the (m, ρ)-alternative decimation operator is defined to be $D_\rho S_\rho$, where

- $S_\rho = S_\rho^+ - S_\rho^- \in \mathbb{R}^{m \times m}$ is the integration operator satisfying

$$\begin{aligned}(S_\rho^+)_{l,j} &= \begin{cases} \frac{1}{\rho} & \text{if} \quad l \geq \rho,\, l-(\rho-1) \leq j \leq l \\ 0 & \text{otherwise,} \end{cases} \\ (S_\rho^-)_{l,j} &= \begin{cases} \frac{1}{\rho} & \text{if} \quad l \leq \rho-1,\, l+1 \leq j \leq m-\rho+l \\ 0 & \text{otherwise.} \end{cases}\end{aligned} \tag{2.7}$$

  Here, the cyclic convention is adopted: for any $s \in \mathbb{Z}$, $s \equiv s+m$.

- $D_\rho \in \mathbb{N}^{\eta \times m}$ is the sub-sampling operator satisfying

$$(D_\rho)_{l,j} = \begin{cases} 1 & \text{if} \quad j = \rho \cdot l \\ 0 & \text{otherwise,} \end{cases}$$

  where $\eta = \lfloor m/\rho \rfloor$.

Remark 2.3.2 (Canonical Decimation $D_\rho \tilde{S}_\rho$ and Alternative Decimation $D_\rho S_\rho$)**.** It is tempting to consider a closely related circulant matrix $\tilde{S}_\rho$ that satisfies $S_\rho = \tilde{S}_\rho - L$, where L is constant on the first $(\rho-1)$ rows and zero otherwise. Visually, $\tilde{S}_\rho$ and

S_ρ has the following form

$$\tilde{S}_\rho = \frac{1}{\rho}\begin{pmatrix} 1 & 0 & \dots & \dots & 0 & 1 & \dots & 1 \\ \vdots & \ddots & \ddots & & & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 & \dots & \dots & 0 & 1 \\ 1 & \dots & \dots & 1 & & & & \\ & \ddots & & & \ddots & & & \\ & & \ddots & & & \ddots & & \\ & & & \ddots & & & \ddots & \\ & & & & 1 & \dots & \dots & 1 \end{pmatrix}, S_\rho = \frac{1}{\rho}\begin{pmatrix} 0 & -1 & \dots & \dots & -1 & 0 & & \\ & & \ddots & & & \ddots & \ddots & \\ & & & -1 & \dots & \dots & -1 & 0 \\ 1 & \dots & \dots & 1 & & & & \\ & \ddots & & & \ddots & & & \\ & & \ddots & & & \ddots & & \\ & & & \ddots & & & \ddots & \\ & & & & 1 & \dots & \dots & 1 \end{pmatrix}. \tag{2.8}$$

Indeed, $D_\rho \tilde{S}_\rho = D_\rho S_\rho$, so there is no difference between the alternative decimation and canonical decimation. However, we will show in Appendix 2.3.6.1 that $D_\rho \tilde{S}_\rho^2 \neq D_\rho S_\rho^2$, and it is necessary to consider $D_\rho S_\rho^2$ instead of $D_\rho \tilde{S}_\rho^2$ for the second order decimation.

Definition 2.3.3 (Frame variation)**.** Given $A = (A_1, \dots, A_p) \in \mathbb{C}^{s \times p}$, the frame variation $\sigma(A)$ is defined to be

$$\sigma(A) = \sum_{t=1}^{p-1} \|A_t - A_{t+1}\|_2.$$

Theorem 2.3.4 (Special Case: Decimation for Harmonic Frames)**.** *Fix the analysis operator* $E = E^{m,k} \in \mathbb{C}^{m \times k}$ *with entries* $E_{l,j} = \frac{1}{\sqrt{k}} exp(-2\pi\imath(n_j l)/m)$*. Suppose* $\{n_l\}_{l=1}^k$ *are distinct integers in* $[-k/2, k/2]$*, then the following statements are true:*

(a) ***Signal reconstruction:*** *The matrix $D_\rho S_\rho E \in \mathbb{C}^{\eta\times k}$ has rank k.*

(b) ***Error estimate****: The dual $F = (D_\rho S_\rho E)^\dagger D_\rho S_\rho$ to E has reconstruction error*

$$\|x - Fq\|_2 \le \frac{\pi}{2}(\sigma(\bar{F}) + \|\bar{F}_\eta\|_2)\|u\|_\infty \frac{1}{\rho},$$

where $\bar{F} = (\bar{F}_1, \dots, \bar{F}_\eta)$ is the canonical dual of the matrix $(\frac{1}{\sqrt{k}} e^{-2\pi\imath\rho l n_j/m})_{l,j} \in \mathbb{C}^{\eta\times k}$.

Moreover, if $\rho \mid m$, then the reconstruction error $\mathscr{E}$ satisfies

$$\mathscr{E} := \|x - Fq\|_2 \le \begin{cases} \frac{\pi^2(k+1)}{\sqrt{3}}\|u\|_\infty \frac{k}{m} & \text{if } m, k \text{ are even and } n_j\text{'s are nonzero,} \\ \frac{\pi}{2}(\frac{2\pi(k+1)}{\sqrt{3}} + 1)\|u\|_\infty \frac{k}{m} & \text{otherwise.} \end{cases}$$

In particular, the error decays linearly with respect to the oversampling ratio m/k.

(c) ***Efficient data storage****: Suppose the length of the quantization alphabet $\mathscr{A}$ is $2L$, then the decimated samples $D_\rho S_\rho q$ can be encoded by a total of $\mathscr{R} = 2\lfloor m/\rho \rfloor \log(2L\rho) = 2\eta \log(2L\rho)$ bits. Furthermore, suppose η is fixed as $m \to \infty$, then as a function of the total number of bits used, the reconstruction error $\mathscr{E}$ is*

$$\mathscr{E}(\mathscr{R}) \le C_{F,L}\|u\|_\infty 2^{-\frac{1}{2\eta}\mathscr{R}},$$

where $C_{F,L} \le \pi L(\sigma(\bar{F}) + \|\bar{F}_\eta\|)$, and $\bar{F}$ is defined above.

For $\rho \mid m$, we have a better estimate

$$\mathscr{E}(\mathscr{R}) \leq C_{k,L}\|u\|_\infty 2^{-\frac{1}{2\eta}\mathscr{R}},$$

where $C_{k,L} \leq \frac{\pi k L}{\eta}(\frac{2\pi(k+1)}{\sqrt{3}}+1)$, independent of ρ. The optimal exponent $\frac{1}{2k}$ will be achieved in the case $\rho = m/k \in \mathbb{N}$.

The more general result is as follows:

Theorem 2.3.5 (Decimation for Unitarily Generated Frames (UGF))**.** *Given Ω, ϕ_0, $\{\lambda_j\}_j$, $\{v_j\}_j$, and $\Phi = \Phi_{m,k}$ as the generator, base vector, eigenvalues, eigenvectors, and the analysis operator of the corresponding UGF, respectively, suppose*

- $\{\lambda_j\}_{j=1}^k \subset [-\eta/2, \eta/2] \cap \mathbb{Z}$,
- $C_{\phi_0} = \min_s |<\phi_0, v_s>|^2 > 0$, *and*
- $\rho \mid m$,

where $\eta = m/\rho$, then the following statements are true:

(a) ***Signal reconstruction:*** *$D_\rho S_\rho \Phi_{m,k} \in \mathbb{C}^{\eta\times k}$ has rank k.*

(b) ***Error estimate:*** *For the dual frame $F = (D_\rho S_\rho \Phi_{m,k})^\dagger D_\rho S_\rho$, the reconstruction error $\mathscr{E}_{m,\rho}$ satisfies*

$$\mathscr{E}_{m,\rho} \leq \frac{\pi}{2\eta C_{\phi_0}}(2\pi \max_{1\leq j\leq k}|\lambda_j| + 1)\|u\|_\infty \frac{1}{\rho}.$$

(c) ***Efficient data storage:*** *Suppose the length of the quantization alphabet is $2L$, then the total number of bits used to record the quantized samples are*

$\mathscr{R} = 2\eta \log(2L\rho)$ bits. Furthermore, suppose $\eta = m/\rho$ is fixed as $m \to \infty$, then as a function of the total number of bits used, $\mathscr{E}_{m,\rho}$ satisfies

$$\mathscr{E}(\mathscr{R}) \leq C_{k,\phi_0,L,\eta} \|u\|_\infty 2^{-\frac{1}{2\eta}\mathscr{R}},$$

where $C_{k,\phi_0,L,\eta} = \frac{\pi L}{\eta C_{\phi_0}}(2\pi \max_{1\leq j\leq k} |\lambda_j| + 1)$, independent of ρ.

Remark 2.3.6. For Theorem 2.3.4 and 2.3.5, if both the signal and the frame are real, then the total number of bits used will be $\mathscr{R} = \eta \log(2L\rho)$ bits, half the amount needed for the complex case.

One additional property of decimation is its multiplicative structure.

Theorem 2.3.7 (The Multiplicative Structure of Decimation Schemes)**.** *Suppose $\rho \mid m$ and $\rho = \rho_1\rho_2$, then the (m, ρ)-decimation is equal to the successive iterations of an (m, ρ_1)-decimation coupled by an $(m/\rho_1, \rho_2)$-decimation.*

Besides the first order alternative decimation in Theorem 2.3.5, it is also possible to generalize the result to the second order decimation. For such a decimation process, the reconstruction error decays quadratically (as opposed to linearly in Theorem 2.3.5) with respect to the oversampling ratio ρ and exponentially with respect to the bit usage.

Theorem 2.3.8 (Second Order Decimation for UGF)**.** *With the same assumptions as Theorem 2.3.5 and the additional requirement that the eigenvalues are nonzero, the following statements are true:*

(a) ***Signal reconstruction:*** *$D_\rho S_\rho^2 \Phi_{m,k} \in \mathbb{C}^{\eta\times k}$ has rank k.*

(b) ***Error estimate:*** *For the dual frame* $F = (D_\rho S_\rho^2 \Phi_{m,k})^\dagger D_\rho S_\rho^2$*, the reconstruction error* $\mathscr{E}_{m,\rho,r}$ *has quadratic error decay rate with respect to the oversampling ratio* ρ*:*

$$\mathscr{E}_{m,\rho,r} \leq \frac{\pi^2}{4\eta C_{\phi_0}} \left(9 + \eta (2\pi \max_{1 \leq j \leq k} |\lambda_j| \frac{1}{\eta})^2 \right) \|u\|_\infty \frac{1}{\rho^2}.$$

(c) ***Efficient data storage:*** *Suppose the length of the quantization alphabet is* $2L$*, then the total number of bits used to record the quantized samples is* $\mathscr{R} = 4\eta \log(2Lm)$ *bits. Furthermore, suppose* $\eta = m/\rho$ *is fixed as* $m \to \infty$*, then as a function of the total number of bits used* $\mathscr{E}_{m,\rho}$ *satisfies*

$$\mathscr{E}(\mathscr{R}) \leq C_{k,\phi_0,L,\eta} \|u\|_\infty 2^{-\frac{1}{2\eta}\mathscr{R}},$$

where $C_{k,\phi_0,L,\eta} = \frac{\pi^2}{4\eta C_{\phi_0}} \left(9 + \eta (2\pi \max_{1 \leq j \leq k} |\lambda_j| \frac{1}{\eta})^2 \right) (2L\eta)^2$*, independent of* ρ*.*

To better demonstrate the ideas in the proof, Theorem 2.3.4 will be proven separately in Section 2.3.1 even though it is essentially a special case of Theorem 2.3.5. Theorem 2.3.5 will be proven in Section 2.3.2, and Theorem 2.3.7 in Section 2.3.3. The proof of Theorem 2.3.8 is given in Section 2.3.4.

2.3.1 Decimation for Finite Harmonic Frames

To prove Theorem 2.3.4, we break down the proof into the following steps: first, we investigate properties of $D_\rho S_\rho E$, the decimated version of the frame E. Then, we examine the effect of $D_\rho S_\rho \Delta$, which is essential for our error estimate.

2.3.1.1 The Scaling Effect of Decimation

Let $E = E^{m,k} = (e_{l,j})_{l,j} = (\frac{1}{\sqrt{k}} \exp(-2\pi \imath n_j l/m))_{l,j}$ where $\{n_j\}_j$ are distinct. For any $\rho \leq m$, we have the following lemma:

Lemma 2.3.9. *S_ρ and E satisfy*

$$\begin{cases} S_\rho E = E\bar{C} & \text{if} \quad n_j \neq 0 \quad \forall j \\ S_\rho E = E\bar{C} - K & \text{if} \quad n_{j_0} = 0 \quad \text{for some } j_0, \end{cases}$$

where $\bar{C} \in \mathbb{C}^{k \times k}$ is a diagonal matrix with entries

$$\bar{C}_{j,j} = \begin{cases} \frac{\sin(\rho n_j \pi/m)}{\rho \sin(n_j \pi/m)} e^{\pi \imath (\rho-1) n_j/m} & \text{if} \quad n_j \neq 0 \\ 1 & \text{if} \quad n_j = 0, \end{cases} \tag{2.9}$$

and K is zero except for the j_0-th column, where

$$K_{l,j_0} = \begin{cases} \frac{m}{\rho\sqrt{k}} & \text{if} \quad 1 \leq l \leq \rho - 1 \\ 0 & \text{otherwise.} \end{cases}$$

In either case $D_\rho S_\rho E = D_\rho E\bar{C}$ as $D_\rho K = 0$.

Remark 2.3.10. In (2.8), one observes that S_ρ differs from an actual circulant matrix $\tilde{S}_\rho$ by a matrix L with $1/\rho$ on every entry of the first $\rho - 1$ rows and zero otherwise. Since $D_\rho L = 0$, we can conclude that $D_\rho S_\rho = D_\rho \tilde{S}_\rho$. Thus, it is possible to consider $D_\rho \tilde{S}_\rho$, which is a more natural formulation of decimation than the alternative decimation.

Proof. We start with the computation on $\rho \le l \le m$. First, suppose $n_j \neq 0$. Then, by (2.7),

$$\begin{aligned}(S_\rho^+ E)_{\rho,j} &= \frac{1}{\rho\sqrt{k}} \sum_{s=1}^{\rho} \exp(-2\pi\imath n_j s/m) \\ &= \frac{1}{\rho\sqrt{k}} e^{-\pi\imath(\rho+1)n_j/m} \frac{e^{\pi\imath(\rho-1)n_j/m}(1 - e^{-2\pi\imath\rho n_j/m})}{1 - \exp(-2\pi\imath n_j/m)} \\ &= \frac{1}{\sqrt{k}} e^{-\pi\imath(\rho+1)n_j/m} \frac{\sin(\rho n_j\pi/m)}{\rho \sin(n_j\pi/m)}.\end{aligned}$$

For $\rho \le l \le m$,

$$\begin{aligned}(S_\rho E)_{l,j} = (S_\rho^+ E)_{l,j} &= (S_\rho^+ E)_{\rho,j} \exp(-2\pi\imath(l-\rho)n_j/m) \\ &= \frac{1}{\rho\sqrt{k}} \exp(-2\pi\imath l n_j/m) \frac{\sin(\rho n_j\pi/m)}{\sin(n_j\pi/m)} e^{\pi\imath(\rho-1)n_j/m} \\ &= E_{l,j} \frac{\sin(\rho n_j\pi/m)}{\rho \sin(n_j\pi/m)} e^{\pi\imath(\rho-1)n_j/m}.\end{aligned}$$

If $n_j = 0$, then $(S_\rho^+ E)_{l,j} = \frac{1}{\sqrt{k}} = E_{l,j}$.

For $l \le \rho$, we make the following observation:

$$(S_\rho)_{l,j} + \frac{1}{\rho} = (S_\rho)_{l+\rho,j+\rho},$$

with the cyclic convention on indices. Then for $l \le \rho - 1$, noting that $\exp(-2\pi\imath n_j(s+$

$m)/m) = \exp(-2\pi\imath n_j s/m)$,

$$\begin{aligned}
(S_\rho E)_{l,j} &= \sum_{s=1}^{m} (S_\rho)_{l,s} E_{s,j} \\
&= \sum_{s=1}^{m} (S_\rho)_{l+\rho,s+\rho} E_{s,j} - \frac{1}{\rho}\sum_{s=1}^{m} E_{s,j} \\
&= \sum_{s=1}^{m} (S_\rho)_{l+\rho,s+\rho} E_{s+\rho,j} \exp(2\pi\imath\rho n_j/m) - \frac{m}{\rho\sqrt{k}}\delta(n_j) \\
&= E_{l+\rho,j} e^{2\pi\imath\rho n_j/m} \frac{\sin(\rho n_j\pi/m)}{\rho\sin(n_j\pi/m)} e^{\pi\imath(\rho-1)n_j/m} - \frac{m}{\rho\sqrt{k}}\delta(n_j) \\
&= E_{l,j} \frac{\sin(\rho n_j\pi/m)}{\rho\sin(n_j\pi/m)} e^{\pi\imath(\rho-1)n_j/m} - \frac{m}{\rho\sqrt{k}}\delta(n_j).
\end{aligned}$$

□

Now we can give the condition for which $D_\rho S_\rho E$ has full rank.

Proposition 2.3.11. *The following statements are equivalent:*

- $D_\rho S_\rho E$ *has full rank.*
- $\{\rho n_j\}_{j=1}^{k}$ *are distinct residues modulo* m*, and* $\rho n_j = 0$ *modulo* m *implies* $n_j = 0$.

Proof. By Lemma 2.3.9, we see that

$$D_\rho S_\rho E = D_\rho E\bar{C} = D \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_m \end{pmatrix} \bar{C} = \begin{pmatrix} E_\rho \\ E_{2\rho} \\ \vdots \\ E_{\eta\rho} \end{pmatrix} \bar{C}.$$

$D_\rho E$ is a sub-matrix of a Vandermonde matrix with parameters $\{\exp(-2\pi\imath\rho n_j/m)\}_{j=1}^{k}$.

Thus this matrix has full rank if and only if $\{\rho n_j\}_{j=1}^k$ are distinct modulo m. On the other hand, $\bar{C}$ is an invertible diagonal matrix if and only if $\bar{C}_{j,j} \neq 0$. It is true when $\rho n_j \neq 0$ for all j except if $n_j = 0$ to begin with.

□

Remark 2.3.12. $|\{-\lfloor \eta/2 \rfloor, \ldots, \lfloor \eta/2 \rfloor\}| \geq \eta$, and if $\{n_j\}_{j=1}^k \subset \{-\lfloor \eta/2 \rfloor, \ldots, \lfloor \eta/2 \rfloor\}$ are distinct residues modulo m, then $\{\rho n_j\}_j$ are distinct since elements of $\{-\lfloor \eta/2 \rfloor, \ldots, \lfloor \eta/2 \rfloor\}$ are in different cosets of $(\mathbb{Z}/m\mathbb{Z})/ker(\sigma)$ where $\sigma : \mathbb{Z}/m\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z}$ satisfies $\sigma(x) = \rho x$.

From Lemma 2.3.9, we see that $D_\rho S_\rho E = D_\rho E \bar{C}$. Thus for any dual $\tilde{F}$ to $D_\rho E \bar{C}$, $\tilde{F} = \bar{C}^{-1}\bar{F}$ where $\bar{F}$ is a dual to $D_\rho E$. The estimate of $\|\bar{C}^{-1}\|_2$ is described in Proposition 2.3.14, and we need a lemma for this proposition:

Lemma 2.3.13. *Given any number $\alpha > 1$, the function*

$$h(x) = h_\alpha(x) = \frac{\sin(\alpha x)}{\alpha \sin(x)}$$

is even and strictly decreasing in $(0, \pi/(2\alpha))$. Moreover, $\min_{x \in [-\pi/2\alpha, \pi/2\alpha]} h_\alpha(x) \geq \frac{2}{\pi}$.

Proof. Given any $\alpha > 1$, note that $\lim_{x \to 0} h(x) = 1$. Taking the derivative of h, we have

$$h'(x) = \frac{\alpha \cos(\alpha x)\sin(x) - \sin(\alpha x)\cos(x)}{\alpha^2 \sin^2(x)} = \frac{\cos(\alpha x)\cos(x)}{\alpha^2 \sin^2(x)} \Big(\alpha \tan(x) - \tan(\alpha x) \Big).$$

The first factor on the right hand side is even and positive in $(0, \pi/(2\alpha))$, while the second one $\alpha \tan(x) - \tan(\alpha x)$ is odd and decreasing in $(0, \pi/(2\alpha))$ by taking yet

another derivative. Thus, the derivative of h is odd and negative in $(0, \pi/(2\alpha)]$. That is, on $I_\alpha = [-\pi/(2\alpha), \pi/(2\alpha)]$, h achieves global maximum at $x = 0$ and minimum at $x = \pi/(2\alpha)$. At the minimum point,

$$h_\alpha\left(\frac{\pi}{2\alpha}\right) = \frac{1}{\alpha \sin(\pi/(2\alpha))} \geq \frac{2}{\pi}$$

by noting that $\sin(z) \leq z$ for any $z > 0$. □

Proposition 2.3.14. *If $\{n_j\}_{j=1}^k$ are concentrated in $[-\eta/2, \eta/2]$ in $\mathbb{Z}/m\mathbb{Z}$, then*

$$\|\bar{C}^{-1}\|_2 \leq \frac{\pi}{2}.$$

Proof. By (2.9), we see that

$$|\bar{C}_{l,l}| = \left|\frac{1}{\rho}\frac{\sin(\rho n_l \pi/m)}{\sin(n_l \pi/m)}\right|$$

with the convention that $\sin(\rho \cdot 0)/(\rho \sin(0)) = 1$. Thus,

$$\|\bar{C}^{-1}\|_2 = \max_{1\leq l\leq k}\left(|\bar{C}(l)|^{-1}\right) = \left(\min_{1\leq l\leq k}\{\left|\frac{\sin(\rho n_l \pi/m)}{\rho \sin(n_l \pi/m)}\right|\}\right)^{-1}.$$

Using the result from Lemma 2.3.13 with $\alpha = \rho \geq 1$, we see that $\|\bar{C}^{-1}\|_2 \leq \frac{\pi}{2}$. □

2.3.1.2 Effect of S_ρ on the Difference Structure Δ

Here, we describe the effect of $D_\rho S_\rho \Delta$ in Proposition 2.3.16, which is directly connected to the proof of Theorem 2.3.4.

Lemma 2.3.15. *$S_\rho = \frac{1}{\rho}\bar{\Delta}_\rho\Delta^{-1}$ where*

$$(\bar{\Delta}_\rho)_{l,j} = \begin{cases} \delta([j-l]) - \delta([\rho+j-l]) & \text{if} \quad j \neq m, \\ \delta([j-l]) & \text{if} \quad j = m, \end{cases}$$

and $\delta : \mathbb{Z}/m\mathbb{Z} \to \{0,1\}$ is the Kronecker delta.

Proof. Let $\bar{\delta} : \mathbb{Z} \to \{0,1\}$ be the Kronecker delta on $\mathbb{Z}$. Then,

$$\begin{aligned}(\bar{\Delta}_\rho\Delta^{-1})_{l,j} &= \sum_{j\leq t\leq m-1} (\delta([t-l]) - \delta([\rho+t-l])) + \delta([j-m]) \\ &= \sum_{j\leq t\leq m} \delta([t-l]) - \sum_{j\leq t\leq m-1} (\bar{\delta}(\rho+t-l) + \bar{\delta}(\rho+t-l-m)).\end{aligned}$$

By definition,

$$\begin{cases} \sum_{j\leq t\leq m} \delta([t-l]) & = 1 \quad \text{if} \quad j \leq l \\ \sum_{j\leq t\leq m-1} \bar{\delta}(\rho+t-l) & = 1 \quad \text{if} \quad l \geq \rho+1,\ j \leq l-\rho \\ \sum_{j\leq t\leq m-1} \bar{\delta}(\rho+t-l-m) & = 1 \quad \text{if} \quad l \leq \rho-1,\ j \leq m-\rho+l. \end{cases}$$

Thus, splitting into the cases $l \leq \rho-1$, $l = \rho$, and $l \geq \rho+1$, we see that

$$\frac{1}{\rho}\bar{\Delta}_\rho\Delta^{-1} = S_\rho,$$

as claimed. □

Proposition 2.3.16. *Given any $n \in \mathbb{N}$, let $\Delta^{(n)} \in \mathbb{N}^{n\times n}$ denote the n-dimensional*

backward difference matrix. For $\rho|m$, one has

$$D_\rho S_\rho \Delta^{(m)} = \frac{1}{\rho}\Delta^{(m/\rho)} D_\rho.$$

Proof. If $\rho \mid m$,

$$D_\rho S_\rho \Delta^{(m)} = \frac{1}{\rho} D_\rho \bar{\Delta}_\rho.$$

Now, note that, for $s \neq m$,

$$\begin{aligned}(D_\rho \bar{\Delta}_\rho)_{l,s} &= (\bar{\Delta}_\rho)_{l\rho,s} \\ &= \delta(s - l\rho) - \delta(s + \rho - l\rho) = \delta(s - l\rho) - \delta(s - (l-1)\rho) \\ &= (\Delta D_\rho)_{l,s}.\end{aligned}$$

For $s = m$, $(D_\rho \bar{\Delta}_\rho)_{l,m} = \delta(m - l\rho) = (\Delta D_\rho)_{l,m}$. □

2.3.1.3 Proof of Theorem 2.3.4

Before proving Theorem 2.3.4, we shall need two more lemmas:

Lemma 2.3.17. *For any $E = E^{n,k}$ with $n \geq k$, suppose $\{n_j\}_j$ are concentrated between $[-k/2, k/2]$, then E^* has frame variation $\sigma(E^*) \leq \frac{2\pi(k+1)}{\sqrt{3}}$.*

Proof.

$$\begin{aligned}
\sigma(E^*) &= \frac{1}{\sqrt{k}}\sum_{l=1}^{n-1}(\sum_{j=1}^{k}|e^{-2\pi\imath l n_j/n} - e^{-2\pi\imath(l+1)n_j/n}|^2)^{1/2} \\
&= \frac{1}{\sqrt{k}}\sum_{l=1}^{n-1}(\sum_{j=1}^{k}|1 - e^{2\pi\imath n_j/n}|)^{1/2} \\
&\leq \frac{1}{\sqrt{k}}\sum_{l=1}^{n-1}(\sum_{j=1}^{k}(2\pi n_j/n)^2)^{1/2} \\
&\leq \frac{1}{\sqrt{k}}2\pi\frac{n-1}{n}(\sum_{j=-k/2}^{k/2} n_j^2)^{1/2} \\
&\leq \frac{1}{\sqrt{k}}2\pi\sqrt{2\cdot\frac{k/2(k/2+1)(k+1)}{3}} \\
&\leq \frac{2\pi(k+1)}{\sqrt{3}}.
\end{aligned}$$

□

Lemma 2.3.18 ([6], Theorem III.7)**.** *Given a stable ΣΔ quantization scheme with a mid-rise uniform quantizer of gap δ, if the frame $T = \{e_j\}_{j=1}^m$ satisfies the zero sum condition*

$$\sum_{j=1}^{m} e_j = 0,$$

then the auxiliary variable u_m has

$$|u_m| = \begin{cases} 0, & \text{if } m \text{ even}, \\ \delta/2, & \text{if } m \text{ odd}. \end{cases}$$

Now we are ready to give the proof of Theorem 2.3.4.

Proof. of Theorem 2.3.4:

Adopting the notations above, we see that the reconstruction error is

$$\begin{aligned}\|x-\tilde{F}(D_\rho S_\rho q)\|_2 &= \|\tilde{F}(D_\rho S_\rho)(y-q)\|_2 \\ &= \|\tilde{F}D_\rho S_\rho \Delta^{(m)} u\|_2 \\ &= \|\frac{1}{\rho}\bar{C}^{-1}\bar{F}\Delta^{(\eta)}D_\rho u\|_2 \leq \frac{1}{\rho}\|\bar{C}^{-1}\|_2\|\bar{F}\Delta^{(\eta)}D_\rho u\|_2,\end{aligned}$$

where the second equality comes from (2.6), and the third equality follows from Proposition 2.3.16 along with the fact that $\tilde{F} = \bar{C}^{-1}\bar{F}$ with $\bar{F}$ being the canonical dual frame to $D_\rho E$. Suppose $\bar{F} = (\bar{F}_1, \cdots, \bar{F}_\eta)$, then

$$\bar{F}\Delta^{(\eta)}D_\rho u = \sum_{s=1}^{\eta-1} u_{s\rho}(\bar{F}_s - \bar{F}_{s+1}) + u_{\eta\rho}\bar{F}_\eta. \tag{2.10}$$

By Proposition 2.3.14 and (2.10),

$$\frac{1}{\rho}\|\bar{C}^{-1}\|_2\|\bar{F}\Delta^{(\eta)}u^{(\eta)}\|_2 \leq \frac{\pi}{2\rho}(\sigma(\bar{F}) + \|\bar{F}_\eta\|_2)\|u\|_\infty.$$

For the case $\rho \mid m$, we note that $E^{m/\rho,k}$ is a tight frame with frame bound $\frac{m}{k\rho}$. In particular, $(E^{m/\rho,k})^* E^{m/\rho,k} = \frac{m}{k\rho} I_k$. Thus, by Lemma 2.3.17 ,

$$\sigma(\bar{F}) \leq \frac{k}{m/\rho}\frac{2\pi(k+1)}{\sqrt{3}}.$$

Thus, we have obtained the following error bound

$$\mathscr{E}_\rho = \|x-\tilde{F}(D_\rho S_\rho q)\|_2 \leq \frac{k}{m/\rho}\frac{\pi}{2\rho}(\frac{2\pi(k+1)}{\sqrt{3}}+1)\|u\|_\infty = \frac{\pi}{2}(\frac{2\pi(k+1)}{\sqrt{3}}+1)\|u\|_\infty\frac{k}{m}. \tag{2.11}$$

Furthermore, by Lemma 2.3.18, if m, k are even, n_j's are all nonzero, and $\rho \mid m$, then $u_{\eta\rho} = u_m = 0$. With that there is a better estimate

$$\mathscr{E}_\rho = \|x - \tilde{F}(D_\rho S_\rho q)\|_2 \le \frac{k}{m/\rho}\frac{\pi}{2\rho}\frac{2\pi(k+1)}{\sqrt{3}}\|u\|_\infty = \frac{\pi^2(k+1)}{\sqrt{3}}\frac{k}{m}\|u\|_\infty. \tag{2.12}$$

Letting $F = \tilde{F} D_\rho S_\rho$, Theorem 2.3.4 (b) is now proven.

For Theorem 2.3.4 (c), note that for mid-rise uniform quantizers $\mathscr{A} = \mathscr{A}_0 + \imath\mathscr{A}_0$ with length $2L$, each entry q_j of $q \in \mathbb{C}^m$ is of the form

$$q_j = \big((2s_j + 1) + \imath(2t_j + 1)\big)\frac{\delta}{2}, \quad -L \le s_j,\, t_j \le L - 1.$$

Then, each entry in $D_\rho S_\rho q$ is the average of ρ entries in q which has the form

$$(D_\rho S_\rho q)_j = \big((2\tilde{s}_j + \rho) + \imath(2\tilde{t}_j + \rho)\big)\frac{\delta}{2\rho}, \quad -L\rho \le \tilde{s}_j,\, \tilde{t}_j \le (L-1)\rho.$$

There are at most $((2L-1)\rho+1)^2 \le (2L\rho)^2$ choices per entry with $\eta = m/\rho$ entries in total. Thus, the vector $D_\rho S_\rho q$ can be encoded by $\mathscr{R} = 2\eta \log(2L\rho)$ bits. Noting that $\frac{1}{m} \le \frac{1}{\eta} \cdot \frac{1}{\rho}$ and

$$e^{-\frac{1}{2\eta}\mathscr{R}} = \frac{1}{2L\rho},$$

for any estimate we have

$$\mathscr{E} \le C\frac{1}{m} \le C\frac{1}{\eta}\frac{1}{\rho} = C\frac{2L}{\eta}e^{-\frac{1}{2\eta}\mathscr{R}},$$

for some $C > 0$. Substituting the suitable constant for each case, we have

$$\mathscr{E}(\mathscr{R}) \leq C_{F,L}\|u\|_\infty 2^{-\frac{1}{2\eta}\mathscr{R}},$$

where $C_{F,L} \leq \pi L(\sigma(\bar{F}) + \|\bar{F}_\eta\|_2)$. If $\rho \mid m$, then by (2.11), (2.12),

$$\mathscr{E}(\mathscr{R}) \leq C_{k,L}\|u\|_\infty 2^{-\frac{1}{2\eta}\mathscr{R}},$$

where $C_{k,L} \leq \frac{\pi k L}{\eta}(\frac{2\pi(k+1)}{\sqrt{3}} + 1)$, independent of ρ.

□

2.3.2 Generalization: Decimation on Unitarily Generated Frames

Upon examining the proof of Theorem 2.3.4, one can see the following interaction between decimation and the existing sampling scheme:

- Commutativity: $D_\rho S_\rho E^{m,k} = E^{m/\rho,k}\bar{C}_{m,\rho}$.
- Scalability: $D_\rho S_\rho \Delta^{(m)} u = \frac{1}{\rho}\Delta^{(\eta)} D_\rho u$.

Fixing the $\Sigma\Delta$ quantization scheme for now, any family of frames satisfying the commutativity condition shall be compatible with decimation, yielding exponential error decay with respect to the bit usage. One example is the unitarily generated frames.

The collection of such elements $T_u = \{\phi_j^{(m)}\}_{j=1}^m$ is the frame of interest.

Lemma 2.3.19. *For the same D_ρ and S_ρ along with the analysis operator $\Phi \in \mathbb{C}^{m\times k}$*

of T_u generated by $(\Omega, \{\lambda_j\}_{j=1}^k, \phi_0)$,

$$D_\rho S_\rho \Phi_{m,k} = \begin{pmatrix} (\phi_\rho^{(m)})^* \\ (\phi_{2\rho}^{(m)})^* \\ \vdots \\ (\phi_{\eta\rho}^{(m)})^* \end{pmatrix} \bar{C}_{m,\rho},$$

where $\eta = \lfloor m/\rho \rfloor$ and $\bar{C}_{m,\rho} = \frac{1}{\rho}\sum_{s=1}^{\rho} U_{(s-\rho)/m}^*$ has eigenvalues $\{e^{\pi\imath(\rho-1)\lambda_j/m}\frac{\sin(\rho\lambda_j\pi/m)}{\rho\sin(\lambda\pi/m)}\}_j$. In particular, if $\rho \mid m$, then

$$D_\rho S_\rho \Phi_{m,k} = \Phi_{m/\rho,k}\bar{C}_{m,\rho}.$$

Proof. First, note that $S_\rho = \tilde{S}_\rho + L$, where L has value $1/\rho$ on the first $\rho - 1$ rows and 0 otherwise, and $D_\rho L = 0$. Moreover, for any $1 \le t \le m$,

$$\begin{aligned} (\tilde{S}_\rho \Phi_{m,k})_t &= (\frac{1}{\rho}\sum_{t-\rho+1}^{t} U_{s/m}\phi_0)^* \\ &= (\frac{1}{\rho}\sum_{t+1}^{t+\rho} U_{(s-\rho)/m}\phi_0)^* \\ &= \phi_t^* \cdot \frac{1}{\rho}\sum_{s=1}^{\rho} U_{(s-\rho)/m}^* = (\Phi_{m,k})_t \cdot \frac{1}{\rho}\sum_{s=1}^{\rho} U_{(s-\rho)/m}^*. \end{aligned} \tag{2.13}$$

Thus, $D_\rho S_\rho \Phi_{m,k} = D_\rho \Phi_{m,k}\bar{C}_{m,\rho} + D_\rho L \Phi_{m,k} = D_\rho \Phi_{m,k}\bar{C}_{m,\rho}$.

Note that we can diagonalize $U_t = BT_tB^*$ where B is a unitary matrix and T_t is a diagonal matrix with entries $\{e^{2\pi\imath\lambda_j t}\}_{j=1}^k$. Then, $B^*\bar{C}_{m,\rho}B = \frac{1}{\rho}\sum_{s=1}^{\rho}(B^*U_{(s-\rho)/m}B)^*$

is a diagonal matrix, with entries

$$(B * \bar{C}_{m,\rho} B)_{j,j} = \frac{1}{\rho} \sum_{s=1}^{\rho} \exp(-2\pi\imath \frac{s-\rho}{m} \lambda_j)$$
$$= \frac{1}{\rho} e^{2\pi\imath \frac{\lambda_j(\rho-1)}{m}} \frac{e^{-2\pi\imath\rho\lambda_j/m} - 1}{e^{-2\pi\imath\lambda_j/m} - 1} = e^{\pi\imath(\rho-1)\lambda_j/m} \frac{\sin \rho\lambda_j\pi/m}{\rho \sin(\lambda_j\pi/m)}.$$

□

Now, we can find the conditions under which $D_\rho S_\rho \Phi_{m,k}$ has full rank:

Proposition 2.3.20. *Let $\{v_s\}_{s=1}^k$ be a set of orthonormal eigenvectors of Ω with eigenvalues $\{\lambda_s\}_{s=1}^k$. Suppose*

- *$\rho \mid m$,*
- *$\{\rho(\lambda_s - \lambda_t)\}_{s \neq t}$ are nonzero integers modulo m, and*
- *the base vector $\phi_0 = \sum_s c_s v_s$ satisfies $c_s \neq 0$ for all s,*

then $\Phi_{m/\rho,k}$ is a frame with frame bounds

$$\left(\frac{m}{\rho} \min_s |c_s|^2\right) \|x\|_2^2 \leq \sum_{s=1}^{m/\rho} |<x, \phi_{s\rho}^{(m)}>|^2 \leq \left(\frac{m}{\rho} \max_s |c_s|^2\right) \|x\|_2^2.$$

In particular, the frame operator $\mathcal{S}_{m/\rho} = \Phi^*\Phi$ satisfies $\|\mathcal{S}_{m/\rho}^{-1}\|_2 \leq \frac{1}{\eta \min |c_i|^2}$.

Proof. Suppose the assumptions above are true, then given an arbitrary $x \in \mathbb{C}^k$,

$$\begin{aligned}
\sum_{s=1}^{m/\rho} |<x, \phi_{s\rho}^{(m)}>|^2 &= \sum_{s=1}^{m/\rho} |\sum_{t=1}^{k} <x, v_t><v_t, \phi_{s\rho}^{(m)}>|^2 \\
&= \sum_{s=1}^{m/\rho} |\sum_{t=1}^{k} <x, v_t><U_{-s\rho/m} v_t, \phi_0>|^2 \\
&= \sum_{s=1}^{m/\rho} |\sum_{t=1}^{k} e^{-2\pi \imath s\rho\lambda_t/m} <x, v_t><v_s, \phi_0>|^2 \\
&= \sum_{j,l=1}^{k} <x, v_j>\overline{<x, v_l>}<v_j, \rho>\overline{<v_l, \phi_0>} \sum_{s=1}^{m/\rho} e^{-2\pi \imath s\rho(\lambda_j - \lambda_l)/m} \\
&= \frac{m}{\rho} \sum_{j=1}^{k} |<x, v_j>|^2 |<v_j, \phi_0>|^2,
\end{aligned}$$

where the second equality follows from the fact that U_t is unitary, the fourth by expanding the sums, and the last one from the following equality

$$\sum_{s=1}^{m/\rho} \exp(-2\pi \imath s\rho(\lambda_j - \lambda_l)/m) = \begin{cases} \frac{m}{\rho} & \text{if} \quad j = l \\ 0 & \text{if} \quad j \neq l. \end{cases}$$

Finally, we have

$$\left(\frac{m}{\rho} \min_s |c_s|^2\right) \|x\|_2^2 \leq \frac{m}{\rho} \sum_{j=1}^{k} |<x, v_j>|^2 |<v_j, \phi_0>|^2 \leq \left(\frac{m}{\rho} \max_s |c_s|^2\right) \|x\|_2^2.$$

□

Moreover, with the same proof in Proposition 2.3.14, we have the estimate on $\|\bar{C}_{m,\rho}^{-1}\|_2$:

Proposition 2.3.21. *If the eigenvalues $\{\lambda_j\}_{j=1}^k$ of the generator Ω are concentrated*

between $[-m/(2\rho), m/(2\rho)]$, *then*

$$\|\bar{C}_{m,\rho}^{-1}\|_2 = \|B^*\bar{C}_{m,\rho}^{-1}B\|_2 = \max_{1\le j\le k}\left\{\left|\frac{\sin(\rho\lambda_j\pi/m)}{\rho\sin(\lambda_j\pi/m)}\right|^{-1}\right\} \le \frac{\pi}{2}.$$

Also, we need to consider the frame variation of $\Phi^*_{m/\rho,k}$.

Lemma 2.3.22. $\sigma(\Phi^*_{m/\rho}) \le 2\pi \max_{1\le j\le k}|\lambda_j|$.

Proof. Following the same process of Lemma 2.3.17, we see that

$$\begin{aligned}
\sigma(\Phi^*_{m/\rho}) &= \sum_{s=1}^{m/\rho-1} \|(U_{s\rho/m} - U_{(s+1)\rho/m})\phi_0\|_2 \\
&= \sum_{s=1}^{m/\rho-1} \|U_{s\rho/m}(1 - U_{\rho/m})\phi_0\|_2 \\
&= \sum_{s=1}^{m/\rho-1} \|(1 - U_{\rho/m})\phi_0\|_2 \\
&= \sum_{s=1}^{m/\rho-1} \|\sum_{j=1}^{k} c_j[1 - e^{2\pi\imath\lambda_j\rho/m}]v_j\|_2 \\
&= \sum_{s=1}^{m/\rho-1} (\sum_{j=1}^{k} |c_j|^2|e^{2\pi\imath\lambda_j\rho/m} - 1|^2)^{1/2} \\
&\le \sum_{s=1}^{m/\rho-1} \left(\sum_{j=1}^{k} |c_j|^2 \cdot \left(2\pi|\lambda_j|\frac{\rho}{m}\right)^2\right)^{1/2} \\
&\le \sum_{s=1}^{m/\rho-1} \left(\max_{1\le j\le k} 2\pi|\lambda_j|\frac{\rho}{m}\right) \cdot \|\phi_0\|_2 \le 2\pi \max_{1\le j\le k}|\lambda_j| \cdot \|\phi_0\|_2.
\end{aligned}$$

□

Now we are ready to prove Theorem 2.3.5.

Proof. of Theorem 2.3.5.

First of all, that $D_\rho S_\rho \Phi_{m,k} = \Phi_{m/\rho,k}\bar{C}_{m,\rho}$ has full rank follows from Proposition 2.3.20 and 2.3.21. For notational clarity, we shall denote $\Phi_{m/\rho,k} = \Phi_{m/\rho}$.

Let $\mathcal{S}_{m/\rho} = \Phi^*_{m/\rho}\Phi_{m/\rho}$ be the corresponding frame operator, then $\|\mathcal{S}^{-1}_{m/\rho}\|_2 \leq \rho/(mC_{\phi_0})$ where $C_{\phi_0} := \min_s |c_s|^2$. Also, note that, by Proposition 2.3.16,

$$\mathcal{S}^{-1}_{m/\rho}\Phi^*_{m/\rho}(D_\rho S_\rho)\Delta^{(m)}u = \frac{1}{\rho}\mathcal{S}^{-1}_{m/\rho}(\Phi^*_{m/\rho}\Delta^{(m/\rho)})D_\rho u.$$

Then, the reconstruction error $\|x - \bar{C}^{-1}_{m,\rho}\mathcal{S}^{-1}_{m/\rho}\Phi^*_{m/\rho}(D_\rho S_\rho)q\|_2$ is

$$\begin{aligned}\|x - \bar{C}^{-1}_{m,\rho}\mathcal{S}^{-1}_{m/\rho}\Phi^*_{m/\rho}(D_\rho S_\rho)q\|_2 &= \|\bar{C}^{-1}_{m,\rho}\mathcal{S}^{-1}_{m/\rho}\Phi^*_{m/\rho}(D_\rho S_\rho)\Delta^{(m)}u\|_2 \\ &\leq \frac{1}{\rho}\|\bar{C}^{-1}_{m,\rho}\|_2\|\mathcal{S}^{-1}_{m/\rho}\|_2(\sigma(\Phi^*_{m/\rho}) + \|\phi^{(m)}_m\|_2)\|D_\rho u\|_\infty \\ &\leq \frac{\pi}{2\rho}\frac{\rho}{mC_{\phi_0}}(\sigma(\Phi^*_{m/\rho}) + \|\phi^{(m)}_m\|_2)\|u\|_\infty,\end{aligned} \tag{2.14}$$

where $\|\bar{C}^{-1}_{m,\rho}\|_2 \leq \pi/2$ by Proposition 2.3.21.

Combining (2.14), Lemma 2.3.22, and the fact that $\|\phi^{(m)}_m\|_2 = \|U_1\phi_0\|_2 = \|\phi_0\|_2 = 1$, the reconstruction error $\mathscr{E}_{m,\rho}$ can be bounded by

$$\begin{aligned}\mathscr{E}_{m,\rho} &\leq \frac{\pi}{2\rho}\frac{\rho}{mC_{\phi_0}}(\sigma(\Phi^*_{m/\rho}) + 1)\|u\|_\infty \\ &\leq \frac{\pi}{2mC_{\phi_0}}(2\pi \max_{1\leq j\leq k}|\lambda_j| + 1)\|u\|_\infty \\ &= \frac{\pi}{2\eta C_{\phi_0}}(2\pi \max_{1\leq j\leq k}|\lambda_j| + 1)\|u\|_\infty\frac{1}{\rho}.\end{aligned}$$

Theorem 2.3.5 (c) follows verbatim from the proof in Theorem 2.3.4. □

2.3.3 The Multiplicative Structure of Decimation Schemes

In this section, we demonstrate the multiplicative structure of alternative decimation. In particular, given $m, \rho, \rho_1, \rho_2 \in \mathbb{N}$ fixed with $\rho = \rho_1\rho_2$ and $\rho \mid m$, consider the following operators:

$$\begin{aligned} &D_\rho \in \mathbb{N}^{(m/\rho)\times m}, && S_\rho \in \mathbb{R}^{m\times m}, \\ &D_{\rho_1} \in \mathbb{N}^{(m/\rho_1)\times m}, && S_{\rho_1} \in \mathbb{R}^{m\times m}, \\ &D_{\rho_2} \in \mathbb{N}^{(m/\rho)\times(m/\rho_1)}, && S_{\rho_2} \in \mathbb{R}^{(m/\rho_1)\times(m/\rho_1)}. \end{aligned}$$

We shall show that $D_\rho S_\rho = D_{\rho_2} S_{\rho_2} D_{\rho_1} S_{\rho_1}$.

Proof. of Theorem 2.3.7:

The (m, ρ)-decimation operator is $D_\rho S_\rho$ while the successive iterations of (m, ρ_1) and $(m/\rho_1, \rho_2)$-decimation combine to be $D_{\rho_2} S_{\rho_2} D_{\rho_1} S_{\rho_1}$.

Note that $D_{\rho_2} D_{\rho_1} = D_\rho$. Then, by Proposition 2.3.16,

$$\begin{aligned} D_{\rho_2} S_{\rho_2} D_{\rho_1} S_{\rho_1} &= (D_{\rho_2}\bar{\Delta}_{\rho_2})(\Delta^{(m/\rho_1)})^{-1}(D_{\rho_1}\bar{\Delta}_{\rho_1})(\Delta^{(m)})^{-1} \\ &= \Delta^{(m/\rho_1\rho_2)} D_{\rho_2}(\Delta^{(m/\rho_1)})^{-1}\Delta^{(m/\rho_1)} D_{\rho_1}(\Delta^{(m)})^{-1} \\ &= \Delta^{(m/\rho)} D_{\rho_2} D_{\rho_1}(\Delta^{(m)})^{-1} \\ &= D_\rho\bar{\Delta}_\rho(\Delta^{(m)})^{-1} = D_\rho S_\rho, \end{aligned}$$

which concludes our proof. □

The multiplicative property implies the possibility to conduct decimation with multiple steps, gradually down-sizing the dimension m. It can be particularly useful

for parallel computation and transmission of data through multiple devices with scarce storage resources. In particular, for each stage, it suffices to choose ρ_j to be a small number dividing m. It reduces the waiting time between each transmission, and the amplification of quantized sample q will not be large after each stage.

Moreover, although the case where $\rho \nmid m$ does not produce this structure for frames, it is now possible to first reduce m to a number closer to k. Only at the last stage do we choose ρ that does not divide m. This yields the same result as direct division m/k by the remark above while possibly gaining sharper estimate on the error.

2.3.4 Extension to Second Order Decimation

So far, we have only defined decimation for the first order $\Sigma\Delta$ quantization, while its counterpart for bandlimited functions, introduced in Section 2.1.3, applies for arbitrary orders. Due to the boundary effect in finite dimensional spaces, it is harder to extend decimation to arbitrary orders. However, there is no issue generalizing this concept to the second order, as stated in Theorem 2.3.8. To prove the theorem, we shall need the following lemmas:

Lemma 2.3.23 (Effect of $D_\rho S_\rho^2$ on the Finite Frame)**.** *If none of the eigenvalues of* $U_{1/m}$ *are* 1*, then*

$$S_\rho \Phi_{m,k} = \Phi_{m,k} \bar{C}_{m,\rho}.$$

where $\bar{C}_{m,\rho} = \frac{1}{\rho} \sum_{s=1}^{\rho} U_{(s-\rho)/m}^*$ *has eigenvalues* $\{e^{\pi\imath(\rho-1)\lambda_j/m} \frac{\sin(\rho\lambda_j\pi/m)}{\rho \sin(\lambda\pi/m)}\}_j$*. In particu-*

lar, for any $r \in \mathbb{N}$,

$$D_\rho S_\rho^r \Phi_{m,k} = \Phi_{m/\rho,k} \bar{C}_{m,\rho}^r.$$

Remark 2.3.24. The proof is very similar to the one of Lemma 2.3.19. However, since we are now dealing with $D_\rho S_\rho^r$, we are no longer able to use the fact that $D_\rho L = 0$. Instead, we impose the condition that $U_{1/m}$ has no eigenvalue equal to 1.

Proof. First, note that if $\mathbb{1} \in \mathbb{C}^m$ is the constant vector with value 1, then

$$\mathbb{1}^* \Phi_{m,k} = (\sum_{s=0}^{m-1} U_{s/m} \phi_0)^* = \phi_0^* B (\sum_{s=0}^{m-1} T_{s/m})^* B^* = 0.$$

Given $1 \leq t \leq m$, note that $S_\rho = \tilde{S}_\rho - L$, where L has value $1/\rho$ on the first $\rho - 1$ rows and 0 otherwise, and $L\Phi_{m,k} = 0$. Then, by (2.13), $S_\rho \Phi_{m,k} = \tilde{S}_\rho \Phi_{m,k} = \Phi_{m,k} \bar{C}_{m,\rho}$. Using induction on r, $S_\rho^r \Phi_{m,k} = \Phi_{m,k} \bar{C}_{m,\rho}^r$, and $D_\rho S_\rho^r \Phi_{m,k} = \Phi_{m/\rho,k} \bar{C}_{m,\rho}^r$. The properties of $\bar{C}_{m,\rho}$ follow from Lemma 2.3.19. □

Lemma 2.3.25. *For any* $r, m, \rho \in \mathbb{N}$,

$$D_\rho \bar{\Delta}_\rho^r = (\Delta^{(m/\rho)})^r D_\rho.$$

Proof. By Proposition 2.3.16,

$$D_\rho \bar{\Delta}_\rho = D_\rho (\bar{\Delta}_\rho (\Delta^{(m)})^{-1}) \Delta^{(m)} = \Delta^{(m/\rho)} D_\rho.$$

Thus, for $r \in \mathbb{N}$, we have, by induction on r,

$$D_\rho \bar{\Delta}_\rho^r = \Delta^{(m/\rho)} D_\rho \bar{\Delta}_\rho^{r-1} = (\Delta^{(m/\rho)})^r D_\rho.$$

□

Lemma 2.3.26. *$\Delta^{-1}\bar{\Delta}_\rho\Delta = \bar{\Delta}_\rho + \mathcal{E}$, where $\mathcal{E}_{l,s} = \delta(s-(m-\rho))$.*

Proof. For $s \neq m$,

$$\begin{aligned}
(\Delta^{-1}\bar{\Delta}_\rho\Delta)_{l,s} &= \sum_{j,n} \Delta^{-1}_{l,j}(\bar{\Delta}_\rho)_{j,n}\Delta_{n,s} \\
&= \sum_{j=1}^{l} (\bar{\Delta}_\rho)_{j,s} - (\bar{\Delta}_\rho)_{j,s+1} \\
&= \sum_{j=1}^{l} \Big[\delta(s-j) - \delta(s+\rho-j) - \delta(s+1-j) + \delta(s+1+\rho-j)\Big] \\
&= \sum_{j=1}^{l} (\delta(s-j) - \delta(s+1-j)) - \sum_{j=1}^{l} (\delta(s+\rho-j) - \delta(s+1+\rho-j)) \\
&= \delta(s-l) - \delta(s+\rho-l) + \delta(s+\rho) = (\bar{\Delta}_\rho)_{l,s} + \delta(s+\rho),
\end{aligned}$$

where the $\delta(s+\rho) = \delta(s-(m-\rho))$ comes from the second term in the second-to-last line. When $s+1+\rho = m+1$, the term $\delta(s+1+\rho-j)$ wraps around, producing an additional -1.

When $s = m$,

$$(\Delta^{-1}\bar{\Delta}_\rho\Delta)_{l,s} = \sum_{j} \Delta^{-1}_{l,j}(\bar{\Delta}_\rho)_{j,m} = \sum_{j=1}^{l} \delta(m-j) = \delta(m-l).$$

Combining the two equations above, we see that $\Delta^{-1}\bar{\Delta}_\rho\Delta = \bar{\Delta}_\rho + \mathcal{E}$. □

Proposition 2.3.27. *For* $\Phi^*_{m/\rho,k} = (\phi_1^{(\eta)} \mid \cdots \mid \phi_\eta^{(\eta)})$,

$$\Phi^*_{m/\rho,k} D_\rho S_\rho^2 \Delta^2 = \frac{1}{\rho^2} \Phi^*_{m/\rho,k} \Delta^2 D_\rho + \frac{1}{\rho^2} V,$$

where V *is zero except for the* $(m-\rho)$*-th column, which is* $\phi_1^{(\eta)}$.

Proof. When $r = 2$, we consider the $(2,\rho)$-decimation operator $D_\rho S_\rho^2$. Then,

$$\begin{aligned} D_\rho S_\rho^2 \Delta^2 &= \frac{1}{\rho^2} D_\rho \bar{\Delta}_\rho \Delta^{-1} \bar{\Delta}_\rho \Delta \\ &= \frac{1}{\rho^2} D_\rho \bar{\Delta}_\rho (\bar{\Delta}_\rho + \mathcal{E}) \\ &= \frac{1}{\rho^2} \Delta^2 D_\rho + \frac{1}{\rho^2} D_\rho \bar{\Delta}_\rho \mathcal{E}, \end{aligned}$$

where the first term in the last line follows from Lemma 2.3.25. Now, $(\bar{\Delta}_\rho \mathcal{E})_{l,s} = \delta(l-\rho)\delta(s+\rho)$, and $(D_\rho \bar{\Delta}_\rho \mathcal{E})_{l,s} = \delta(l-1)\delta(s+\rho)$. Thus,

$$\Phi^*_{m/\rho,k} D_\rho S_\rho^2 \Delta^2 = \frac{1}{\rho^2} \Phi^*_{m/\rho,k} \Delta^2 D_\rho + \frac{1}{\rho^2} V.$$

□

Lemma 2.3.28. *For any* $r \in \mathbb{N}$, $\sum_{s=1}^{m/\rho} \|\Phi^*_{m/\rho} \Delta^r v_s\|_2 \leq r 2^r + \eta(2\pi \max_{1 \leq j \leq k} |\lambda_j| \frac{1}{\eta})^r$, *where* $(v_s)_j = \delta(s-j)$, *the* s*-th canonical coordinate.*

Proof. Note that for any $A = (A_1, \ldots, A_n)$, $A\Delta e_s = A_s - A_{s+1}$. Thus,

$$\begin{aligned}
\sum_{s<m/\rho-r} \|\Phi^*_{m/\rho}\Delta^r v_s\|_2 &= \sum_{s<m/\rho-r} \|\sum_{t=s}^{s+r}(-1)^t\binom{r}{t-s}U_{t\rho/m}\phi_0\|_2 \\
&= \sum_{s<m/\rho-r} \|U_{s\rho/m}\sum_{t=0}^{r}(-1)^t\binom{r}{t}U_{t\rho/m}\phi_0\|_2 \\
&= \sum_{s<m/\rho-r} \|\sum_{t=0}^{r}(-1)^t\binom{r}{t}U_{t\rho/m}\phi_0\|_2 \\
&= \sum_{s<m/\rho-r} \|\sum_{j=1}^{k}c_j[\sum_{t=0}^{r}(-1)^t\binom{r}{t}e^{2\pi\imath t\lambda_j\rho/m}]v_j\|_2 \\
&= \sum_{s<m/\rho-r} (\sum_{j=1}^{k}|c_j|^2|e^{2\pi\imath\lambda_j\rho/m}-1|^{2r})^{1/2} \\
&\leq \sum_{s<m/\rho-r} \left(\sum_{j=1}^{k}|c_j|^2\cdot\left(2\pi|\lambda_j|\frac{\rho}{m}\right)^{2r}\right)^{1/2} \\
&\leq \sum_{s<m/\rho-r} \left(\max_{1\leq j\leq k}2\pi|\lambda_j|\frac{1}{\eta}\right)^r\cdot\|\phi_0\|_2 \\
&\leq \eta(2\pi\max_{1\leq j\leq k}|\lambda_j|\frac{1}{\eta})^r\cdot\|\phi_0\|_2,
\end{aligned}$$

where we note that $m/\rho = \eta$.

For $s \geq m/\rho - r$, with trivial estimates one has

$$\sum_{s\geq m/\rho-r}\|\Phi^*_{m/\rho}\Delta^r e_s\|_2 \leq r\|\sum_{j=1}^{k}|c_j|\sum_{t=0}^{r}\binom{r}{t}v_j\| \leq r2^r.$$

□

Proposition 2.3.29 ("Frame Variation" Estimate)**.**

$$\|\Phi^*_{m/\rho,k}D_\rho S^2_\rho\Delta^2 u\|_2 \leq \left(9+\eta\left(2\pi\max|\lambda_j|\frac{1}{\eta}\right)^2\right)\frac{1}{\rho^2}.$$

Proof. Let $\{v_s\}_s$ be the canonical basis of $\mathbb{C}^\eta$. Then, by Proposition 2.3.27, we have

$$\begin{aligned}
\|\Phi^*_{m/\rho,k} D_\rho S_\rho^2 \Delta^2 u\|_2 &= \frac{1}{\rho^2} \|\Phi^*_{m/\rho,k} \Delta^2 D_\rho u + V u\|_2 \\
&\leq \frac{1}{\rho^2} \left(\sum_{s=1}^{m/\rho} \|\Phi^*_{m/\rho} \Delta^2 v_s\|_2 + \|\phi_1\|_2 \right) \|u\|_\infty \\
&\leq \frac{1}{\rho^2} \left(8 + \eta \left(2\pi \max |\lambda_j| \frac{1}{\eta} \right)^2 + 1 \right) \\
&= \left(9 + \eta \left(2\pi \max |\lambda_j| \frac{1}{\eta} \right)^2 \right) \frac{1}{\rho^2}.
\end{aligned}$$

□

Lemma 2.3.30 (Total Number of Bits Used)**.** *Given a mid-rise quantizer $\mathscr{A} = \mathscr{A}_0 + \imath\mathscr{A}_0$ with length $2L$ and $r \in \mathbb{N}$, if $q \in \mathscr{A}^m$ is a quantized sample from the alphabet, then $D_\rho S_\rho^r q \in \mathbb{C}^\eta$ can be encoded by $\eta \cdot 2r \log(2Lm)$ bits.*

Proof. Given the assumption above, each entry q_j of q is a number of the form

$$q_j = \big((2s_j + 1) + \imath(2t_j + 1)\big) \frac{\delta}{2}, \quad -L \leq s_j,\, t_j \leq L - 1.$$

Then, each entry in $S_\rho q$ is the average of ρ entries in q, which has the form

$$(S_\rho q)_j = \big((2\tilde{s}_j + \rho) + \imath(2\tilde{t}_j + \rho)\big) \frac{\delta}{2\rho}, \quad -Lm \leq \tilde{s}_j,\, \tilde{t}_j \leq (L-1)m.$$

There are at most $((2L-1)m + 1)^2 \leq (2Lm)^2$ choices per entry. Note that there are $(2Lm)^2$ choices instead of $(2L\rho)^2$ as we need to account for the first $\rho - 1$ rows, which sums $m - \rho$ terms. Iterating r times, there are $(2Lm)^{2r}$ choices for each entry

of $S_\rho^r q$. Thus, the vector $D_\rho S_\rho^r q$ can be encoded by $\mathscr{R} = \eta \cdot 2r \log(2Lm)$ bits. $\square$

Proof. of Theorem 2.3.8:

To estimate the reconstruction error, we note that

$$D_\rho S_\rho^2 \Phi_{m,k} = \Phi_{m/\rho,k} \bar{C}_{m,\rho}^2,$$

which follows from Lemma 2.3.23. Moreover, $(D_\rho S_\rho^2 \Phi_{m,k})^\dagger = \bar{C}_{m,\rho}^{-2} \mathcal{S}^{-1} \Phi_{m/\rho,k}^*$, where $\mathcal{S} = \Phi_{m/\rho,k}^* \Phi_{m/\rho,k}$ has lower frame bound $\frac{m}{\rho} C_{\phi_0}$. Since $\|\bar{C}_{m,\rho}^{-1}\|_2 \leq \frac{\pi}{2}$, the reconstruction error is

$$\begin{aligned}
\mathscr{E}_{m,\rho} &= \|x - \bar{C}_{m,\rho} \mathcal{S}^{-1} \Phi_{m/\rho,k}^* q\|_2 \\
&= \|\bar{C}_{m,\rho}^{-2} \mathcal{S}^{-1} \Phi_{m/\rho,k}^* D_\rho S_\rho^2 (\Delta^{(m)})^2 u\|_2 \\
&\leq \frac{1}{\rho^2} \|\bar{C}^{-1}\|_2^2 \|\mathcal{S}^{-1}\|_2 \|\Phi_{m/\rho}^* D_\rho S_\rho^2 \Delta^2 u\|_2 \\
&\leq \frac{\pi^2}{4\eta C_{\phi_0}} \left(9 + \eta (2\pi \max_{1 \leq j \leq k} |\lambda_j| \frac{1}{\eta})^2\right) \|u\|_\infty \frac{1}{\rho^2},
\end{aligned}$$

where $\{v_j\}_j \subset \mathbb{C}^m$ denotes the canonical basis in $\mathbb{C}^m$, the first inequality comes from Proposition 2.3.29, and the second follows from Lemma 2.3.28. Here, we see that the error decays quadratically with respect to the oversampling rate ρ.

As for the bits used, note that $\frac{1}{m} = \frac{1}{\eta} \cdot \frac{1}{\rho}$ and

$$e^{-\frac{1}{2\eta}\mathscr{R}} = \frac{1}{(2Lm)^2} = \frac{1}{(2L\eta)^2} \frac{1}{\rho^2},$$

where $\mathscr{R} = \eta \cdot 4\log(2Lm)$ comes from Lemma 2.3.30. Thus, we have

$$\mathscr{E}(\mathscr{R}) \leq \frac{\pi^2}{4\eta C_{\phi_0}}\left(9 + \eta(2\pi \max_{1\leq j\leq k} |\lambda_j|\frac{1}{\eta})^2\right)\|u\|_\infty \frac{1}{\rho^2} \leq C_{k,\phi_0,L,\eta}\|u\|_\infty 2^{-\frac{1}{2\eta}\mathscr{R}},$$

where $C_{k,\phi_0,L,\eta} \leq \frac{\pi^2}{4\eta C_{\phi_0}}\left(9 + \eta(2\pi \max_{1\leq j\leq k} |\lambda_j|\frac{1}{\eta})^2\right)(2L\eta)^2$, independent of ρ.

□

Lemma 2.3.26 shows that Δ^{-1} and $\bar{\Delta}_\rho$ do not commute, and such non-commutativity limits the potential to generalize alternative decimation to higher orders. For the sake of demonstration, we show explicit calculation in Section 2.3.5 which highlights the difficulty in the generalization of our results. Thus, to achieve exponential error decay with respect to the bit usage for higher order $\Sigma\Delta$ quantization schemes, we need to employ different approaches. The new scheme we propose will be introduced in Section 2.4.

2.3.5 Limitation of Alternative Decimation: Third Order Decimation

The non-commutativity between $\bar{\Delta}_\rho$ and Δ^{-1} results in incomplete difference scaling when applying $D_\rho S_\rho^r$ on Δ^r, creating substantial error terms. This phenomenon already occurs for $r = 3$.

Proposition 2.3.31. *Given $m, \rho \in \mathbb{N}$ with $\rho \mid m$, the third order decimation satisfies $D_\rho S_\rho^3 \Delta^3 = \frac{1}{\rho^3}(\Delta^{(\eta)})^3 D_\rho + O(\rho^{-2})$. In particular, $D_\rho S_\rho^3$ only yields quadratic error decay with respect to the oversampling ratio ρ.*

First, by noting that $\Delta^{-1}\bar{\Delta}_\rho\Delta = \mathcal{E}$ as in Lemma 2.3.26, one has

$$
\begin{aligned}
D_\rho S_\rho^3 \Delta^3 &= \frac{1}{\rho^3} D_\rho \bar{\Delta}_\rho \Delta^{-1} \bar{\Delta}_\rho \Delta^{-1} \bar{\Delta}_\rho \Delta^2 \\
&= \frac{1}{\rho^3} D_\rho \bar{\Delta}_\rho (\Delta^{-1} \bar{\Delta}_\rho \Delta) \Delta^{-2} \bar{\Delta}_\rho \Delta^2 \\
&= \frac{1}{\rho^3} D_\rho \bar{\Delta}_\rho (\bar{\Delta}_\rho + \mathcal{E}) \Delta^{-1} (\bar{\Delta}_\rho + \mathcal{E}) \Delta \\
&= \frac{1}{\rho^3} D_\rho \bar{\Delta}_\rho (\bar{\Delta}_\rho + \mathcal{E}) (\bar{\Delta}_\rho + \mathcal{E} + \Delta^{-1} \mathcal{E} \Delta) \\
&= \frac{1}{\rho^3} D_\rho \bigg(\bar{\Delta}_\rho^3 + \bar{\Delta}_\rho^2 \mathcal{E} + \bar{\Delta}_\rho^2 (\Delta^{-1} \mathcal{E} \Delta) + \bar{\Delta}_\rho \mathcal{E} \bar{\Delta}_\rho + \bar{\Delta}_\rho \mathcal{E}^2 + \bar{\Delta}_\rho \mathcal{E} (\Delta^{-1} \mathcal{E} \Delta) \bigg).
\end{aligned}
\tag{2.15}
$$

We shall calculate all terms one-by-one.

Lemma 2.3.32. *We have the following equalities:*

(1) :

$$
(D_\rho \bar{\Delta}_\rho^2 \mathcal{E})_{l,s} = \delta(s - (m - \rho)) \bigg(\delta(l-1) - \delta(l-2) \bigg),
$$

(2) :

$$
(D_\rho \bar{\Delta}_\rho^2 (\Delta^{-1} \mathcal{E} \Delta))_{l,s} = \begin{cases} -\rho & \text{if} \quad (l,s) = (1, m - \rho - 1) \\ \rho & \text{if} \quad (l,s) = (1, m - \rho) \\ 0 & \text{otherwise} \end{cases},
$$

(3) :

$$
(D_\rho \bar{\Delta}_\rho \mathcal{E} \bar{\Delta}_\rho)_{l,s} = \delta(l-1) \bigg(\delta(s - (m - \rho)) - \delta(s - (m - 2\rho)) \bigg),
$$

(4) :

$$
(D_\rho \bar{\Delta}_\rho \mathcal{E}^2)_{l,s} = \delta(l-1) \delta(s - (m - \rho)),
$$

(5) :

$$(D_\rho \bar{\Delta}_\rho \mathcal{E}(\Delta^{-1}\mathcal{E}\Delta))_{l,s} = (m-\rho)\delta(l-1)\Big(\delta(s-(m-\rho)) - \delta(s-(m-\rho-1)\Big),$$

where given $n \in \mathbb{N}$, $[n] := \{1, \dots, n\}$. *In particular,* $D_\rho\big(\bar{\Delta}_\rho^2(\Delta^{-1}\mathcal{E}\Delta) + \bar{\Delta}_\rho \mathcal{E}(\Delta^{-1}\mathcal{E}\Delta)\big) = O(m)$, *and* $D_\rho(\bar{\Delta}_\rho^2\mathcal{E} + \bar{\Delta}_\rho\mathcal{E}\bar{\Delta}_\rho + \bar{\Delta}_\rho\mathcal{E}^2) = O(1)$.

Proof. We will first compute each term without the effect of D_ρ since D_ρ is the sub-sampling matrix retaining only the $t\rho$-th rows for $t \in [\eta]$.

(1), (3) First, note that $(\bar{\Delta}_\rho \mathcal{E})_{l,s} = \delta(l-\rho)\delta(s+\rho)$, so

$$(\bar{\Delta}_\rho^2 \mathcal{E})_{l,s} = \delta(s+\rho)(\bar{\Delta}_\rho)_{l,\rho} = \delta(s-(m-\rho))(\delta(l-\rho) - \delta(l-2\rho)).$$

Similarly,

$$(\bar{\Delta}_\rho \mathcal{E} \bar{\Delta}_\rho)_{l,s} = \delta(l-\rho)(\bar{\Delta}_\rho)_{m-\rho,s} = \delta(l-\rho)(\delta(s-(m-\rho)) - \delta(s-(m-2\rho))).$$

(5) Now, to compute $\Delta^{-1}\mathcal{E}\Delta$, we see that, for $s \neq m$,

$$(\Delta^{-1}\mathcal{E}\Delta)_{l,s} = \sum_{j=1}^{l}(\mathcal{E}_{j,s} - \mathcal{E}_{j,s+1}) = l(\delta(m-\rho-s) - \delta(m-\rho-(s+1))),$$

and $(\Delta^{-1}\mathcal{E}\Delta)_{l,m} = 0$. In particular,

$$\Delta^{-1}\mathcal{E}\Delta = \begin{pmatrix} 0 & \dots & 0 & -1 & 1 & 0 & \dots & 0 \\ \vdots & & \vdots & -2 & 2 & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & -m & m & 0 & \dots & 0 \end{pmatrix},$$

where the nonzero columns occur at the $(m-\rho-1)$ and $(m-\rho)$-th positions.

For $\bar{\Delta}_\rho\mathcal{E}(\Delta^{-1}\mathcal{E}\Delta)$,

$$\begin{aligned}(\bar{\Delta}_\rho\mathcal{E}(\Delta^{-1}\mathcal{E}\Delta))_{l,s} &= \delta(l-\rho)(\Delta^{-1}\mathcal{E}\Delta)_{m-\rho,s} \\ &= \delta(l-\rho)(m-\rho)(\delta(s-(m-\rho)) - \delta(s-(m-\rho-1))).\end{aligned}$$

(4) Note that $\bar{\Delta}_\rho\mathcal{E}^2 = \bar{\Delta}_\rho\mathcal{E}$. The result then follows from the calculation on the first term.

(2) Finally, as $\Delta^{-1}\mathcal{E}\Delta$ only has non-zero entries on the $(m-\rho-1)$ and $(m-\rho)$-th columns, and the two columns differ by a sign, it suffices to calculate the $(m-\rho)$-th column of $\bar{\Delta}_\rho^2(\Delta^{-1}\mathcal{E}\Delta)$.

$$\begin{aligned}(\bar{\Delta}_\rho(\Delta^{-1}\mathcal{E}\Delta))_{l,m-\rho} &= \sum_{j=1}^{m} j(\bar{\Delta}_\rho)_{l,j} \\ &= \begin{cases} l-(l-\rho) = \rho & \text{if} \quad l > \rho \\ l-(l-\rho+m) = -(m-\rho) & \text{if} \quad l < \rho \\ l = \rho & \text{if} \quad l = \rho. \end{cases}\end{aligned}$$

Then,

$$(\bar{\Delta}_\rho^2(\Delta^{-1}\mathcal{E}\Delta))_{l,m-\rho} = \sum_{j=1}^{m}(\bar{\Delta}_\rho)_{l,j}(\bar{\Delta}_\rho(\Delta^{-1}\mathcal{E}\Delta))_{j,m-\rho}$$
$$= \begin{cases} -m & \text{if} \quad l \in [2\rho-1]\backslash\{\rho\} \\ \rho & \text{if} \quad l = \rho \\ 0 & \text{otherwise.} \end{cases}$$

□

Proof. of Proposition 2.3.31:

From (2.15) and Lemma 2.3.32, we see that

$$D_\rho S_\rho^3 \Delta^3 = \frac{1}{\rho^3} D_\rho \bar{\Delta}_\rho^3 + \frac{\eta}{\rho^2}\mathcal{E}_1 + \frac{1}{\rho^3}\mathcal{E}_2 = \frac{1}{\rho^3} D_\rho \bar{\Delta}_\rho^3 + O(\rho^{-2}),$$

where

$$(\mathcal{E}_1)_{l,s} = \frac{1}{m}\left(D_\rho(\bar{\Delta}_\rho^2(\Delta^{-1}\mathcal{E}\Delta)+\bar{\Delta}_\rho\mathcal{E}(\Delta^{-1}\mathcal{E}\Delta))\right)_{l,s} = \begin{cases} -1 & \text{if} \quad (l,s) = (1, m-\rho-1) \\ 1 & \text{if} \quad (l,s) = (1, m-\rho) \\ 0 & \text{otherwise,} \end{cases}$$

and

$$(\mathcal{E}_2)_{l,s} = \left(D_\rho(\bar{\Delta}_\rho^2\mathcal{E}+\bar{\Delta}_\rho\mathcal{E}\bar{\Delta}_\rho+\bar{\Delta}_\rho\mathcal{E}^2)\right)_{l,s} = \begin{cases} -1 & \text{if} \quad (l,s) = (2, m-\rho) \text{ or } (1, m-2\rho) \\ 3 & \text{if} \quad (l,s) = (1, m-\rho) \\ 0 & \text{otherwise.} \end{cases}$$

□

Even in higher order cases, alternative decimation still only yields quadratic error decay with respect to the oversampling ratio, as can be seen in Figure 2.2d and 2.2e.

Alternative decimation is limited by this incomplete cancellation, but canonical decimation has even worse error decay. Contrary to the quadratic decay for alternative decimation, canonical decimation only has linear decay for high order $\Sigma\Delta$ quantization. The same thing applies to plain $\Sigma\Delta$ quantization, as can be seen in Figure 2.2b.

2.3.6 Comparison Between Alternative and Canonical Decimation

Here, we present numerical evidence that the alternative decimation on frames has linear and quadratic error decay rate for the first and the second order, respectively. Moreover, it is shown that the canonical decimation, as described in Remark 2.3.2, is not suitable for our purpose when $r \geq 2$.

Recall that given m, r, ρ, one can define the canonical decimation operator $D_\rho \tilde{S}_\rho^r \in \mathbb{R}^{\eta \times m}$, where $\tilde{S}_\rho \in \mathbb{R}^{m \times m}$ is a circulant matrix.

In our experiment, we look at three different quantization schemes: alternative decimation, canonical decimation, and plain $\Sigma\Delta$. Given observed data $y \in \mathbb{C}^m$ from a frame $E \in \mathbb{C}^{m \times k}$ and $r \in \mathbb{N}$, one can determine the quantized samples $q \in \mathbb{C}^m$ by

$$y - q = \Delta^r u$$

for some bounded u. The three schemes differ in the choice of dual frames:

- Alternative decimation: $\tilde{x} = (D_\rho S_\rho^r E)^\dagger D_\rho S_\rho^r q = F_a q$.
- Canonical decimation: $\tilde{x} = (D_\rho \tilde{S}_\rho^r E)^\dagger D_\rho \tilde{S}_\rho^r q = F_c q$.
- Plain $\Sigma\Delta$: $\tilde{x} = E^\dagger q = F_p q$.

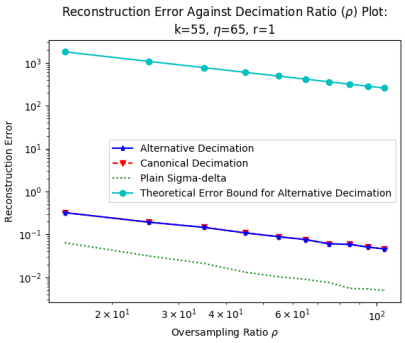
For each experiment, we use the mid-rise quantizer $\mathscr{A}$ and fix $k = 55, \delta = 0.5, L = 100$, and $\eta = 65$. For each ρ, we set $m = \rho\eta$ and pick 10 randomly generated vectors $\{x^j\}_{j=1}^{10} \subset \mathbb{C}^k$. $\Sigma\Delta$ quantization on each signal gives $\{q^j\}_{j=1}^{10} \subset \mathbb{C}^m$. The maximum reconstruction error over the 10 experiments is recorded, namely

$$\mathscr{E}_i = \max_{1 \le j \le 10} \|x^j - F_i q^j\|_2, \quad i \in \{a, c, p\}.$$

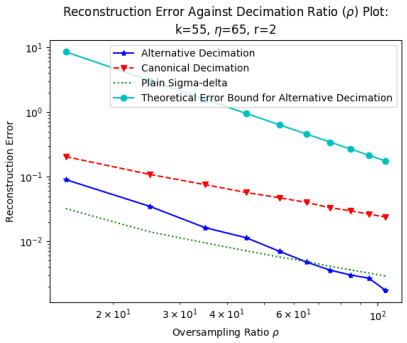
The frame in our experiment is

$$(E^{m,k})_{l,j} = (E)_{l,j} = \frac{1}{\sqrt{k}} (\exp(-2\pi\imath(l+1)(j+1)/m))_{l,j}.$$

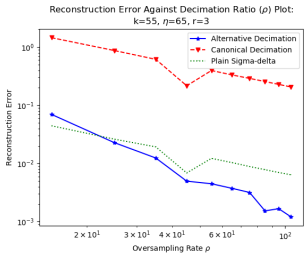
First, we shall compare alternative decimation with plain $\Sigma\Delta$ quantization from Figure 2.2. For $r = 1$, alternative decimation performs worse than plain $\Sigma\Delta$ quantization, as plain $\Sigma\Delta$ quantization benefits from the smoothness of the frame elements, having decay rate $O((\frac{m}{k})^{-5/4})$ proven in [6]. However, for $r \ge 2$, alternative decimation supersedes plain $\Sigma\Delta$ quantization as the better scheme. This can be explained by the boundary effect in finite-dimensional spaces that results in incomplete cancellation for backward difference matrices. We are interested in the



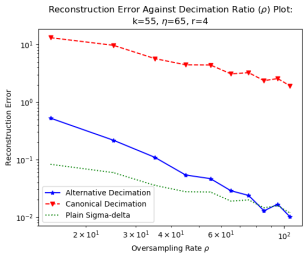
(a) $r = 1$.



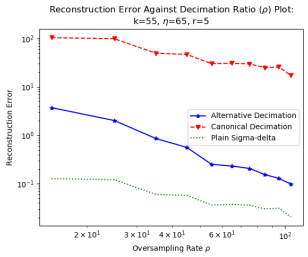
(b) $r = 2$.



(c) $r = 3$.



(d) $r = 4$.



(e) $r = 5$.

Figure 2.2: The log-log plot for reconstruction error against the decimation ratio ρ for different quantization schemes. In the case $r = 1$, alternative decimation coincides with canonical decimation. For $r \geq 2$, alternative decimation has better error decay rate than both canonical decimation and plain $\Sigma\Delta$ quantization.

case $r = 1$ or 2. As we can see, the theoretical error bound does not have a tight constant, although the decay rate is consistent with our experimental result.

2.3.6.1 Necessity of Alternative Decimation

The main difference between the alternative decimation operator $D_\rho S_\rho^r$ and the canonical one $D_\rho \tilde{S}_\rho^r$ lies in the scaling effect on difference structures. We have $\tilde{S}_\rho^r = (S_\rho + L)^r$ with ρL having unit entries on the first $\rho - 1$ rows and 0 everywhere else.

In Figure 2.2, we can see the performance drop-off when switching from alternative decimation to canonical decimation for $r \geq 2$. we can see that canonical decimation incurs much worse reconstruction error than the alternative one, while generally having worse decay rate. For demonstration, we show explicitly the difference between alternative and canonical decimation schemes for $r = 2$:

$$\begin{aligned}
\tilde{S}_\rho^2 \Delta^2 &= (S_\rho + L)^2 \Delta^2 \\
&= S_\rho^2 \Delta^2 + (L S_\rho + S_\rho L + L^2)\Delta^2 \\
&= S_\rho^2 \Delta^2 + L(S_\rho + L^2)\Delta^2 + S_\rho L \Delta^2.
\end{aligned}$$

Since $D_\rho L = 0$, we are left with $D_\rho S_\rho L \Delta^2$. Now,

$$(L\Delta^2)_{l,j} = \begin{cases} \frac{-1}{\rho} & \text{if} \quad 1 \leq l \leq \rho - 1,\ j = m - 1, \\ \frac{1}{\rho} & \text{if} \quad 1 \leq l \leq \rho - 1,\ j = m, \\ 0 & \text{otherwise.} \end{cases}$$

Then, we see that

$$(DS_\rho L\Delta^2)_{l,j} = \begin{cases} \frac{-(\rho-1)}{\rho^2} & \text{if} \quad l=1,\, j=m-1, \\ \frac{\rho-1}{\rho^2} & \text{if} \quad l=1,\, j=m, \\ 0 & \text{otherwise.} \end{cases}$$

We see that $D_\rho \tilde{S}_\rho^2 \Delta^2 = O(\rho^{-1})$, hence the linear decay for $r = 2$.

2.4 Adapted Decimation

In Theorem 2.1.4, we see that signal decimation coupled with the r-th $\Sigma\Delta$ quantization scheme in A/D conversion yields polynomial error decay rate of degree r with respect to the oversampling ratio. Moreover, it yields exponential error decay rate the bit-rate. The question we seek to address is whether it is possible to translate decimation from A/D conversion to finite frame quantization. This adaptation proves to be non-trivial, as the r-th order $\Sigma\Delta$ quantization does not yield much more than linear error decay rate for finite frames in general as opposed to polynomial degree r, [6, 47].

With the introduction of *alternative decimation*, we were able to adapt signal decimation to finite frames up to the second order $\Sigma\Delta$ quantization, yielding quadratic error decay rate with respect to the oversampling ratio. Here, we further generalize the concept of decimation and extends the decimation on finite frames to arbitrary polynomial degrees.

We have seen in Theorem 2.3.8 that alternative decimation is only useful up

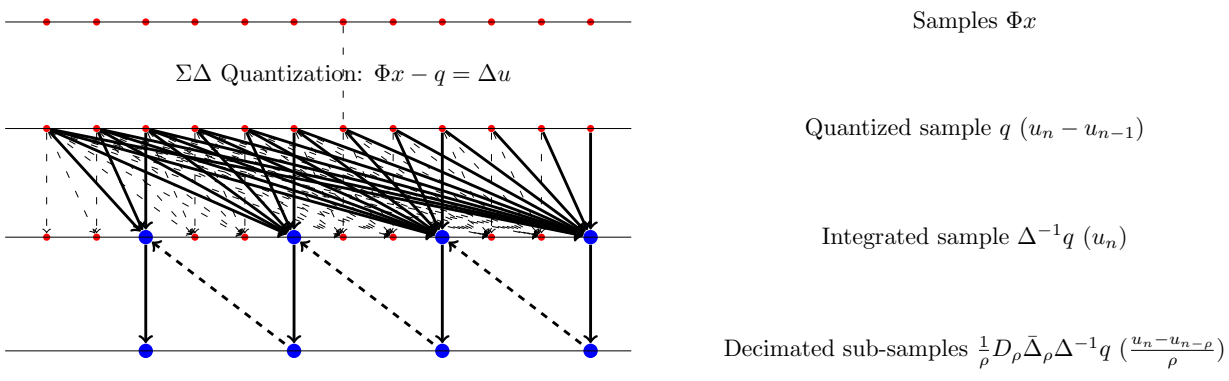


Figure 2.3: Illustration of the first order adapted (alternative) decimation scheme for finite frames. After obtaining the quantized samples $\{q_n\}_n$ in the first step, one starts by integrating quantized samples in the second step. Finite difference of step size ρ followed by sub-sampling are then taken in the third step. The effect on the reconstruction (replacing q_n with $y_n - q_n$) is illustrated in parentheses. Note that both the recursivity and the boundary effect (see bottom left) can be seen in this diagram.

to the second order. Thus, we aim to extend our results to arbitrary orders, and the solution we present here is called *adapted decimation*.

Definition 2.4.1 (Adapted Decimation)**.** Given $r, m, \rho \in \mathbb{N}$, the (r, m, ρ)-adapted decimation operator is defined to be

$$A_r = \frac{1}{\rho^r} D_\rho \bar{\Delta}_\rho^r \Delta^{-r},$$

where $\Delta \in \mathbb{N}^{m\times m}$ is the usual backward difference matrix, $\bar{\Delta}_\rho \in \mathbb{R}^{m\times m}$ satisfies $(\bar{\Delta}_\rho)_{l,s} = \frac{1}{\rho}(\delta(l-s) - \delta(l+\rho-s) + \delta(s-m)\delta(l-\rho))$, and $D_\rho \in \mathbb{N}^{m/\rho\times m}$ has $(D_\rho)_{l,s} = \delta(s - l\rho)$.

Remark 2.4.2 (Comparison between Alternative and Adapted Decimation)**.** While coinciding for $r = 1$, A_r is different from $D_\rho S_\rho^r$ in the following way: $S_\rho = \frac{1}{\rho}\bar{\Delta}_\rho\Delta^{-1}$,

and thus

$$D_\rho S_\rho^2 = \frac{1}{\rho^2} D_\rho (\bar{\Delta}_\rho \Delta^{-1})^2 \neq \frac{1}{\rho^2} D_\rho \bar{\Delta}_\rho^2 \Delta^{-2} = A_2.$$

The non-commutativity between $\bar{\Delta}_\rho$ and Δ^{-1} limits the success of the alternative decimation, see Proposition A.1 in [47]. Adapted decimation essentially factorizes the alternative decimation and re-arranges the terms. In doing so, the reconstruction error rate can now be of polynomial degree r. However, it also complicates the effect of decimation on finite frames, as will be seen in Section 2.4.2. For the illustration, see Figure 2.3.

It will be shown that, for unitarily generated frames $\Phi \in \mathbb{C}^{m\times k}$ satisfying conditions specified in Theorem 2.4.3 and *any* $r \in \mathbb{N}$, an r-th order $\Sigma\Delta$ quantization coupled with the corresponding adapted decimation has r-th order polynomial reconstruction error decay rate with respect to the ratio ρ. As for the data storage, decimation allows for highly efficient storage, making the error decay exponentially with respect to the bit usage.

Theorem 2.4.3. *Given* Ω, ϕ_0, $\{\lambda_j\}_j$, $\{v_j\}_j$, *and* $\Phi = \Phi_{m,k}$ *as the generator, base vector, eigenvalues, eigenvectors, and the corresponding UGF, respectively, and* $r \in \mathbb{N}$ *fixed. Suppose*

- $\rho \mid m$,
- $\eta = m/\rho \geq 3rk$,
- $\{\lambda_j\}_{j=1}^k \subset [-\eta/2, \eta/2] \cap \mathbb{Z}\backslash\{0\}$, *and*
- $C_{\phi_0} = \min_s |<\phi_0, v_s>|^2 > 0$,

then the following statements are true:

(a) **Recursivity:** *For all* $s \in \{1, \ldots, \eta\}$, *there exists* $\{c_j^s\}_{j=1}^{s\rho}$ *such that* $(A_r q)_s = \sum_{j=1}^{s\rho} c_j^s q_j$.

(b) **Signal reconstruction:** $A_r \Phi_{m,k}$ *is a frame.*

(c) **Error estimate:** *Given the dual frame* $F = (A_r \Phi_{m,k})^\dagger A_r$, *where for any* M, $M^\dagger = (M^* M)^{-1} M^*$ *is defined to be the pseudo-inverse of* M. *Then the reconstruction error* $\mathscr{E}_{m,\rho} = \|x - Fq\|_2$ *satisfies*

$$\mathscr{E}_{m,\rho} \leq \left(\frac{4}{k\eta C_{\phi_0}} (\pi^2 \eta)^r \right) \|u\|_\infty \frac{1}{\rho^r}. \tag{2.16}$$

(d) **Efficient data storage:** *Suppose the length of the quantization alphabet is* $2L$, *then the total bits used to record the quantized samples* $A_r q$ *are* $\mathscr{R} = 2\eta r \log(2m) + 2\eta \log(2L)$ *bits. Furthermore, as a function of bits used at each entry,* $\mathscr{E}_{m,\rho}$ *satisfies*

$$\mathscr{E}(\mathscr{R}) \leq C_{k,\eta,\phi_0,L} \|u\|_\infty 2^{-\frac{1}{2\eta}\mathscr{R}}, \tag{2.17}$$

where $C_{k,\eta,\phi_0,L} = \frac{8L}{k\eta C_{\phi_0}} (2\pi^2)^r$, *independent of* ρ.

We shall prove Theorem 2.3.5 in several steps. First, we split $A_r \Phi_{m,k}$ into one main term and many residual terms in Section 2.4.2.1 . Then, we compute the cancellation among residual terms in Section 2.4.2.2. We compute the lower frame bound of $A_r \Phi_{m,k}$ in Section 2.4.3 before proving the theorem itself in Section 2.4.6.

2.4.1 Roadmap of the Proof

In this subsection, we shall identify the key components regarding the proof of Theorem 2.4.3. Then, we will provide estimates for those components in Sections 4.2-4.5 before finishing the proof in Section 2.4.6.

To estimate the reconstruction error $\mathscr{E}_{m,\rho} = \|x - (A_r\Phi_{m,k})^\dagger A_r q\|_2$ in (2.16), we re-write the form of A_r, making the estimate simpler. In particular, we claim that $\bar{\Delta}_\rho$ scales down to the usual backward-difference matrix under the under-sampling matrix D_ρ:

Lemma 2.4.4. *Given $m, \rho \in \mathbb{N}$ with $\eta = m/\rho \in \mathbb{N}$,*

$$D_\rho \bar{\Delta}_\rho = \Delta^{(\eta)} D_\rho,$$

where $\Delta^{(\eta)}$ is the η-dimensional backward difference matrix.

Proof. Note that, for $s \neq m$,

$$\begin{aligned}(D_\rho\bar{\Delta}_\rho)_{l,s} &= (\bar{\Delta}_\rho)_{l\rho,s} \\ &= \delta(s - l\rho) - \delta(s + \rho - l\rho) = \delta(s - l\rho) - \delta(s - (l-1)\rho) \\ &= (\Delta D_\rho)_{l,s}.\end{aligned}$$

For $s = m$, $(D_\rho\bar{\Delta}_\rho)_{l,m} = \delta(m - l\rho) = (\Delta D_\rho)_{l,m}$. □

Then, the reconstruction error $\mathscr{E}_{m,\rho}$ satisfies

$$
\begin{aligned}
\mathscr{E}_{m,\rho} &= \|x - (A_r\Phi_{m,k})^\dagger A_r q\|_2 \\
&= \|(A_r\Phi_{m,k})^\dagger A_r(\Phi_{m,k}x - q)\|_2 \\
&= \|(A_r\Phi_{m,k})^\dagger \frac{1}{\rho^r} D_\rho \bar{\Delta}^r \Delta^{-r}(\Delta^r u)\|_2 \\
&= \frac{1}{\rho^r}\|\big((A_r\Phi_{m,k})^* A_r\Phi_{m,k}\big)^{-1}(A_r\Phi_{m,k})^*\Delta^r D_\rho u\|_2 \\
&\leq \|\big((A_r\Phi_{m,k})^* A_r\Phi_{m,k}\big)^{-1}\|_2 \cdot \|(A_r\Phi_{m,k})^*\Delta^r\|_{\infty,2} \cdot \|u\|_\infty \frac{1}{\rho^r},
\end{aligned}
\tag{2.18}
$$

where the fourth equality follows from Lemma 2.4.4. We have seen from Remark 2.1.5 that $\|((A_r\Phi_{m,k})^* A_r\Phi_{m,k})^{-1}\|_2$ is the reciprocal of the lower frame bound of $A_r\Phi_{m,k}$. Thus, in order to estimate (2.18), we need only to answer two questions:

- Is $A_r\Phi_{m,k}$ a frame? What is the lower frame bound of $A_r\Phi_{m,k}$?
- What is $\|(A_r\Phi_{m,k})^*\Delta^r\|_{\infty,2}$?

The lower frame bound of $A_r\Phi_{m,k}$ will be calculated in Section 2.4.3, specifically in Proposition 2.4.15. As for the estimate in the second question, it is given in Proposition 2.4.16 of Section 2.4.4.

Aside from the reconstruction error estimate, we also need to calculate the number of bits needed to record the decimated sample $A_r q$. We shall show that $A_r q$ can be efficiently stored in $O(\log \rho)$ instead of $O(\rho)$ bits. The explicit estimate will be done in Proposition 2.4.17 of Section 2.4.5.

2.4.2 Expansion of $A_r\Phi_{m,k}$

In [47], one has, for any $r \in \mathbb{N}$, the alternative decimation satisifes

$$D_\rho S_\rho^r \Phi_{m,k} = \frac{1}{\rho^r} D_\rho (\bar{\Delta}_\rho \Delta^{-1})^r \Phi_{m,k} = \Phi_{\eta,k} (\tilde{D}\tilde{C})^r$$

where $\tilde{D}, \tilde{C} \in \mathbb{C}^k$ will be defined in Section 2.4.2.1. The form is rather simple thanks to the alternating applications of $\bar{\Delta}_\rho$ and Δ^{-1}. For adapted decimation, we have $A_r = \frac{1}{\rho^r} D_\rho \bar{\Delta}_\rho^r \Delta^{-r}$, and the displaced order of applications creates residual terms other than $\Phi_{\eta,k}(\tilde{D}\tilde{C})^r$. In this section, we observe this phenomenon and examine the effect of the residual terms.

2.4.2.1 The Effect of Adapted Decimation on the Frame

We start by introducing the following notation:

Definition 2.4.5. Given $l, s \in \mathbb{N}$, the l-by-s constant matrix $1_{l,s}$ has constant 1 on all entries.

The following two lemmas are needed for us to describe $A_r\Phi_{m,k}$ in Proposition 2.4.8.

Lemma 2.4.6. *Given $\Phi = \Phi_{m,k} \in \mathbb{C}^{m\times k}$ with base vector ϕ_0, we have*

$$\Delta^{-1}\Phi = (\Phi - 1_{m,k}V)\tilde{C},$$

where $\tilde{C}$ and U_t are simultaneously diagonalizable with $B^*\tilde{C}B = \tilde{C}_0 = diag(\frac{1}{1-e^{2\pi\imath\lambda_s/m}})_{1\le s\le m}$

and $V = diag(\phi_0)$.

Proof. For any $1 \leq t \leq m$, the t-th row of $\Delta^{-1}\Phi_{m,k}$ can be written as

$$\begin{aligned}(\Delta^{-1}\Phi_{m,k})_t &= (\sum_{s=1}^{t} U_{s/m}\phi_0)^* \\ &= (\sum_{s=1}^{t} BT_{s/m}B^*\phi_0)^* \\ &= (B\sum_{s=1}^{t} T_{s/m}B^*\phi_0)^*,\end{aligned}$$

where we note that $U_t = BT_tB^*$ can be diagonalized by the unitary matrix $B = B_\Phi$, and $T_t = e^{2\pi\imath\Lambda t} = diag(\exp(2\pi\imath\lambda_s t))_s$. Now,

$$\begin{aligned}\sum_{s=1}^{t}(T_{s/m})_\sigma &= \sum_{s=1}^{t} e^{2\pi\imath\lambda_\sigma s/m} \\ &= \frac{e^{2\pi\imath\lambda_\sigma t/m} - 1}{e^{2\pi\imath\lambda_\sigma/m} - 1} \\ &= (T_{t/m})_\sigma \frac{1}{e^{2\pi\imath\lambda_\sigma/m} - 1} - \frac{1}{e^{2\pi\imath\lambda_\sigma/m} - 1} \\ &= (\tilde{C}_0 T_{t/m} - \tilde{C}_0)_\sigma,\end{aligned}$$

Then,

$$\begin{aligned}(\Delta^{-1}\Phi_{m,k})_t &= (B\sum_{s=1}^{t} T_{s/m}B^*\phi_0)^* \\ &= (B\tilde{C}_0B^*U_{t/m}\phi_0)^* - (B\tilde{C}_0B^*\phi_0)^* \\ &= \phi_t^*(B\tilde{C}_0B^*)^* - \phi_0^*(B\tilde{C}_0B^*)^* \\ &= (\Phi_{m,k})_t\tilde{C} - \phi_0^*\tilde{C}.\end{aligned}$$

Thus, $\Delta^{-1}\Phi_{m,k} = \Phi_{m,k}\tilde{C} - 1_{m,k}V\tilde{C}$. □

Lemma 2.4.7.

$$\bar{\Delta}_\rho \Phi = \Phi \tilde{D} + \bar{\Delta}_\rho 1_{m,k} V,$$

where $B^* \tilde{D} B = diag(1 - e^{2\pi\imath\rho n_s/m})_{1\le s\le m}$.

Proof. For any $1 \le t \le m$,

$$\begin{aligned}
(\bar{\Delta}_\rho \Phi_{m,k})_t &= (U_{t/m}\phi_0 - U_{(t-\rho)/m}\phi_0)^* + \delta(t-\rho)\phi_0^* \\
&= (B(I_k - T_{-\rho/m})B^* U_{t/m}\phi_0)^* + \delta(t-\rho)\phi_0^* \\
&= \phi_t^* B(I_k - T_{\rho/m})B^* + \bar{\Delta}_\rho 1_{m,k} V \\
&= (\Phi\tilde{D})_t + (\bar{\Delta}_\rho 1_{m,k} V)_t.
\end{aligned}$$

□

Combining Lemma 2.4.6 and 2.4.7, one has the following expansion:

Proposition 2.4.8. *Given* $r, m, \rho \in \mathbb{N}$,

$$\rho^r A_r \Phi_{m,k} = D_\rho \bar{\Delta}_\rho^r \Delta^{-r} \Phi_{m,k} = D_\rho \left[\Phi_{m,k}\tilde{D}^r\tilde{C}^r + \sum_{j=0}^{r-1} \bar{\Delta}_\rho^{r-j} 1_{m,k} V \tilde{D}^j \tilde{C}^r - \bar{\Delta}_\rho^r \sum_{j=0}^{r-1} \Delta^{-j} 1_{m,k} V \tilde{C}^{r-j} \right]. \tag{2.19}$$

Remark 2.4.9. Note that $\tilde{D}\tilde{C} = \tilde{C}\tilde{D}$ as they are simultaneously diagonalizable by B_Φ, and thus $\tilde{D}^r\tilde{C}^r = (\tilde{D}\tilde{C})^r$.

Proof. First, we claim that, for $1 \le q \le r$, $\Delta^{-q}\Phi = \Phi\tilde{C}^q - \sum_{j=0}^{q-1} \Delta^{-j} 1_{m,k} V \tilde{C}^{q-j}$.

For $q = 1$, $\Delta^{-1}\Phi = \Phi\tilde{C} - 1_{m,k}V\tilde{C}$ by Lemma 2.4.6. For $q > 1$,

$$\begin{aligned}\Delta^{-q}\Phi &= \Delta^{-1}\left(\Phi\tilde{C}^{q-1} - \sum_{j=0}^{q-2}\Delta^{-j}1_{m,k}V\tilde{C}^{q-1-j}\right)\\ &= \Phi\tilde{C}^{q} - 1_{m,k}V\tilde{C}^{q} - \sum_{s=1}^{q-1}\Delta^{-s}1_{m,k}V\tilde{C}^{q-s}\\ &= \Phi\tilde{C}^{q} - \sum_{j=0}^{q-1}\Delta^{-j}1_{m,k}V\tilde{C}^{q-j}.\end{aligned}$$

As for the effect of $\bar{\Delta}_\rho$, we claim that $\bar{\Delta}_\rho^q\Phi = \Phi\tilde{D}^q + \sum_{j=0}^{q-1}\bar{\Delta}_\rho^{q-j}1_{m,k}V\tilde{D}^j$ for $1 \leq q \leq r$.

For $q = 1$, $\bar{\Delta}_\rho\Phi = \Phi\tilde{D} + \bar{\Delta}_\rho 1_{m,k}V$ by Lemma 2.4.7. For $q > 1$,

$$\begin{aligned}\bar{\Delta}_\rho^q\Phi &= \bar{\Delta}_\rho\left(\Phi\tilde{D}^{q-1} + \sum_{j=0}^{q-2}\bar{\Delta}_\rho^{q-1-j}1_{m,k}V\tilde{D}^j\right)\\ &= \Phi\tilde{D}^q + \bar{\Delta}_\rho 1_{m,k}V\tilde{D}^{q-1} + \sum_{j=0}^{q-2}\bar{\Delta}_\rho^{q-j}1_{m,k}V\tilde{D}^j\\ &= \Phi\tilde{D}^q + \sum_{j=0}^{q-1}\bar{\Delta}_\rho^{q-j}1_{m,k}V\tilde{D}^j.\end{aligned}$$

From the two assertions above, we get

$$\begin{aligned}\bar{\Delta}_\rho^r\Delta^{-r}\Phi &= \bar{\Delta}_\rho^r\left(\Phi\tilde{C}^r - \sum_{j=0}^{r-1}\Delta^{-j}1_{m,k}V\tilde{C}^{r-j}\right)\\ &= \Phi\tilde{D}^r\tilde{C}^r + \sum_{j=0}^{r-1}\bar{\Delta}_\rho^{r-j}1_{m,k}V\tilde{D}^j\tilde{C}^r - \bar{\Delta}_\rho^r\sum_{j=0}^{r-1}\Delta^{-j}1_{m,k}V\tilde{C}^{r-j}.\end{aligned}$$

□

2.4.2.2 Cancellation Between Residual Terms of $A_r\Phi_{m,k}$

From (2.19), we can divide $A_r\Phi_{m,k}$ into two parts: $\frac{1}{\rho^r}D_\rho\Phi_{m,k}\tilde{D}^r\tilde{C}^r$ being the main term, and the rest being residual terms. In this section, we shall investigate the behavior of the residual terms.

To facilitate the cancellation, we define an auxiliary double-sequence $\{a_{l,s}\}_{l\geq 0, s\in\mathbb{Z}}$ recursively by

$$a_{l,s} = \begin{cases} 1 & \text{if} \quad l = 0,\ s \geq 1 \\ 0 & \text{if} \quad l = 0,\ s \leq 0 \\ \sum_{j\leq s} a_{l-1,j} & \text{if} \quad l > 0. \end{cases}$$

Let $D_\rho\bar{\Delta}_\rho^{r-j}1_{m,k}V\tilde{D}^j\tilde{C}^r = I_j^{(2)}$ and $D_\rho\bar{\Delta}_\rho^r\Delta^{-j}1_{m,k}V\tilde{C}^{r-j} = I_j^{(3)}$. We first examine the form of each $I_j^{(3)}$ before calculating the cancellation between $I_j^{(2)}$ and $I_j^{(3)}$.

Lemma 2.4.10. *For any $j \in \mathbb{N}$ and $1 \leq l \leq m$,*

$$(\Delta^{-j}1_{m,k})_{l,s} = a_{j,l}.$$

Proof. First, it can easily be seen that $a_{l,s} = 0$ for all $s \leq 0$ by induction on l. Then, by definition and induction on j,

$$(\Delta^{-j}1_{m,k})_{l,s} = \sum_{n=1}^{l}(\Delta^{-j+1}1_{m,k})_{n,s} = \sum_{n=1}^{l} a_{j-1,n} = \sum_{n\leq l} a_{j-1,n} = a_{j,l}.$$

□

Lemma 2.4.11. *For $1 \leq \kappa \leq q$ and $1 \leq l \leq \eta$,*

$$(\bar{\Delta}_\rho^\kappa \Delta^{-q} 1_{m,1})_{l\rho} = \sum_{s_1,\ldots,s_\kappa=1}^{\rho} a_{q-\kappa,(l-\kappa)\rho+s_1+\cdots+s_\kappa}.$$

Proof. We shall prove this by induction on κ. For $\kappa = 1$ and $l > 1$,

$$(\bar{\Delta}_\rho \Delta^{-q} 1_{m,1})_{l\rho} = (\Delta^{-q} 1_{m,1})_{l\rho} - (\Delta^{-q} 1_{m,1})_{(l-1)\rho} = a_{q,l\rho} - a_{q,(l-1)\rho} = \sum_{s_1=1}^{\rho} a_{q-1,(l-1)\rho+s_1}.$$

For $l = 1$,

$$(\bar{\Delta}_\rho \Delta^{-q} 1_{m,1})_{\rho} = (\Delta^{-q} 1_{m,1})_{\rho} = a_{q,\rho} = a_{q,\rho} - a_{q,0} = \sum_{s_1=1}^{\rho} a_{q-1,0+s_1}.$$

For $1 < \kappa \leq q$ and $l > 1$,

$$\begin{aligned}(\bar{\Delta}_\rho^\kappa \Delta^{-q} 1_{m,1})_{l\rho} &= (\bar{\Delta}_\rho^{\kappa-1} \Delta^{-q} 1_{m,1})_{l\rho} - (\bar{\Delta}_\rho^{\kappa-1} \Delta^{-q} 1_{m,1})_{(l-1)\rho} \\ &= \sum_{s_1,\ldots,s_{\kappa-1}=1}^{\rho} (a_{q-\kappa+1,(l-\kappa+1)\rho+s_1+\cdots+s_{\kappa-1}} - a_{q-\kappa+1,(l-\kappa)\rho+s_1+\cdots+s_{\kappa-1}}) \\ &= \sum_{s_1,\ldots,s_\kappa}^{\rho} a_{q-\kappa,(l-\kappa)\rho+s_1+\cdots+s_\kappa}.\end{aligned}$$

As for $l = 1$,

$$\begin{aligned}(\bar{\Delta}_\rho^\kappa \Delta^{-q} 1_{m,1})_\rho &= (\bar{\Delta}_\rho^{\kappa-1} \Delta^{-q} 1_{m,1})_\rho \\ &= \sum_{s_1,\dots,s_{\kappa-1}=1}^{\rho} a_{q-\kappa+1,(1-\kappa+1)\rho+s_1+\cdots+s_{\kappa-1}} \\ &= \sum_{s_1,\dots,s_{\kappa-1}=1}^{\rho} a_{q-\kappa+1,(1-\kappa+1)\rho+s_1+\cdots+s_{\kappa-1}} - a_{q-\kappa+1,(0-\kappa+1)\rho+s_1+\cdots+s_{\kappa-1}} \\ &= \sum_{s_1,\dots,s_\kappa}^{\rho} a_{q-\kappa,(1-\kappa)rho+s_1+\cdots+s_\kappa},\end{aligned}$$

where the third equality follows from the fact that $s_1 + \cdots + s_{\kappa-1} \leq (\kappa - 1)\rho$. □

Proposition 2.4.12. *For $1 \leq l \leq r$,*

$$D_\rho \bar{\Delta}_\rho^l 1_{m,k} V \tilde{D}^{r-l} \tilde{C}^r - D_\rho \bar{\Delta}_\rho^r \Delta^{-r+l} 1_{m,k} V \tilde{C}^l = \Delta^l \Big(1_{\eta,k} V (\tilde{D}^{r-l} \tilde{C}^{r-l} - Id) + E_{r-l} \Big) \tilde{C}^l,$$

where $E_{r-l} = \tilde{B} 1_{\eta,k} V$, and $\tilde{B}$ is a diagonal matrix with $|\tilde{B}_{i,i}| \leq \rho^{r-l}$ for all $i \leq r$ and $\tilde{B}_{i,i} = 0$ otherwise.

Proof. From Lemma 2.4.11, we see that $(\bar{\Delta}_\rho^q \Delta^{-q} 1_{m,1})_{l\rho} = \sum_{s_1,\dots,s_q=1}^{\rho} a_{0,(l-q)\rho+s_1+\cdots+s_q}$. Thus, $(\bar{\Delta}_\rho^q \Delta^{-q} 1_{m,1})_{l\rho} = |Z_{l,q}|$, where

$$Z_{l,q} = \{(s_1, \dots, s_q) \in \mathbb{N}^q : 1 \leq s_1, \dots, s_1 \leq \rho,\ s_1 + \cdots + s_q > (q-l)\rho\}.$$

Note that $|Z_{l,q}| \leq \rho^q$, and $|Z_{l,q}| = \rho^q$ if $l \geq q$. Thus, $D_\rho \bar{\Delta}_\rho^q \Delta^{-q} 1_{m,1} = \rho^q 1_{\eta,1} - \tilde{b}$, where $\|\tilde{b}\|_\infty \leq \rho^q$ and $\tilde{b}_j = 0$ for all $j \geq q$. Then, we have

$$
\begin{aligned}
&D_\rho \bar{\Delta}_\rho^l 1_{m,k} V \tilde{D}^{r-l} \tilde{C}^r - D_\rho \bar{\Delta}_\rho^r \Delta^{-r+l} 1_{m,k} V \tilde{C}^l \\
&= D_\rho \bar{\Delta}_\rho^l \Big(1_{m,k} V \tilde{D}^{r-l} \tilde{C}^{r-l} - \bar{\Delta}_\rho^{r-l} \Delta^{-(r-l)} 1_{m,k} V \Big) \tilde{C}^l \\
&= \Delta^l D_\rho \Big(1_{m,k} V \tilde{D}^{r-l} \tilde{C}^{r-l} - \bar{\Delta}_\rho^{r-l} \Delta^{-(r-l)} 1_{m,k} V \Big) \tilde{C}^l \\
&= \Delta^l \Big(1_{\eta,k} V \tilde{D}^{r-l} \tilde{C}^{r-l} - D_\rho \bar{\Delta}_\rho^{r-l} \Delta^{-(r-l)} 1_{m,k} V \Big) \tilde{C}^l \\
&= \Delta^l \Big(1_{\eta,k} V \tilde{D}^{r-l} \tilde{C}^{r-l} - \rho^{r-l} 1_{\eta,k} V + \tilde{B} 1_{\eta,k} V \Big) \tilde{C}^l \\
&= \Delta^l \Big(1_{\eta,k} V (\tilde{D}^{r-l} \tilde{C}^{r-l} - \rho^{r-l} Id) + E_{r-l} \Big) \tilde{C}^l.
\end{aligned}
$$

□

2.4.3 Lower Frame Bound Estimate

Now, we are able to answer the first question in Section 2.4.1.

Lemma 2.4.13. *The 2-norm of* $(\frac{1}{\rho}\tilde{D}\tilde{C})^{-1}$ *satisfies* $\|(\frac{1}{\rho}\tilde{D}\tilde{C})^{-1}\|_2 \leq \frac{\pi}{2}$.

Proof. To prove the lemma, it suffices to show that for any unit-norm vector v, $\|\frac{1}{\rho}\tilde{D}\tilde{C}v\|_2 \geq \frac{2}{\pi}$. Note that $\tilde{D}$ and $\tilde{C}$ are simultaneously diagonalizable by the hermitian matrix B, so for any such v,

$$
\begin{aligned}
\|\frac{1}{\rho}\tilde{D}\tilde{C}v\|_2 &= \|\frac{1}{\rho} B(B^*\tilde{D}B)(B^*\tilde{C}B)B^* v\|_2 \\
&= \left\| diag\left(\frac{1 - e^{2\pi\imath\rho\lambda_s/m}}{\rho(1 - e^{2\pi\imath\lambda_s/m})} \right) (B^* v) \right\|_2 \\
&\geq \min_{s\in\{1,\ldots,k\}} \left| \frac{1 - e^{2\pi\imath\rho\lambda_s/m}}{\rho(1 - e^{2\pi\imath\lambda_s/m})} \right| \\
&= \min_s \left| \frac{\sin(\pi\lambda_s/\eta)}{\rho \sin(\pi\lambda_s/m)} \right| \geq \min_{t\in[-\eta/2,\eta/2]} \left| \frac{\sin(\pi t/\eta)}{\rho \sin(\pi t/m)} \right| \geq \frac{2}{\pi},
\end{aligned}
$$

where in the second equality, we note that since B is unitary, $\|MB\|_2 = \|BM\|_2 = \|M\|_2$ for any matrix M, and $\|B^*v\|_2 = \|v\|_2 = 1$. The second-to-last inequality comes from the assumption that $\{\lambda_s\}_{s=1}^k \subset [-\eta/2, \eta/2]$, and the final inequality can be obtained with simple calculus, see Lemma 4.5 in [47].

□

Lemma 2.4.14 (Proposition 5.2, [47]). *Given the assumption in Theorem 2.4.3 and n satisfying $n \mid m$ and $m/n \geq k$, $\Phi^*_{m/n,k}$ has lower frame bound larger than $\frac{m}{n}\min_s |<\phi_0, v_s>|^2 = \frac{m}{n}C_{\phi_0}$.*

Using Lemma 2.4.13 and 2.4.14, we are able to prove the following proposition:

Proposition 2.4.15. *Suppose $\eta = m/\rho \geq k \cdot 3r$, then $A_r\Phi_{m,k}$ is a frame with lower frame bound larger than $kC_{\phi_0}(\frac{2}{\pi})^{2r}$, where $\phi_0 = \sum_s c_s v_s$.*

Proof. First, note that

$$
\begin{aligned}
&D_\rho \bar{\Delta}_\rho^r \Delta^{-r} \Phi_{m,k} \\
&= D_\rho \left[\Phi_{m,k}\tilde{D}^r\tilde{C}^r + (\bar{\Delta}_\rho^r 1_{m,k} V + \cdots + \bar{\Delta}_\rho 1_{m,k} V \tilde{D}^{r-1})\tilde{C}^r - \bar{\Delta}_\rho^r (1_{m,k} V \tilde{C}^r + \cdots + (\Delta^{-r+1} 1_{m,k} V)\tilde{C}) \right] \\
&= \Phi_{\eta,k}\tilde{D}^r\tilde{C}^r + D_\rho \sum_{l=1}^{r} (\bar{\Delta}_\rho^l 1_{m,k} V \tilde{D}^{r-l}\tilde{C}^r - \bar{\Delta}_\rho^r \delta^{-r+l} 1_{m,k} V \tilde{C}^l) \\
&= \Phi_{\eta,k}\tilde{D}^r\tilde{C}^r + \sum_{l=1}^{r} \Delta^l D_\rho \left[1_{m,k} V \tilde{D}^{r-l}\tilde{C}^{r-l} - \bar{\Delta}_\rho^{r-l} \Delta^{-(r-l)} 1_{m,k} V \right] \tilde{C}^l \\
&= \Phi_{\eta,k}\tilde{D}^r\tilde{C}^r + \sum_{l=1}^{r} \Delta^l [1_{\eta,k} V (\tilde{D}^{r-l}\tilde{C}^{r-l} - I_k)]\tilde{C}^l + \Delta^l E_{r-l}\tilde{C}^l.
\end{aligned}
$$

Now, note that $\Delta^l 1_{\eta,k}$ has nonzero entries on only the first l rows. For $\Delta^l E_{r-l}$, only the first $r + l$ entries can be nonzero. Thus, the $l \cdot \lfloor \eta/k \rfloor$-th rows of $A_r\Phi_{m,k}$ is

equal to the one of $\frac{1}{\rho^r}\Phi_{\eta,k}\tilde{D}^r\tilde{C}^r$. Now, the lower frame bound of $A_r\Phi_{m,k}$ is larger than the one of any of its sub-frame. In particular, its lower frame bound is larger than the one of $\frac{1}{\rho^r}\Phi_{k,k}\tilde{D}^r\tilde{C}^r$, which is $kC_{\phi_0}(\frac{2}{\pi})^{2r}$, since for any unit-norm vector v,

$$\|\frac{1}{\rho^r}\Phi_{k,k}\tilde{D}^r\tilde{C}^r v\|_2^2 \geq kC_{\phi_0}\|(\frac{1}{\rho}\tilde{D}\tilde{C})^r v\|_2^2 \geq kC_{\phi_0}\left(\frac{2}{\pi}\right)^{2r}.$$

□

2.4.4 Frame Variation Bound

In (2.18), we also need to estimate $\|(A_r\Phi_{m,k})^*\Delta^r\|_{\infty,2}$.

Proposition 2.4.16.

$$\|(A_r\Phi_{m,k})^*\Delta^r\|_{\infty,2} \leq 2^{2r+2}\eta^{r-1}.$$

Proof. From Proposition 2.4.8 and 2.4.12, we see that

$$D_\rho\bar{\Delta}^r\Delta^{-r}\Phi_{m,k} = \Phi_{\eta,k}\tilde{D}^r\tilde{C}^r + \sum_{l=1}^{r}\Delta^l\left(1_{\eta,k}V(\tilde{D}^{r-l}\tilde{C}^{r-l} - \rho^{r-l}Id) + E_{r-l}\right)\tilde{C}^l.$$

Thus,

$$\|(A_r\Phi_{m,k})^*\Delta^r\|_{\infty,2} = \|\frac{1}{\rho^r}(D_\rho\bar{\Delta}^r\Delta^{-r}\Phi_{m,k})^*\Delta^r\|_{\infty,2}$$
$$\leq \|\frac{1}{\rho^r}\tilde{D}^r\tilde{C}^r\|_2\|\Phi_{\eta,k}^*\Delta^r\|_{\infty,2} + 2\sum_{l=1}^{r}\|\frac{1}{\rho^l}\tilde{C}^l\|_2\|\frac{1}{\rho^{r-l}}\tilde{D}^{r-l}\tilde{C}^{r-l} - Id\|_2\|V^*1_{k,\eta}\Delta^{l+r}\|_{\infty,2},$$

where we observe that $\|\frac{1}{\rho^{r-l}}E_{r-l}^*\Delta^{r+l}\|_{\infty,2} \leq \|V^*1_{k,\eta}\Delta^{r+l}\|_{\infty,2}$.

Now, $\|\Phi^*_{\eta,k}\Delta^r\|_{\infty,2} \leq r2^r + \eta(2\pi \max_{1\leq j\leq k} |\lambda_j|\frac{1}{\eta})^r$ by Lemma 2.3.22, and $\|V^* 1_{k,\eta}\Delta^{l+r}\|_{\infty,2} = 2^{l+r-1}\|\phi_0\|_2 = 2^{l+r-1}$. Moreover, $\|\frac{1}{\rho^r}\tilde{D}^r\tilde{C}^r\|_2 \leq 1$, $\|\frac{1}{\rho^{r-l}}\tilde{D}^{r-l}\tilde{C}^{r-l} - Id\|_2 \leq 2$, and $\|\frac{1}{\rho^l}\tilde{C}^l\|_2 \leq \eta^l$. Thus,

$$\begin{aligned}\|\frac{1}{\rho^r}(D_\rho\bar{\Delta}^r\Delta^{-1}\Phi_{m,k})^*\Delta^r\|_{\infty,2} &\leq r2^r + \eta(2\pi \max_{1\leq j\leq k} |\lambda_j|\frac{1}{\eta})^r + 2^{r+1}\frac{(2\eta)^r - 1}{2\eta - 1} \\ &\leq r2^r + \eta(2\pi \max_{1\leq j\leq k} |\lambda_j|\frac{1}{\eta})^r + 2^{2r+1}\eta^{r-1} \leq 2^{2r+2}\eta^{r-1},\end{aligned}$$

independent of m. □

2.4.5 Data Storage Efficiency

Given a mid-rise quantizer with length $2L$ and the quantized sample $q \in \mathbb{C}^m$, one needs $\log(2L)$ bits to record each entry of q. Thus, a total of $m\log(2L) = O(\rho)$ bits is needed to fully record q as $\rho \to \infty$. In this section, we shall show that with the application of adapted decimation, we may now record the decimated signal in $O(\log(\rho))$ bits, drastically fewer than originally needed.

Proposition 2.4.17. *Given a mid-rise quantizer with length $2L$, it is possible to encode $D_\rho\bar{\Delta}^r_\rho\Delta^{-r}q$ with $2\eta r\log(2m) + 2\eta\log(2L)$ bits in total.*

Proof. Note that for mid-rise uniform quantizers $\mathscr{A} = \mathscr{A}_0 + \imath\mathscr{A}_0$ with length $2L$, each entry q_j of q is a number of the form

$$q_j = \big((2s_j + 1) + \imath(2t_j + 1)\big)\frac{\delta}{2}, \quad -L \leq s_j,\, t_j \leq L - 1.$$

Then, each entry in $\Delta^{-1}q$ is the summation of at most m entries in q, which has the

form

$$(\Delta^{-1}q)_j = \big((2\tilde{s}_j+\rho)+\imath(2\tilde{t}_j+\rho)\big)\frac{\delta}{2}, \quad -Lm \le \tilde{s}_j,\, \tilde{t}_j \le (L-1)m.$$

Iterating r times, we see that

$$(\Delta^{-r}q)_j = \big((2\tilde{s}_j+\rho)+\imath(2\tilde{t}_j+\rho)\big)\frac{\delta}{2}, \quad -Lm^r \le \tilde{s}_j,\, \tilde{t}_j \le (L-1)m^r.$$

As for $\bar{\Delta}_\rho^r\Delta^{-r}q$, we see that, for any $v \in \mathbb{C}^m$, each entry of $\bar{\Delta}_\rho v$ contains at most 2 entries of v. Thus,

$$(\bar{\Delta}_\rho^r\Delta^{-r}q)_j = \big((2\tilde{s}_j+\rho)+\imath(2\tilde{t}_j+\rho)\big)\frac{\delta}{2}, \quad -L(2m)^r \le \tilde{s}_j,\, \tilde{t}_j \le (L-1)(2m)^r$$

Now, there are at most $((2L-1)(2m)^r+1)^2 \le (2L(2m)^r)^2$ choices per entry with $\eta = m/\rho$ entries in total for $D_\rho\bar{\Delta}_\rho^r\Delta^{-r}q$. Thus, it can be encoded by $\mathscr{R} = 2\eta r\log(2m) + 2\eta\log(2L)$ bits.

□

2.4.6 Proof of Theorem 2.4.3

Proof. of Theorem 2.4.3:

By Lemma 2.4.4,

$$\rho^r A_r q = D_\rho\bar{\Delta}_\rho^r\Delta^{-r}q = \Delta^r D_\rho(\Delta^{-r}q).$$

Since Δ and Δ^{-1} are lower-triangular, we see that, for any $1 \le s \le \eta$, there exists $\{a_j^s\}_{j=1}^s$ and $\{b_j^l\}_{j,l}$ such that

$$(A_r q)_s = \sum_{j=1}^{s} a_j^s (D_\rho \Delta^{-r} q)_j = \sum_{j=1}^{s} a_j^s (\Delta^{-r} q)_{j\rho} = \sum_{j=1}^{s} a_j^s \sum_{l=1}^{j\rho} b_l^j q_l = \sum_{\xi=1}^{s\rho} c_\xi q_\xi,$$

proving the first claim. The second assertion follows from Proposition 2.4.15.

Given $\Phi = \Phi_{m,k}$, $A = A_r = \frac{1}{\rho^r} D_\rho \bar{\Delta}_\rho^r \Delta^{-r}$, and $\mathcal{S} = (A\Phi)^* A\Phi$, the reconstruction error can be estimated as follows:

$$\begin{aligned}
\mathscr{E} = \|\mathcal{S}^{-1}(A\Phi)^* Aq - x\|_2 &= \|\mathcal{S}^{-1}(A\Phi)^* A\Delta^r u\|_2 \\
&= \frac{1}{\rho^r} \|\mathcal{S}^{-1}(A\Phi)^* D_\rho \bar{\Delta}_\rho^r u\|_2 \\
&= \frac{1}{\rho^r} \|\mathcal{S}^{-1}(A\Phi)^* \Delta^r D_\rho u\|_2 \\
&\le \frac{1}{\rho^r} \|\mathcal{S}^{-1}\|_2 \|(A\Phi)^* \Delta^r\|_{\infty,2} \|D_\rho u\|_\infty \\
&\le \frac{1}{\rho^r} \left(kC_{\phi_0}(\frac{2}{\pi})^{2r}\right)^{-1} 2^{2r+2} \eta^{r-1} \|u\|_\infty \\
&= \left(\frac{4}{k\eta C_{\phi_0}} (\pi^2 \eta)^r\right) \|u\|_\infty \frac{1}{\rho^r},
\end{aligned}$$

where the second inequality comes from Proposition 2.4.15 and Proposition 2.4.16.

As for the data storage, we see from Proposition 2.4.17 that one can encode the data $A_r q$ with $\mathscr{R} = 2\eta r \log(2m) + 2\eta \log(2L)$ bits in total.

Note that

$$e^{\frac{-\mathscr{R}}{2\eta}} = (2m)^{-r} \cdot \frac{1}{2L} = \frac{1}{2L} \left(\frac{\eta}{2}\right)^r \cdot \frac{1}{\rho^r}.$$

Thus, as the function of bits used, the reconstruction error satisfies

$$\begin{aligned}\mathscr{E}(\mathscr{R}) &\leq \left(\frac{4}{k\eta C_{\phi_0}}(\pi^2\eta)^r\right)\|u\|_\infty\frac{1}{\rho^r}\\ &= C_{k,\eta,\phi_0,L}\|u\|_\infty\frac{1}{2L}\left(\frac{\eta}{2}\right)^r\frac{1}{\rho^r}\\ &= C_{k,\eta,\phi_0,L}\|u\|_\infty e^{\frac{-\mathscr{R}}{2\eta}},\end{aligned}$$

where $C_{k,\eta,\phi_0,L} = \frac{8L}{k\eta C_{\phi_0}}(2\pi^2)^r$.

□

Chapter 3: Compressive Sensing

3.1 Introduction and Motivation

In this chapter we estimate the following sum: given a prime $p \in \mathbb{N}$ and $n \in \mathbb{Z}/p\mathbb{Z}$, suppose that $M_1, M_2 \subset \mathbb{Z}/p\mathbb{Z}$ are two sets of consecutive numbers with $|M_1| \leq |M_2| \leq \sqrt{p}$. We would like to estimate

$$|\sum_k \sum_{m_1\in M_1}\sum_{m_2\in M_2} \chi[k+m_1-m_2]\chi[k]e^{2\pi\imath kn/p}e^{-2\pi\imath m_2 n/p}|, \tag{3.1}$$

where $\chi : \mathbb{Z}/p\mathbb{Z} \to \mathbb{C}$ is a non-principal character.

The sum in (3.1) is related to deterministic compressive sensing, character sums, and Weil's exponential sum estimates. From all prior works, one can easily

derive an upper bound of $p^{3/2}$ for (3.1). However, as such an estimate is not sufficient for our purpose, we shall prove that it is possible to improve the estimate to $p^{3/2-\alpha}$ under certain mild assumptions, where $\alpha \in (0, 1/2)$ depends on $|M_1|$ and n.

Motivated by this, we aim to construct deterministic matrices with bottleneck-breaking RIP from the Gabor system of Legendre symbols. Our formulation follows from (3.5): given a prime $p \in \mathbb{N}$, consider $\{u_{l,j}\}_{l,j\in\mathbb{Z}/p\mathbb{Z}} \subset \mathbb{C}^p$ where $u_{l,j}[k] = \frac{1}{\sqrt{p}}\chi[k-l]e^{-2\pi\imath kj/p}$ with χ being the Legendre symbol. Fix disjoint $\Omega_1, \Omega_2 \subset \mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$ where $|\Omega_1|, |\Omega_2| \leq \sqrt{p}$, define $\pi_2(\Omega_i) = \{j \in \mathbb{Z}/p\mathbb{Z} : \exists l \in \mathbb{Z}/p\mathbb{Z} \text{ such that } (l, j) \in \Omega_i\}$ and $\Omega_i(j) = \{l \in \mathbb{Z}/p\mathbb{Z} : (l, j) \in M_i\}$ for $i = 1, 2$. Then,

$$\begin{aligned}
&\left| < \sum_{(m_1,n_1)\in\Omega_1} u_{m_1,n_1}, \sum_{(m_2,n_2)\in\Omega_2} u_{m_2,n_2} > \right| \\
&= \left| \frac{1}{p} \sum_{n_1\in\pi_2(\Omega_1)} \sum_{n_2\in\pi_2(\Omega_2)} \sum_{k\in\mathbb{Z}/p\mathbb{Z}} \sum_{m_1\in\Omega_1(n_1)} \sum_{m_2\in\Omega_2(n_2)} \chi[k+m_1-m_2]\chi[k]e^{2\pi\imath k(n_1-n_2)/p}e^{-2\pi\imath m_2(n_1-n_2)/p} \right| \\
&\leq \frac{1}{p} \sum_{n_1\in\pi_2(\Omega_1)} \sum_{n_2\in\pi_2(\Omega_2)} \left| \sum_{k\in\mathbb{Z}/p\mathbb{Z}} \sum_{m_1\in\Omega_1(n_1)} \sum_{m_2\in\Omega_2(n_2)} \chi[k+m_1-m_2]\chi[k]e^{2\pi\imath k(n_1-n_2)/p}e^{-2\pi\imath m_2(n_1-n_2)/p} \right|.
\end{aligned} \tag{3.2}$$

Note that the expression in inside the final absolute value of (3.2) is exactly (3.1) when $\Omega_1(n_1), \Omega_2(n_2)$ are consecutive numbers. In order to use Lemma 3.2.6, we aim to show that (3.1) is less than $p^{3/2-\alpha}$ for some $\alpha > 0$.

3.2 Preliminaries

3.2.1 Restricted Isometry Property

Introduced in [13] and refined in [12], the Restricted Isometry Property (RIP) is defined as follows:

Definition 3.2.1. An $n \times m$ matrix A satisfies (S, δ_S)-RIP if the following statement is true: Let A_T, $T \subset \{1, \ldots, m\}$ be the $n \times |T|$ submatrix obtained by extracting the columns of A which corresponds to the elements in T. Then for any subset T with $|T| \leq S$ and any coefficient sequence $\{c_j\}_{j \in T}$, we have

$$(1 - \delta_S)\|c\|_2^2 \leq \|A_T c\|_2^2 \leq (1 + \delta_S)\|c\|_2^2. \tag{3.3}$$

For sampling schemes satisfying RIP, one is able to retrieve sparse signals efficiently from highly incomplete measurements because of the equivalence between the following optimization problems:

$$\min \|x\|_{\ell_0} \quad \text{subject to } Ax = b, \tag{P_0}$$

where $\|x\|_{\ell_0}$ denotes the number of nonzero entries of x, and

$$\min \|x\|_{\ell_1} \quad \text{subject to } Ax = b. \tag{P_1}$$

(P_0) and (P_1) do not yield the same solution in general, but for matrices satisfying

RIP with small constant δ, the two problems will be equivalent provided that the signal itself is sparse, [15]. (P_0) is a non-convex optimization problem, whereas (P_1) is convex and is readily solvable. Thus, solving (P_1) is much more preferable to solving (P_0).

Using probabilistic estimates, one can show that given $\epsilon > 0$, there exists a random matrix $A \in \mathbb{C}^{M\times N}$ satisfies (S, δ_S)-RIP with $M^{1-\epsilon} \ll S \ll M$ with exponentially high probability.

3.2.2 Square-root Bottleneck

Compressive sensing has found great success in probabilistic settings. However, deterministically one is not able to obtain such strong results: very few methods are available other than the coherence estimate, and it is extremely hard to extend the order S to $S \gg \sqrt{M}$. We describe the method and its limitation below:

Definition 3.2.2 (Coherence parameter)**.** Given a matrix $\Phi = (\phi_1 \mid \phi_2 \mid \cdots \mid \phi_r)$ with unit column vectors, the coherence parameter μ of Φ is defined to be

$$\mu := \max_{i\neq j} |<\phi_i, \phi_j>|.$$

Proposition 3.2.3 ([9], Proposition 1)**.** *Given a matrix $\Phi \in \mathbb{C}^{n\times m}$ with unit norm columns $\{\phi_i\}_i$. If the coherence of Φ is μ, then Φ satisfies $(k, (k-1)\mu)$-RIP for all k.*

Proof. Given any k-sparse vector $x \in \mathbb{C}^m$, let $T \subset \{1, \ldots, m\}$ be the set of its

non-zero entries. Then, one has

$$\begin{aligned} |\|\Phi x\|^2 - \|x\|^2| &= |2 \sum_{r<s\in T} x_r x_s <\phi_r, \phi_s>| \\ &\leq \mu((\sum_j |x_r|)^2 - \|x\|^2) \leq \mu(|T|-1)\|x\|^2 \leq \mu(k-1). \end{aligned}$$

□

Following the proposition above, it is favorable to find a matrix with coherence as low as possible. However, the coherence parameter is bounded below by the following universal bound:

Proposition 3.2.4. *Given any matrix $\Phi \in \mathbb{C}^{n\times m}$ with unit norm columns, the coherence parameter μ is lower bounded by*

$$\mu \geq C\sqrt{\frac{\log m}{n\log(n/\log m)}} \geq \frac{C}{\sqrt{n}},$$

for $\log m \leq n \leq m/2$ and a fixed $C > 0$.

Proposition 3.2.4 shows that one can construct an (S, δ_S)-RIP with $S \sim \sqrt{n}$ by constructing matrices with low coherence. However, anything more than that is significantly harder. In fact, few techniques are available other than the coherence approach, making the explicit construction of matrices satisfying (S, δ_S)-RIP with $S \gg \sqrt{n}$ extremely hard. Such difficulty is denoted as the square-root bottleneck.

Bourgain et al. [9] proposed a new class of matrices satisfying RIP of high order, breaking the bottleneck by constructing a family of matrices satisfying (S, δ_S)-RIP with $S \sim M^{1/2+\epsilon}$, where ϵ is of the order of 10^{-28}. Mixon [51] improved the ϵ to

the order of 10^{-24}, more than $8,000$ times better than the original result. One key ingredient of their proofs is the following notion of flat RIP.

Definition 3.2.5 (flat RIP)**.** Let $u_1, \ldots, u_N$ be the columns of an $n \times N$ matrix Φ. Suppose that for every j, $\|u_j\|_2 = 1$. Φ satisfies the (k, δ)-flat RIP if for any disjoint $J_1, J_2 \subset \{1, \ldots, N\}$ with $|J_1|, |J_2| \leq k$ we have

$$| < \sum_{j \in J_1} u_j, \sum_{i \in J_2} u_i > | \leq \delta(|J_1||J_2|)^{1/2}. \tag{3.4}$$

The following lemma takes a slightly weaker form of flat RIP.

Lemma 3.2.6. *Let $k \geq 2^{10}$ and s be any positive integer. Assume that the coherence parameter of Φ is $\mu \leq 1/k$, and for some δ and any disjoint J_1, J_2 with $|J_1|, |J_2| \leq k$, one has*

$$\left| < \sum_{j_1 \in J_1} u_{j_1}, \sum_{j_2 \in J_2} u_{j_2} > \right| \leq \delta k, \tag{3.5}$$

then Φ satisfies RIP of order $(2sk, 44s\delta \log k)$-RIP.

By Lemma 3.2.6, matrices satisfying flat RIP also satisfy RIP of high order, which provides insights on how to approach this problem from a new direction.

3.2.3 Björck Sequence

For each prime number p, the *Legendre symbol* modulo p is the function $\chi = \left(\frac{\cdot}{p}\right) : \mathbb{Z}/p\mathbb{Z} \to \{-1, 0, 1\}$ given by

$$\chi[k] = \left(\frac{k}{p}\right) = \begin{cases} +1 & \text{if } k \equiv m^2 \,(\text{mod } p) \text{ for some } m \in \mathbb{Z}/p\mathbb{Z}^\times \\ 0 & \text{if } k \equiv 0 \,(\text{mod } p) \\ -1 & \text{if } k \not\equiv m^2 \,(\text{mod } p) \text{ for all } m \in \mathbb{Z}/p\mathbb{Z}^\times \end{cases}$$

Let the set $\mathcal{Q}$ be there nonzero *quadratic residues* modulo p, and $\mathcal{Q}^C$ be the quadratic nonresidues modulo p. Note that $\mathcal{Q} = \chi^{-1}(1)$, and $\mathcal{Q}^C = \chi^{-1}(-1)$.

Definition 3.2.7. The *Björck* sequence $\{u_p[k]\}_{k \in \mathbb{Z}/p\mathbb{Z}}$ of length p, where p is an odd prime, is defined as follows:

For any $k \in \mathbb{Z}/p\mathbb{Z}$, if $p \equiv 1 \,(\text{mod } 4)$, then

$$u_p[k] = \exp(\imath\theta\chi(k)) = \exp\left(\imath\theta\left(\frac{k}{p}\right)\right), \quad \text{where } \theta = \arccos\left(\frac{1}{1+\sqrt{p}}\right).$$

If $p \equiv 3 \,(\text{mod } 4)$, then

$$u_p[k] = \begin{cases} \exp(\imath\phi) & \text{if } k \in \mathcal{Q}^C \subset (\mathbb{Z}/p\mathbb{Z})^\times, \text{ where } \phi = \arccos(\frac{1-p}{1+p}). \\ 1 & \text{otherwise,} \end{cases}$$

The Björck sequence $\{u_p[k]\}_k$ is an example of a *constant amplititude zero auto-correlation* (CAZAC) sequence. The definition of a CAZAC sequence is as

follows:

Definition 3.2.8. A sequence $v = \{v_k\}_{k\in\mathbb{Z}/p\mathbb{Z}}$ is CAZAC if

- There exists $C \geq 0$ such that $|v_k| = C$ for all k.
- For any $t \in (\mathbb{Z}/p\mathbb{Z})\backslash\{0\}$, one has $<v, \tau_t v> = 0$, where $(\tau_t v)_k = v_{k-t}$ with cyclic convention.

Now, for any odd prime p, consider the following Gabor frame $\Psi_p = (\psi_{m,n})_{m,n\in\mathbb{Z}/p\mathbb{Z}} \in \mathbb{C}^{p\times p^2}$, where

$$\psi_{m,n}[k] = \frac{1}{\sqrt{p}} u_p[k+m]e_p(kn),$$

where $e_p(x) := \exp(-2\pi\imath x/p)$. Then for any $(m_1, n_1), (m_2, n_2) \in (\mathbb{Z}/p\mathbb{Z})^2$, one has

$$\begin{aligned}
<\psi_{m_1,n_1}, \psi_{m_2,n_2}> &= \frac{1}{p}\sum_{k\in\mathbb{Z}/p\mathbb{Z}} \Big(u_p[k+m_1]e_p(kn_1)\Big)\Big(\overline{u_p[k+m_2]}e_p(-kn_2)\Big) \\
&= \frac{1}{p}\sum_{k\in\mathbb{Z}/p\mathbb{Z}} u_p[k+m_1]\overline{u_p[k+m_2]}e_p(k(n_1-n_2)) \\
&= \frac{1}{p}\sum_{t\in\mathbb{Z}/p\mathbb{Z}} u_p[t+(m_1-m_2)]\overline{u_p[t]}e_p(t(n_1-n_2))e_p(-m_2(n_1-n_2)) \\
&= A_p(u_p)[m_1-m_2, n_1-n_2]e_p(-m_2(n_1-n_2)),
\end{aligned}$$

where $A_p(u_p) = \frac{1}{p}\sum_{t\in\mathbb{Z}/p\mathbb{Z}} u_p[t+(m_1-m_2)]\overline{u_p[t]}e_p(t(n_1-n_2))$ is the ambiguity function of u_p.

For Björck sequences, one has the following estimate:

Theorem 3.2.9 (Theorem 3.8 in [7]). *For any* $(m, n) \in (\mathbb{Z}/p\mathbb{Z}\times\mathbb{Z}/p\mathbb{Z})\backslash\{(0,0)\}$,

one has

$$|A_p(u_p)[m,n]| \leq \frac{2}{\sqrt{p}} + \begin{cases} \frac{4}{p} & \text{if } p \equiv 1 \,(mod\ 4) \\ \frac{4}{p^{3/2}} & \text{if } p \equiv 3 \,(mod\ 4) \end{cases}.$$

In particular, $|A_p(u_p)[m,n]| \leq 3/\sqrt{p}$, which implies that the coherence of Ψ_p is $\mu_\Psi \leq 3/\sqrt{p}$.

3.2.4 Reduction to Legendre Symbols

For given $\mathcal{N} \subset \mathbb{Z}/p\mathbb{Z}$ and disjoint $\Omega_1, \Omega_2 \subset \mathbb{Z}/p\mathbb{Z} \times \mathcal{N}$, we introduce the following notation:

Definition 3.2.10. For $i = 1, 2$, define $\Omega_i(n) = \{m \in \mathbb{Z}/p\mathbb{Z} : (m,n) \in \Omega_i\}$.

Then, one has

$$\sum_{(m_1,n_1)\in\Omega_1} \sum_{(m_2,n_2)\in\Omega_2} <u_{m_1,n_1}, u_{m_2,n_2}> = \sum_{n_1,n_2\in\mathcal{N}} \sum_{m_2\in\Omega_2(n_2)} \sum_{m_1\in\Omega_1(n_1)} A_p(u_p)[m_1-m_2, n_1-n_2]e_p(-m_2(n_1-n_2)). \tag{3.6}$$

Moreover, we may assume $m_1 \neq m_2$, $n_1 \neq n_2$ as $A_p(u_p)[0,n] = A_p(u_p)[m,0] = 0$ for all $m, n \neq 0$.

Now, fixing m_2, n_1, n_2, we see that the innermost sum of (3.6) becomes

$$\sum_{m_1\in\Omega_1(n_1)} A_p(u_p)[m_1-m_2, n_1-n_2]e_p(-m_2(n_1-n_2)) =$$
$$\frac{1}{p} \sum_{m_1\in\Omega_1(n_1)} \sum_{k\in\mathbb{Z}/p\mathbb{Z}} u_p[k+m_1-m_2]\overline{u_p[k]}e_p(k(n_1-n_2))e_p(-m_2(n_1-n_2)).$$

Note that for $p \equiv 1 (\mathrm{mod}\ 4)$,

$$\Re(u_p[k]) = \begin{cases} \frac{1}{1+\sqrt{p}} & \text{if} \quad k \neq 0 \\ 1 & \text{if} \quad k = 0 \end{cases},$$

and $\Im(u_p[k]) = \frac{\sqrt{p+2\sqrt{p}}}{1+p}\chi[k]$, a multiple of the Legendre symbol.

Proposition 3.2.11. *Given $\mathcal{N}, \Omega_1, \Omega_2$, the estimate in (3.6) can be written as*

$$|\sum_{(m_1,n_1)\in\Omega_1}\sum_{(m_2,n_2)\in\Omega_2} <u_{m_1,n_1}, u_{m_2,n_2}>| \leq \sum_{n_1,n_2\in\mathcal{N}}\sum_{m_2\in\Omega_2(n_2)}|\sum_{m_1\in\Omega_1(n_1)} A_p(\chi)[m_1-m_2, n_1-n_2]| + \frac{8}{9}.$$

Before proving the proposition above, we first derive the following lemma:

Lemma 3.2.12. *Given fixed m_2, n_1, n_2, one has*

$$\begin{cases} |\frac{1}{p}\sum_{m_1\in\Omega_1(n_1)}\sum_{k\in\mathbb{Z}/p\mathbb{Z}}\Re(u_p[k+m_1-m_2])\Re(u_p[k])e_p(k(n_1-n_2))| \leq \frac{2|\Omega_1(n_1)|}{p} \\ |\frac{1}{p}\sum_{m_1\in\Omega_1(n_1)}\sum_{k\in\mathbb{Z}/p\mathbb{Z}}\Im(u_p[k+m_1-m_2])\Re(u_p[k])e_p(k(n_1-n_2))| \leq \frac{3|\Omega_1(n_1)|}{p} \\ |\frac{1}{p}\sum_{m_1\in\Omega_1(n_1)}\sum_{k\in\mathbb{Z}/p\mathbb{Z}}\Re(u_p[k+m_1-m_2])\Im(u_p[k])e_p(k(n_1-n_2))| \leq \frac{3|\Omega_1(n_1)|}{p} \end{cases}.$$

Proof. For the first part, note that $\Re(u_p[k]) = \frac{1}{1+\sqrt{p}}$ if $k \neq 0$ and is equal to 1 if $k = 0$. Thus, one has

$$\begin{aligned} &\sum_{m_1\in\Omega_1(n_1)}\Re(u_p[k+m_1-m_2])\Re(u_p[k])e_p(k(n_1-n_2)) \\ &= \frac{1}{1+\sqrt{p}}\sum_{m_1\in\Omega_1(n_1)}\sum_{k\in\mathbb{Z}/p\mathbb{Z}}e_p(k(n_1-n_2)) + \frac{\sqrt{p}}{1+\sqrt{p}}\sum_{m_1\in\Omega_1(n_1)}\Big(1+e_p((m_2-m_1)(n_1-n_2))\Big) \\ &= \frac{\sqrt{p}}{1+\sqrt{p}}\Big(|\Omega_1(n_1)| + \sum_{m_1\in\Omega_1(n_1)}e_p((m_2-m_1)(n_1-n_2))\Big). \end{aligned}$$

As for the second part, we have

$$\sum_{k\in\mathbb{Z}/p\mathbb{Z}} \Im(u_p[k+m_1-m_2])\Re(u_p[k])e_p(k(n_1-n_2))$$
$$= \frac{1}{1+\sqrt{p}}\sum_{k\in\mathbb{Z}/p\mathbb{Z}} \Im(u_p[k+m_1-m_2])e_p(k(n_1-n_2)) + \frac{\sqrt{p}}{1+\sqrt{p}}\Im(u_p[m_1-m_2])$$
$$= \frac{1}{\imath}\left(\frac{\sqrt{p}}{1+\sqrt{p}}\hat{u}_p[n_1-n_2]e_p((m_2-m_1)(n_1-n_2)) - \frac{\sqrt{p}}{1+\sqrt{p}}e_p((m_2-m_1)(n_1-n_2))\right)$$
$$+ \frac{\sqrt{p}}{1+\sqrt{p}}\Im(u_p[m_1-m_2])e_p(k(n_1-n_2)).$$

Now, $|\hat{u}_p| = \Im(u_p[m_1-m_2]) = 1$ since $m_1 \neq m_2$. Thus, the magnitude of the three term does not exceed 3. The third estimate follows verbatim from the second one. □

Proof. of Proposition 3.2.11:

By noting

$$u_p[k+m_1-m_2]\overline{u_p[k]} = \Re(u_p[k+m_1-m_2])\Re(u_p[k]) + \imath\Im(u_p[k+m_1-m_2])\Re(u_p[k])$$
$$- \imath\Re(u_p[k+m_1-m_2])\Im(u_p[k]))$$
$$- \imath\frac{\sqrt{p+2\sqrt{p}}}{1+\sqrt{p}}\chi[k+m_1-m_2]\chi[k],$$

we have

$$|\sum_{m_1\in\Omega_1(n_1)} A_p(u_p)[m_1-m_2,n_1-n_2]| \leq |\sum_{m_1\in\Omega_1(n_1)} A_p(\chi)[m_1-m_2,n_1-n_2]| + \frac{8|\Omega_1(n_1)|}{p}.$$

Then, by summing over m_2, n_1, n_2, we see that

$$
\begin{aligned}
&| \sum_{(m_1,n_1)\in\Omega_1} \sum_{(m_2,n_2)\in\Omega_2} <u_{m_1,n_1}, u_{m_2,n_2}>| \\
&\leq \sum_{n_1,n_2\in\mathcal{N}} \sum_{m_2\in\Omega_2(n_2)} \sum_{m_1\in\Omega_1(n_1)} |A_p(\chi)[m_1-m_2, n_1-n_2] e_p(-m_2(n_1-n_2))| \\
&\quad + \sum_{n_1,n_2\in\mathcal{N}} \sum_{m_2\in\Omega_2(n_2)} \frac{8|\Omega_1(n_1)|}{p} \\
&= \sum_{n_1,n_2\in\mathcal{N}} \sum_{m_2\in\Omega_2(n_2)} \sum_{m_1\in\Omega_1(n_1)} |A_p(\chi)[m_1-m_2, n_1-n_2]| + \frac{8|\Omega_1||\Omega_2|}{p},
\end{aligned}
$$

and that the last term is less than $8/9$.

□

3.2.5 Character Sum Estimates

Besides the practical interests in compressive sensing, estimation of character sums is also intriguing in its own. Let $\chi : \mathbb{Z}/p\mathbb{Z} \to \mathbb{C}$ be a non-principal character on $(\mathbb{Z}/p\mathbb{Z})^*$ with the extension $\chi[0] = 0$. Polya-Vinogradov inequality states that

$$
|\sum_{M\leq k\leq M+N} \chi[k]| \leq \sqrt{p}\log p
$$

for any arbitrary M, N. Chung [21] investigated the cancellation within the sum

$$
\sum_{a\in S} \sum_{b\in T} \chi[a+b]
$$

where $S, T \subset \mathbb{Z}/p\mathbb{Z}$. In particular, the following estimate is given:

$$|\sum_{a\in S}\sum_{b\in T}\chi[a+b]| \leq \sqrt{p|S||T|}(1-\frac{|S|}{p})^{1/2}(1-\frac{|T|}{p})^{1/2}.$$

Note that the estimate is only nontrivial for $|S|, |T| \gg \sqrt{p}$. Chung also commented on a conjecture for the case $|S| \ll \sqrt{p}$: for any fixed $\epsilon > 0$ and $|S| > p^{\epsilon}$, there exists $\delta > 0$ such that

$$|\sum_{a,b\in S}\chi[a-b]| < |S|^{2-\delta}.$$

Friedlander and Iwaniec [36] gave a partial answer to the conjecture above, proving the inequality when S is contained in an interval I of length $\ll \sqrt{p}$ and satisfies $|S| \geq I^{r/(r+1)}p^{1/4r+\epsilon}$ for some $r \geq 2$ using the Burgess estimate. Note that the results here do not apply to (3.1) even if $\Omega_1(n_1) = \Omega_2(n_2)$, since there is an additional summation over $\mathbb{Z}/p\mathbb{Z}$.

3.2.6 Weil's Exponential Sum Estimate

Using Weil's estimate, one has the following inequalities [5, 7, 52, 58]:

Theorem 3.2.13. *Given a prime p with $0 < d_1 < \cdots < d_k < p$, one has*

$$|\sum_{n=0}^{p-1}\chi[n+d_1]\cdots\chi[n+d_k]| \leq 9kp^{1/2}.$$

Theorem 3.2.14. *Given a prime p and $m, n \in \mathbb{Z}/p\mathbb{Z}\backslash\{0\}$, one has*

$$\left| \sum_{k \in \mathbb{Z}/p\mathbb{Z}} \chi[k]\chi[k+m]e^{-2\pi\imath kn/p} \right| \leq 2\sqrt{p}.$$

In particular, the sum (3.1) has the trivial estimate $\sqrt{p}|M_1||M_2|$. When $|M_1|, |M_2| \sim \sqrt{p}$, we will have that (3.1)$\leq p^{3/2}$.

In our case, the summation is three dimensional, complicating the issue. However, we shall show that if we add sufficiently large spins on the sum, there are indeed additional cancellations occurring.

3.3 Main Results

Theorem 3.3.1. *Let p be a prime, and $n \in \mathbb{Z}/p\mathbb{Z}$. Suppose $n \sim p^{1/2+\delta}$, where $\delta \in (0, 1/2)$, and $M_1, M_2 \subset \mathbb{Z}/p\mathbb{Z}$ consist of consecutive numbers such that $|M_1|, |M_2| \leq \sqrt{p}$. Furthermore, if $|M_2|/|M_1|, |M_1|$ are even, and $|M_1| \sim p^{1/2-\sigma}$, $\sigma \in [0, 1/2)$ such that $\delta > \sigma$, then*

$$\sum_{s \neq 0, -n} \left|\frac{\sin(\pi|M_1|s/p)}{\sin(\pi s/p)}\right| \left|\frac{\sin(\pi|M_2|(s+n)/p)}{\sin(\pi(s+n)/p)}\right| = O(p^{3/2-\alpha}), \tag{3.7}$$

where $\alpha = \sigma + (\delta - \sigma)/2$, and the big-O notation $A(p) = O(p^{3/2-\alpha})$ means that there exists a constant K, independent of p, such that $\limsup_{p:prime} \frac{A(p)}{p^{3/2-\alpha}} \leq K$.

From this theorem, we derive the following corollaries:

Corollary 3.3.2. *With the assumptions above, we have*

$$|\sum_{k}\sum_{m_1\in\Omega_1(n_1)}\sum_{m_2\in\Omega_2(n_2)}\chi[k+m_1-m_2]\chi[k]e^{2\pi\imath kn/p}e^{-2\pi\imath m_2n/p}| = O(p^{3/2-\alpha}), \quad (3.8)$$

where $n = n_1 - n_2$.

Corollary 3.3.3. *With the same assumptions above, we have, for a fixed* $k \in \mathbb{Z}/p\mathbb{Z}$,

$$|\sum_{m_1\in\Omega_1(n_1)}\sum_{m_2\in\Omega_2(n_2)}\chi[k+m_1-m_2]e^{2\pi\imath m_2n/p}| = O(p^{1-\alpha}).$$

Proof. of Corollary 3.3.2:

Given $n \in \mathbb{Z}/p\mathbb{Z}$, we compute

$$\begin{aligned}
&\sum_{k}\sum_{m_1\in\Omega_1(n_1),m_2\in\Omega_2(n_2)}\chi[k+m_1-m_2]\chi[k]e^{2\pi\imath kn/p}e^{-2\pi\imath m_2n/p}\\
&=\sum_{k,m_1,m_2}\left(\frac{1}{\sqrt{p}}\sum_{s}\chi[s]e^{2\pi\imath(k+m_1-m_2)s/p}\right)e^{-2\pi\imath m_2n/p}\chi[k]e^{2\pi\imath kn/p}\\
&=\sum_{s}\chi[s]\left(\frac{1}{\sqrt{p}}\sum_{k}\chi[k]e^{2\pi\imath k(n+s)/p}\right)\left(\sum_{m_1}e^{2\pi\imath m_1s/p}\right)\left(\sum_{m_2}e^{-2\pi\imath m_2(s+n)/p}\right)\\
&=\sum_{s}\chi[s]\chi[n+s]\left(\sum_{m_1}e^{2\pi\imath m_1s/p}\right)\left(\sum_{m_2}e^{-2\pi\imath m_2(s+n)/p}\right)\\
&=\sum_{s\neq 0,-n}\chi[s]\chi[n+s]\left(\sum_{m_1}e^{2\pi\imath m_1s/p}\right)\left(\sum_{m_2}e^{-2\pi\imath m_2(s+n)/p}\right).
\end{aligned}$$

Assuming $\Omega_1(n_1), \Omega_2(n_2)$ are both intervals in $\mathbb{Z}/p\mathbb{Z}$, we see that

$$|\sum_{m_j\in\Omega_j(n_j)}e^{2\pi\imath m_jt/p}| = |\frac{\sin(\pi|M_j|t/p)}{\sin(\pi t/p)}|,$$

where $j = 1, 2$. Thus, taking the absolute value on both sides, we get this estimate.

□

The proof of Corollary 3.3.3 follows verbatim.

Remark 3.3.4. Using Hölder's inequality and the Fourier transform of the Fejér's kernel, we can show that the expression in (3.7) is less than $p\sqrt{|\Omega_1(n_1)||\Omega_2(n_2)|}$, which equals $p^{3/2}$ when $|\Omega_1(n_1)| = |\Omega_2(n_2)| = \sqrt{p}$.

To prove Theorem 3.3.1, we will approximate $\sin(\pi|M_j|(s+t_j)/p)$ and $\sin(\pi(s+t_j)/p)$ with piece-wise linear functions. Then, by summing over all pieces, we shall show that the contribution as a whole is less than $p^{3/2-\alpha}$.

Definition 3.3.5. We define the following piece-wise polynomials $p_1^u, p_1^l, p_2^u, p_2^l$ as

$$\begin{cases} p_1^u(s) = 2\||M_1|s/p\|, & p_1^l(s) = \|s/p\|, \\ p_2^u(s) = 2\||M_2|(s+n)/p\|, & p_2^l(s) = \|(s+n)/p\|, \end{cases}$$

where $\|t\| := \min_{n\in\mathbb{Z}} |t-n|$.

Note that

$$|\frac{\sin(\pi|M_1|s/p)}{\sin(\pi s/p)}||\frac{\sin(\pi|M_2|(s+n)/p)}{\sin(\pi(s+n)/p)}| \leq \frac{p_1^u(s)p_2^u(s)}{p_1^l(s)p_2^l(s)}.$$

As we assume that $|M_2| \geq |M_1|$, the piece-wise linear function of $|\sin(\pi|M_2|(s+n)/p)|$ changes directions most frequently. Thus, we first start with the intervals in which the function does not change direction before expanding into larger intervals. In particular, we define the following intervals:

Definition 3.3.6. An interval in $\mathbb{Z}/p\mathbb{Z}$ with the form $[\frac{pj}{|M_2|} - n, \frac{p(j+1)}{|M_2|} - n]$, $j \in \{-|M_2|/2, \ldots, |M_2|/2\}$ is called an y_j-interval, by which we denote I_j^y.

An interval in $\mathbb{Z}/p\mathbb{Z}$ with the form $[\frac{pi}{|M_1|}, \frac{p(i+1)}{|M_1|}]$, $i \in \{-|M_1|/2, \ldots, |M_1|/2\}$ is called an x_j-interval, by which we denote I_i^x.

Here, we abuse the notation by denoting the set of numbers $\{a \in \mathbb{Z}/p\mathbb{Z} : a \in I\} \equiv I$ where $I \subset \mathbb{R}$ is an interval.

Given $s \in I_j^y \subset I_i^x$, we denote $x_i, y_j \in \mathbb{Z}$ by the integers such that $p_1^u(s) = |\frac{|M_1|s}{p} - x_i|$, $p_2^u(s) = |\frac{|M_2|(s+n)}{p} - y_j|$.

3.4 Proof of Theorem 3.3.1

In this section, we track only the main terms occurring during the calculation.

First, we see that

$$\left|\frac{\sin(\frac{\pi|M_1|s}{p})}{\sin(\frac{\pi s}{p})}\right| \cdot \left|\frac{\sin(\frac{\pi|M_2|(s+n)}{p})}{\sin(\frac{\pi(s+n)}{p})}\right| \leq \frac{4p^2}{\pi^2} \frac{|\frac{|M_1|s}{p} - x_i||\frac{|M_2|(s+n)}{p} - y_j|}{s(s+n)} = \frac{p_1^u(s)p_2^u(s)}{p_1^l(s)p_2^l(s)},$$

where $x_i = x_i(s) \in \{-\lceil\frac{|M_1|}{2}\rceil, -\lceil\frac{|M_1|}{2}\rceil+1, \ldots, \lceil\frac{|M_1|}{2}\rceil\} \cap 2\mathbb{Z}$, $y_j = y_j(s,n) \in \{\lfloor\frac{|M_2|n}{p}\rfloor, \ldots, \lceil\frac{|M_2|}{2} + \frac{|M_2|n}{p}\rceil\} \cap 2\mathbb{Z}$.

Note that $I_j^y \subset I_i^x \iff y_j \in [x_i|M_2|/|M_1| + |M_2|n/p, x_{i+1}|M_2|/|M_1| + |M_2|n/p - 1] =: J_i^x$. Then,

$$\sum_{s\neq 0,-n} |\frac{\sin(\pi|M_1|s/p)}{\sin(\pi s/p)}||\frac{\sin(\pi|M_2|(s+n)/p)}{\sin(\pi(s+n)/p)}| \leq \sum_{x_i=-|M_1|/2}^{|M_1|/2} \sum_{y_j\in J_i^x} \sum_{s\in I_j^y} \frac{p_1^u(s)p_2^u(s)}{p_1^l(s)p_2^l(s)}.$$

In our proof, we would like to smooth out $\{x_i\}_i, \{y_j\}_j$ by $\{z_i = i\}_i, \{w_j = j\}_j$ to simplify the approximation process. By doing so, we split the sum into the following parts:

$$\begin{aligned}\sum_{s\neq 0,n} \frac{p_1^u(s)p_2^u(s)}{p_1^l(s)p_2^l(s)} &= \sum_{|i|<p^\epsilon} \sum_{s\in I_i^x} \frac{p_1^u(s)p_2^u(s)}{p_1^l(s)p_2^l(s)} + \sum_{|i|>p^\epsilon} \sum_{j\in J_i^x} \sum_{s\in I_j^y} \frac{(-1)^{i+j}4p^2}{\pi^2} \frac{(\frac{|M_1|s}{p} - i)(\frac{|M_2|s}{p} - j)}{p_1^l(s)p_2^l(s)} \\ &+ \sum_{|i|>p^\epsilon} \sum_{j\in J_i^x : j \text{ odd}} \sum_{s\in I_j^y} \frac{(-1)^i 4p^2}{\pi^2} \frac{\frac{|M_1|s}{p} - i}{p_1^l(s)p_2^l(s)} + \sum_{|i|>p^\epsilon : i \text{ even}} \sum_{j\in J_i^x : j \text{ odd}} \sum_{s\in I_j^y} \frac{4p^2}{\pi^2 s(s+n)} \\ &=: E_1 + S + E_2 + E_3.\end{aligned} \tag{3.9}$$

We shall estimate on each of the four terms to show that (3.9) is of order $O(p^{3/2-\alpha})$.

Proposition 3.4.1. *We have the following estimates:*

(a)

$$E_1 = O(p^{3/2-\delta+\epsilon}).$$

(b)

$$E_2 + E_3 = O(p^{3/2-\delta}\log(p)).$$

(c)

$$S = O(p^{3/2-\sigma-\epsilon}) + O(p^{3/2-\delta}\log(p)).$$

With the estimates in Proposition 3.4.1, we can prove Theorem 3.3.1:

Proof. of Theorem 3.3.1:

From Proposition 3.4.1, we see that

$$\sum_{s\neq 0,n} \frac{p_1^u(s)p_2^u(s)}{p_1^l(s)p_2^l(s)} = O(p^{3/2-\sigma-\epsilon}) + O(p^{3/2-\delta}\log p) + O(p^{3/2-\delta+\epsilon}) = O(p^{3/2-\alpha_\epsilon})$$

where $\alpha_\epsilon = \min\{\epsilon+\sigma, \delta-\epsilon\}$. Since the choice of ϵ is arbitrary, we can optimize α to be $\sigma + (\delta-\sigma)/2$, which is what we claimed.

□

We first consider the case when s is positive. The case when s is negative is similar, and the proof for positive indices can be modified verbatim. We consider the term S in (3.9) to be the main term, while the rest are considered as correction terms. We shall first compute all three correction terms before dealing with the main term.

3.5 Estimates of Correction Terms

First, we shall prove Proposition 3.4.1 (a).

Proof. of Proposition 3.4.1 (a):

Assuming that $|x_i| \leq p^\epsilon$ and $|M_1| \sim p^{1/2-\sigma}$, $\sigma \in (0, 1/2)$, we have $|s| \leq \frac{px_i}{|M_1|} \sim p^{1/2+\epsilon+\sigma}$. Note that $n \sim p^{1/2+\delta}$ where $\delta > \epsilon+\sigma$, $\delta \in (0, 1/2)$. Thus,

$$\sum_{|s|\leq p^{1/2+\epsilon+\sigma}} |\frac{\sin(\pi|M_1|s/p)}{\sin(\pi s/p)}||\frac{\sin(\pi|M_2|(s+n)/p)}{\sin(\pi(s+n)/p)}| \leq |M_1| \sum_{|s|\leq p^{1/2+\epsilon+\sigma}} \frac{p}{\pi(s+n)}$$
$$\leq p|M_1| \log(\frac{n+p^{1/2+\epsilon+\sigma}}{n-p^{1/2+\epsilon+\sigma}})$$
$$= p|M_1| \log(1+\frac{1}{np^{-1/2-\epsilon-\sigma}-1})$$
$$\sim p|M_1|p^{-\delta+\epsilon+\sigma} \sim p^{3/2-\delta+\epsilon}.$$

(3.10)

Around the singular point $s=-n$, we make sure to take out an even number of y_j-intervals so the cancellations still occur in the remaining x_i-interval. Thus, the summation range is $|s+n| \leq \frac{kp}{|M_2|}$ for some $k \in \mathbb{N}$. Then,

$$\sum_{|s+n|\leq \frac{p}{|M_2|}} |\frac{\sin(\pi|M_1|s/p)}{\sin(\pi s/p)}||\frac{\sin(\pi|M_2|(s+n)/p)}{\sin(\pi(s+n)/p)}| \leq |M_2| \sum_{|s+n|\leq p^{1/2+\epsilon+\sigma}} \frac{p}{\pi|s|}$$
$$\leq p|M_2| \log(\frac{n+p/|M_2|}{n-p/|M_2|})$$
$$= p|M_2| \log(1+\frac{1}{n|M_2|/p-1})$$
$$\sim p^{2-1/2-\delta+\epsilon} = p^{3/2-\delta+\epsilon}.$$

(3.11)

□

To prove Proposition 3.4.1 (b), we need the following lemma:

Lemma 3.5.1. *Let $f, g : \mathbb{Z} \to \mathbb{R}$ be $f(s) = \frac{1}{s}$ and $f(s) = \frac{1}{s(s+t)}$ for some $t \in \mathbb{R}$. If*

$1 < a < a+1 < b$ *is such that* $\frac{b}{a} = 1 + r$ *for some* $r \in (0,1)$*, then*

$$\begin{cases} \sum_{a \le s \le b} f(s) &= r + O(r^2) + O(b^{-1}) \\ \sum_{a \le s \le b} g(s) &= \frac{r(r+t/a)}{t(1+r+t/a)} + O(\frac{r^2}{t}). \end{cases}$$

Proof. Since both f and g are monotone in (a,b), we may approximate the summation of both f and g with their respective integrals. Moreover,

$$|\sum_{a \le s \le b} \frac{1}{s} - \int_a^b \frac{1}{x} dx| \le \int_{b-1}^{b+1} \frac{1}{x} dx = \log(1 + \frac{2}{b-1}) = \frac{2}{b-1} + O(b^{-2}).$$

Thus,

$$\sum_{a \le s \le b} \frac{1}{s} = \log(1+r) + O(b^{-1}) = r + O(b^{-1}).$$

Note that $\frac{1}{b} < \frac{1}{a} < \frac{b-a}{a} = r$, so $\sum_{a \le s \le b} f(s) = O(r)$. For g, we have

$$\begin{aligned} |\sum_{a \le s \le b} \frac{1}{s(s+t)} - \int_a^b \frac{1}{s(s+t)} ds| &\le \int_{b-1}^{b+1} \frac{1}{s(s+t)} ds \\ &= \frac{1}{t} \log\left(\frac{(b+1)(b+t-1)}{(b-1)(b+t+1)}\right) \\ &= \frac{1}{t} \log\left((1 + \frac{2}{b-1})(1 - \frac{2}{b+t+1})\right) \\ &= \frac{1}{t}(\frac{2}{b-1} - \frac{2}{b+t+1}) + O(\frac{1}{tb^2}). \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{a \le s \le b} \frac{1}{s(s+t)} &= \frac{1}{t} \log\left((1+r)(1 - \frac{(b-a)}{b+t})\right) \\ &= \frac{r}{t}(1 - \frac{1}{(1+r)+t/a}) + O(\frac{r^2}{t}) = \frac{r(r+t/a)}{t(1+r+t/a)} + O(\frac{r^2}{t}). \end{aligned}$$

□

Proof. of Proposition 3.4.1 (b):

Fixing i, consider

$$\begin{aligned}
E_2 &= p^2 \sum_{(2k+1)\in J_i^x} \sum_{s\in I_j^y} \frac{\frac{|M_1|s}{p} - i}{s(s+n)} \\
&\leq \sum_{s\in I_i^x} \frac{\frac{|M_1|s}{p} - i}{s(s+n)} \\
&= \sum_{s\in I_i^x} p|M_1| \frac{1}{s+n} - \frac{p^2 i}{s(s+n)} \\
&= \sum_{s\in I_i^x} p|M_1| \frac{1}{s+n} - \frac{p^2 i}{n}(\frac{1}{s} - \frac{1}{s+n}) \\
&= p|M_1| \frac{\frac{p}{|M_1|}}{\frac{pi}{|M_1|}+n} + O(p|M_1|r^2) + O(|M_1|^2 i^{-1}) \\
&- \frac{p^2 i}{n}\left(\frac{1}{i} - \frac{\frac{p}{|M_1|}}{\frac{pi}{|M_1|}+n}\right) + O(p^2 n^{-1} x_i^{-1}) \\
&= p|M_1| \frac{1}{i + \frac{|M_1|n}{p}} - \frac{p^2 i}{n} \frac{\frac{|M_1|n}{p}}{i(i+\frac{|M_1|n}{p})} + O(p^2 n^{-1} x_i^{-1}) \\
&= O(p^2 n^{-1} i^{-1}).
\end{aligned} \tag{3.12}$$

As for E_3, by Lemma 3.5.1,

$$\begin{aligned}
E_3 &= p^2 \sum_{i\in 2\mathbb{Z}+1: i<|M_1|/2} \sum_{(2k+1)\in J_i^x} \sum_{s\in I_j^y} \frac{1}{s(s+n)} \\
&\leq \sum_{i<|M_1|/2} p^2 \sum_{s\in I_i^x} \frac{1}{s(s+n)} \\
&\leq \sum_{i<|M_1|/2} \frac{p^2}{ni} + O(p^2 n^{-1} i^{-2}) = O(p^2 n^{-1} \log(|M_1|)).
\end{aligned} \tag{3.13}$$

Combining (3.12) and (3.13), we see that the total contribution is $p^{2-(1/2+\delta)}\log(|M_1|) = O(p^{3/2-\delta}\log(p))$. □

3.6 Estimates of the Main Term S

To estimate S in (3.9), we start by computing the expression of the sum in one y_j-interval.

3.6.1 Estimates within y_j-Intervals

Lemma 3.6.1. *Given $j > 0$, define the error term $E_y(j)$ vt*

$$E_y(j) := \sum_{s\in I_j^y} \frac{4p^2}{\pi^2}\frac{(\frac{|M_1|s}{p} - i)(\frac{|M_2|(s+n)}{p} - j)}{s(s+n)} - \left(\frac{-2p^3 i}{\pi^2|M_2|\tilde{y}_j^2} + \frac{2np^2 i}{\tilde{y}_j^2 j} + \frac{2p|M_1|}{\pi^2 j}\right),$$

where $\tilde{y}_j = \frac{pj}{|M_2|} - n$, and α is as defined in Theorem 3.3.1.

Then, the total contribution of E_y is

$$\sum_{|i|>p^\epsilon}\sum_{j\in J_i^x} E_y(j) = O(p^{3/2-\alpha}).$$

To estimate $E_y(j)$, we will make the following re-arrangements first.

For $s \in I_j^y$ all $p_1^u, p_2^u, p_1^l, p_2^l$ are linear and none changes sign. Thus,

$$\sum_{s\in I_j^y}\frac{p_1^u(s)p_2^u(s)}{p_1^l(s)p_2^l(s)} = \pm\sum_{s\in I_j^y}\frac{4p^2}{\pi^2}\frac{(\frac{|M_1|s}{p} - i)(\frac{|M_2|(s+n)}{p} - j)}{s(s+n)}.$$

Thus, we would like to compute

$$\sum_{s\in I_j^y} \frac{4p^2}{\pi^2} \frac{(\frac{|M_1|s}{p} - i)(\frac{|M_2|(s+n)}{p} - j)}{s(s+n)}$$

$$= \sum_{\frac{pj}{|M_2|}-n \le s \le \frac{pj+1}{|M_2|}-n} \left\{ \frac{4}{\pi^2}|M_1||M_2| - \frac{4p|M_1|j}{\pi^2}\frac{1}{s+n} - \frac{4p|M_2|i}{\pi^2}\frac{1}{s} + \frac{4p^2}{\pi^2}\frac{ij}{s(s+n)} \right\}$$

$$= \frac{4|M_1||M_2|}{\pi^2}(\tilde{y}_{j+1} - \tilde{y}_j + f_1(j)) - \frac{4p|M_1|j}{\pi^2}(\log(\frac{j+1}{j}) + f_2(j)) - \frac{4p|M_2|i}{\pi^2}(\log(\frac{\tilde{y}_{j+1}}{\tilde{y}_j}) + f_3(j))$$

$$+ \frac{4p^2ij}{n\pi^2}\left(\log(\frac{\tilde{y}_{j+1}}{\tilde{y}_j}) - \log(\frac{\tilde{y}_{j+1}+n}{\tilde{y}_j+n}) + f_4(j) \right),$$

where we recall that $\tilde{y}_j = \frac{py_j}{|M_2|} - n$.

Note that $\tilde{y}_{j+1} - \tilde{y}_j = \frac{p}{|M_2|}$, and also $\log(\frac{j+1}{j}) = \log(1+\frac{1}{j}) = \frac{1}{j} - \frac{1}{2j^2} + O(j^{-3})$.

Thus,

$$\frac{4|M_1||M_2|}{\pi^2}(\tilde{y}_{j+1} - \tilde{y}_j) - \frac{4p|M_1|j}{\pi^2}\log(\frac{j+1}{j}) = \frac{2p|M_1|}{\pi^2 j} + O(|M_1|py_j^{-2}).$$

Now,

$$\log(\frac{\tilde{y}_{j+1}}{\tilde{y}_j}) = \log(1 + \frac{\frac{p}{|M_2|}}{\tilde{y}_j}) = \frac{p}{|M_2|}\frac{1}{\frac{pj}{|M_2|}-n} - \frac{p^2}{2|M_2|^2}\frac{1}{\tilde{y}_j^2} + O(p^3|M_2|^{-3}\tilde{y}_j^{-3})$$

$$= \frac{p}{|M_2|}\frac{1}{\frac{pj}{|M_2|}-n} - \frac{p^2}{2|M_2|^2}\frac{1}{\tilde{y}_j^2} + \frac{p^3}{3|M_2|^3}\frac{1}{\tilde{y}_j^3} + O(p^4|M_2|^{-4}\tilde{y}_j^{-4}),$$

and

$$\begin{aligned} \log(\frac{\tilde{y}_{j+1}+n}{\tilde{y}_j+n}) = \log(\frac{j+1}{j}) &= \frac{1}{j} - \frac{1}{2j^2} + O(j^{-3}) \\ &= \frac{1}{j} - \frac{1}{2j^2} + \frac{1}{3j^3} + O(j^{-4}). \end{aligned} \tag{3.14}$$

Thus, we see that, using (3.14),

$$\begin{aligned}
&-\frac{4p|M_2|i}{\pi^2}\frac{p}{|M_2|}\frac{1}{\frac{pj}{|M_2|}-n}+\frac{4p^2ij}{n\pi^2}\frac{p}{|M_2|}\frac{1}{\frac{pj}{|M_2|}-n}\\
&=\frac{-4p^2j}{\pi^2}\frac{1}{\frac{pj}{|M_2|}-n}+\frac{4p^2i}{n\pi^2}+\frac{4p^2i}{\pi}\frac{1}{\frac{pj}{|M_2|}-n}\\
&=\frac{4p^2i}{n\pi^2}.
\end{aligned}$$

Combining all of the above, we get that

$$\begin{aligned}
&\frac{4|M_1||M_2|}{\pi^2}(\tilde{y}_{j+1}-\tilde{y}_j+f_1(j))-\frac{4p|M_1|j}{\pi^2}(\log(\frac{j+1}{j})+f_2(j))-\frac{4p|M_2|i}{\pi^2}(\log(\frac{\tilde{y}_{j+1}}{\tilde{y}_j})+f_3(j))\\
&+\frac{4p^2ij}{n\pi^2}\left(\log(\frac{\tilde{y}_{j+1}}{\tilde{y}_j})-\log(\frac{\tilde{y}_{j+1}+n}{\tilde{y}_j+n})+f_4(j)\right)\\
&=\frac{4|M_1||M_2|}{\pi^2}f_1(j)+\frac{2p|M_1|}{\pi^2 j}+O(|M_1|pj^{-2})-\frac{4p|M_1|j}{\pi^2}f_2(j)+\frac{2p^3i}{\pi^2|M_2|\tilde{y}_j^2}-\frac{4p|M_2|i}{\pi^2}f_3(j)\\
&+O(|M_3|^{-2}p^4\tilde{y}_j^{-3})+\frac{4p^2i}{n\pi^2}-\frac{2p^4ij}{n|M_2|^2\pi^2\tilde{y}_j^2}-\frac{4p^2i}{n\pi^2}+\frac{2p^2i}{n\pi^2 j}+\frac{4p^2ij}{n\pi^2}f_4(j)+\frac{4p^2ij}{3n\pi^2}f_4'(j)\\
&+O(n^{-1}p^6|M_2|^{-4}ij\tilde{y}_j^{-4})+O(n^{-1}p^2ij^{-3})\\
&=-\frac{2p^4ij}{n|M_2|^2\pi^2\tilde{y}_j^2}+\frac{2p^2x_i}{n\pi^2y_j}+\frac{2p|M_1|}{\pi^2y_j}+\frac{2p^3x_i}{\pi^2|M_2|\tilde{y}_j^2}\\
&+\frac{4|M_1||M_2|}{\pi^2}f_1(j)-\frac{4p|M_1|j}{\pi^2}f_2(j)-\frac{4p|M_2|i}{\pi^2}f_3(j)+\frac{4p^2ij}{n\pi^2}f_4(j)+\frac{4p^2ij}{3n\pi^2}f_4'(j)\\
&+O(|M_1|pj^{-2})+O(|M_2|^{-2}p^4\tilde{y}_j^{-3})+O(n^{-1}p^6|M_2|^{-4}ij\tilde{y}_j^{-4})+O(n^{-1}p^2ij^{-3}),
\end{aligned}\tag{3.15}$$

where

$$\begin{aligned}\frac{4p^2ij}{3n\pi^2}f_4'(j) &= \frac{4p^2ij}{3n\pi^2}\left(\frac{p^3}{|M_2|^3}\frac{1}{\tilde{y}_j^3}-\frac{1}{j^3}\right)\\ &= \frac{4p^2ij}{3n\pi^2}\left(\frac{3\frac{p^2}{|M_2|^2}j^2n-3\frac{p}{|M_2|}jn^2+n^3}{\tilde{y}_j^3j^3}\right)\\ &= O(p^4|M_2|^{-2}ij^{-1}\tilde{y}_j^{-3})+O(p^3|M_2|^{-1}nij^{-2}\tilde{y}_j^{-3})+O(p^2n^2ij^{-3}\tilde{y}_j^{-3}).\end{aligned}$$

In (3.15), we have four explicit terms remaining, namely

$$\frac{-2p^4ij}{n\pi^2|M_2|^2\tilde{y}_j^2}+\frac{2p^2i}{n\pi^2j}+\frac{2p|M_1|}{\pi^2j}+\frac{2p^3i}{\pi^2|M_2|\tilde{y}_j^2}. \tag{3.16}$$

Further simplifying the expressions, we have

$$\begin{aligned}\frac{-2p^4ij}{n\pi^2|M_2|^2\tilde{y}_j^2}+\frac{2p^2i}{n\pi^2j}+\frac{2p^3i}{\pi^2|M_2|\tilde{y}_j^2} &= \frac{2p^2}{n\pi^2}\left[\frac{\tilde{y}_j^2-\frac{p^2}{|M_2|^2}j^2}{\tilde{y}_j^2j}\right]+\frac{2p^3i}{\pi^2|M_2|\tilde{y}_j^2}\\ &= \frac{2p^2}{n\pi^2}\frac{-\frac{2p}{|M_2|}nj+n^2}{\tilde{y}_j^2j}+\frac{2p^3i}{\pi^2|M_2|\tilde{y}_j^2}\\ &= \frac{-4p^3i}{\pi^2|M_2|\tilde{y}_j^2}+\frac{2np^2i}{\tilde{y}_j^2j}+\frac{2p^3i}{\pi^2|M_2|\tilde{y}_j^2}\\ &= \frac{-2p^3i}{\pi^2|M_2|\tilde{y}_j^2}+\frac{2np^2i}{\tilde{y}_j^2j}.\end{aligned}$$

To this point, we have computed all the main terms, and we have

$$\begin{aligned}E(j) =& \frac{4|M_1||M_2|}{\pi^2}f_1(j)-\frac{4p|M_1|j}{\pi^2}f_2(j)-\frac{4p|M_2|i}{\pi^2}f_3(j)+\frac{4p^2ij}{n\pi^2}f_4(j)+\frac{4p^2ij}{3n\pi^2}f_4'(j)\\ &+O(|M_1|pj^{-2})+O(|M_2|^{-2}p^4\tilde{y}_j^{-3})+O(n^{-1}p^6|M_2|^{-4}ij\tilde{y}_j^{-4})+O(n^{-1}p^2ij^{-3}).\end{aligned} \tag{3.17}$$

To estimate the effect of f_1, f_2, f_3, f_4, f_4', we refer to the following proposition which shall be proved in Section 3.7.

Proposition 3.6.2. *The following estimates hold:*

(a)

$$\sum_{|i|>p^\epsilon}\sum_{y_j\in J_i^x}\frac{4p^2ij}{3n\pi^2}|f_4'(j)| = \sum_{|i|>p^\epsilon} O(p|M_1|i^{-3}) = O(p^{3/2-\sigma-2\epsilon}).$$

(b)

$$\sum_{i=-|M_1|/2}^{|M_1|/2}\sum_{j\in J_i^x}\frac{4p|M_2|i}{\pi^2}|f_3(j)| + \frac{4p^2ij}{n\pi^2}|f_4(j)| = O(|M_1||M_2|\log p).$$

(c)

$$\sum_{y_j=1}^{|M_2|}\frac{4p|M_1|j}{\pi^2}f_2(j) = O(|M_1||M_2|\log|M_2|).$$

(d) $\sum_{x\geq p^\epsilon}\sum_{y\in J_i^x}|M_1||M_2|f_1(j) = O(|M_2|^2) = O(p)$ *if* $|M_1|$ *is even.*

Proposition 3.6.2 shows that the first five terms in (3.17) sums up to be of the order $O(p^{3/2-\sigma-2\epsilon})$. Thus, it remains to show that the final four terms in (3.17) can be well controlled.

Proof. of Lemma 3.6.1: Note that $p|M_2|^-1\tilde{y}_j^{-1} = \frac{1}{j-\frac{n|M_2|}{p}} =: \frac{1}{j-t}$. Thus,

- $$\sum_{|i|>p^\epsilon}\sum_{j\in J_i^x}p|M_1|j^{-2} = p|M_1|\sum_{|j|>\frac{|M_2|}{|M_1|}p^\epsilon}j^{-2} = O(p^{3/2-\sigma-\epsilon}).$$

- $$\sum_{|i|>p^\epsilon}\sum_{j\in J_i^x}p^4|M_2|^{-2}\tilde{y}_j^{-3} = p|M_2|^{-1}\sum_{|j|>\frac{|M_2|}{|M_1|}p^\epsilon}\frac{1}{(j-t)^3} = O(p^{1/2}).$$

- $$\begin{aligned}\sum_{|i|>p^\epsilon}\sum_{j\in J_i^x} n^{-1}p^6|M_2|^{-4}ij\tilde{y}_j^{-4} &= \frac{p^2}{n}\sum_{|i|>p^\epsilon}\sum_{j\in J_i^x} i\left(\frac{1}{(j-t)^4}+\frac{t}{(j-t)^3}\right)\\ &= \frac{p^2}{n}\sum_{|i|>p^\epsilon}\sum_{k=\frac{|M_2|i}{|M_1|}}^{\frac{|M_2|(i+1)}{|M_1|}} i\left(\frac{1}{k^4}+\frac{t}{k^3}\right)\\ &\le \frac{p^2}{n}\sum_{|i|>p^\epsilon}\frac{3|M_1|^3}{|M_2|^3}\frac{i((i+1)^3-i^3)}{i^3(i+1)^3}+\frac{2|M_1|^2}{|M_2|^2}\frac{it((i+1)^2-i^2)}{i^2(i+1)^2}\\ &\le \frac{20p^2}{n}\frac{|M_1|^3}{|M_2|^3}p^{-2\epsilon}+\frac{|M_1|^2}{|M_2|^2}\frac{n|M_2|}{p}p^{-\epsilon}\\ &= O(\frac{20p^2}{n}(p^{-2\epsilon}+p^{\delta-\sigma-\epsilon})) = O(p^{3/2-\sigma-\epsilon}).\end{aligned}$$

- $$\begin{aligned}\sum_{|i|>p^\epsilon}\sum_{j\in J_i^x} n^{-1}p^2ij^{-3} &\le \frac{p^2}{n}\sum_{|i|>p^\epsilon}\frac{|M_1|^2 i}{|M_2|^2}\left(\frac{1}{i^2}-\frac{1}{(i+1)^2}\right)\\ &= \frac{p^2}{n}\sum_{|i|>p^\epsilon}\frac{|M_1|^2}{|M_2|^2}\frac{2i+1}{i(i+1)^2}\\ &= O(p^{3/2-\delta-\epsilon}).\end{aligned}$$

Combining all the terms above, we see that $\sum_{|i|>p^\epsilon}\sum_{j\in J_i^x} E(j) = O(p^{3/2-\sigma-\epsilon})$. Choosing $\epsilon = (\delta-\sigma)/2$, we see that it is indeed of the order $p^{3/2-\alpha}$. □

3.6.2 Estimates within x_i-Intervals

Within a given I_i^x, p_1^u, p_1^l, p_2^l do not change signs, but p_2^u does between I_j^y and I_{j+1}^y. Thus, the main terms in Lemma 3.6.1 flip signs across different y_j-intervals.

Note that between consecutive y_j-intervals, either $y_{j+1} = y_j$ or $y_{j+1} = y_j + 2$

by construction. Moreover, $y_0 = x_0 = 0$. In this section, we replace $\{y_j\}_j$ by $\{z_j\}_j$ where $z_j = j$. Then, we have

$$y_j - z_j = \begin{cases} 1 & \text{if} \quad j \in 2\mathbb{Z}+1 \\ 0 & \text{if} \quad j \in 2\mathbb{Z}. \end{cases}$$

In particular, we may split the sum into

$$\sum_{j\in J_i^x}\sum_{s\in I_j^y}\frac{p_1^u(s)p_2^u(s)}{p_1^l(s)p_2^l(s)} = \sum_{j\in J_i^x}(-1)^j F(j) + \frac{4p^2}{\pi^2}\sum_{(2k+1)\in J_i^x}\sum_{s\in I_j^y}\frac{\frac{|M_1|s}{p}-x_j}{s(s+n)},$$

where

$$F(j) = \sum_{s\in I_j^y}\frac{(\frac{|M_1|x}{p}-x_j)(\frac{|M_2|(s+n)}{p}-j)}{s(s+n)}.$$

We shall derive the following estimate from Lemma 3.5.1:

Lemma 3.6.3. *The contribution of the correction term satisfies*

$$\frac{4p^2}{\pi^2}\sum_{i>p^\epsilon}\sum_{(2k+1)\in J_i^x}\sum_{s\in I_j^y}\frac{(-1)^i(\frac{|M_1|s}{p}-x_j)}{s(s+n)} = O(p^{3/2-\delta}\log(p)).$$

Proof. Fixing i, consider the inner sum

$$\begin{aligned}
p^2 \sum_{(2k+1)\in J_i^x} \sum_{s\in I_j^y} \frac{\frac{|M_1|s}{p} - i}{s(s+n)} &\le \sum_{s\in I_i^x} \frac{\frac{|M_1|s}{p} - i}{s(s+n)} \\
&= \sum_{s\in I_i^x} p|M_1| \frac{1}{s+n} - \frac{p^2 i}{s(s+n)} \\
&= \sum_{s\in I_i^x} p|M_1| \frac{1}{s+n} - \frac{p^2 i}{n}\left(\frac{1}{s} - \frac{1}{s+n}\right) \\
&= p|M_1| \frac{\frac{p}{|M_1|}}{\frac{pi}{|M_1|}+n} + O(p|M_1|r^2) + O(|M_1|^2 i^{-1}) \\
&- \frac{p^2 i}{n}\left(\frac{1}{i} - \frac{\frac{p}{|M_1|}}{\frac{pi}{|M_1|}+n}\right) + O(p^2 n^{-1} x_i^{-1}) \\
&= p|M_1| \frac{1}{i + \frac{|M_1|n}{p}} - \frac{p^2 i}{n} \frac{\frac{|M_1|n}{p}}{i(i+\frac{|M_1|n}{p})} + O(p^2 n^{-1} x_i^{-1}) \\
&= O(p^2 n^{-1} i^{-1}).
\end{aligned} \tag{3.18}$$

Again, we approximated $\{x_i\}_i$ by $\{w_i = i\}_i$. The contribution of the difference is, by Lemma 3.5.1,

$$\begin{aligned}
&p^2 \sum_{i\in 2\mathbb{Z}+1 : i<|M_1|/2} \sum_{(2k+1)\in J_i^x} \sum_{s\in I_j^y} \frac{1}{s(s+n)} \\
&\le \sum_{i<|M_1|/2} p^2 \sum_{s\in I_i^x} \frac{1}{s(s+n)} \\
&\le \sum_{i<|M_1|/2} \frac{p^2}{ni} + O(p^2 n^{-1} i^{-2}) = O(p^2 n^{-1} \log(|M_1|)).
\end{aligned} \tag{3.19}$$

Combining (3.18) and (3.19), we see that the total contribution is $p^{2-(1/2+\delta)} \log(|M_1|) \sim p^{3/2-\delta} \log(p)$. □

$|M_2|n/p$ will not be an integer unless $n = 0$. Suppose for now that $|M_1| | |M_2|$.

Then we see that there will be $|M_2|/|M_1| - 1$ complete y-intervals within. Also, the left and right incomplete y-intervals will combine to have the same length of a complete y-interval.

Define $g_1(y_j) = \frac{-2p^3 x_i}{\pi^2 |M_2| \tilde{y}_j^2}, g_2(y_j) = \frac{2np^2 x_i}{\tilde{y}_j^2 y_j}, g_3(y_j) = \frac{2p|M_1|}{\pi^2 y_j}$. All three terms are decreasing with respect to y_j. Thus,

$$\sum_{y \in J_i^x} \sum_{s \in I_j^y} \frac{p_1^u(s) p_2^u(s)}{p_1^l(s) p_2^l(s)} \leq \sum_{l=1}^{3} | \sum_{y_j \in J_i^x} (-1)^{y_j} g_l(y_j)| + \sum_{y_j \in J_i^x} |E(y_j)|$$

consists of three alternating series.

Recal that $I_j^y \subset I_i^x \iff y_j \in [\frac{x_i |M_2|}{|M_1|} + \frac{|M_2| n}{p}, \frac{x_{i+1} |M_2|}{|M_1|} + \frac{|M_2| n}{p} - 1]$, and $\left|J_x^i\right| = |M_2|/|M_1|$. Thus, the case when $|M_2|/|M_1|$ is an even number will be superior to the one with odd numbers.

With the three terms carrying over, we need the following lemma:

Lemma 3.6.4. *Within an x_i-interval, the contribution is*

$$\sum_{s \in I_i^x} \frac{p_1^u(s) p_2^u(s)}{p_1^l(s) p_2^l(s)} = O(p|M_1| i^{-2}) + O(p^{3/2-\delta} i^{-1}) + \sum_{y \in J_i^x} E_y(j).$$

Proof. First, note that

$$\left| \sum_{j=\frac{i|M_2|}{|M_1|}+\frac{|M_2|n}{p}+1}^{\frac{i+1|M_2|}{|M_1|}+\frac{|M_2|n}{p}} (-1)^j \left(\frac{-2p^3 i}{\pi^2 |M_2| \tilde{y}_j^2} + \frac{2np^2 i}{\tilde{y}_j^2 j} + \frac{2p|M_1|}{\pi^2 j} \right) \right|$$

$$\leq |\sum_{s \in I_i^x} \frac{p_1^u(s) p_2^u(s)}{p_1^l(s) p_2^l(s)}|$$

$$\leq \left| \sum_{j=\frac{i|M_2|}{|M_1|}+\frac{|M_2|n}{p}}^{\frac{i+1|M_2|}{|M_1|}+\frac{|M_2|n}{p}-1} (-1)^j \left(\frac{-2p^3 i}{\pi^2 |M_2| \tilde{y}_j^2} + \frac{2np^2 i}{\tilde{y}_j^2 j} + \frac{2p|M_1|}{\pi^2 j} \right) \right|.$$

Since $|M_2|/|M_1| \in 2\mathbb{N}$, we can see that, for $\frac{-2p^3 i}{\pi^2 |M_2| \tilde{y}_j^2}$,

$$\begin{aligned}
\sum_{j=\frac{i|M_2|}{|M_1|}+\frac{|M_2|n}{p}}^{\frac{i+1|M_2|}{|M_1|}+\frac{|M_2|n}{p}-1} \frac{(-1)^j}{\tilde{y}_j^2} &= \frac{|M_2|^2}{p^2} \sum_j \frac{(-1)^j}{(j - \frac{|M_2|n}{p})^2} \\
&= \frac{|M_2|^2}{p^2} \sum_{z_j=\frac{i|M_2|}{|M_1|}}^{\frac{i+1|M_2|}{|M_1|}} \frac{(-1)^j}{z_j^2} \\
&\leq \frac{|M_2|^2}{2p^2} \left[\left(\frac{|M_1|}{i|M_2|} - \frac{|M_1|}{(i+1)|M_2|} \right) - \left(\frac{1}{\frac{i|M_2|}{|M_1|}+1} - \frac{1}{\frac{(i+1)|M_1|}{|M_2|}+1} \right) \right] \\
&= \frac{|M_2|^2}{2p^2} \left[\frac{1}{\frac{i|M_2|}{|M_1|}(\frac{i|M_2|}{|M_1|}+1)} - \frac{1}{\frac{(i+1)|M_2|}{|M_1|}(\frac{(i+1)|M_2|}{|M_1|}+1)} \right] \\
&= \frac{|M_2|^2}{2p^2} \left[\frac{\frac{|M_2|}{|M_1|}(\frac{i|M_2|}{|M_1|}+\frac{i|M_2|}{|M_1|}+1)+\frac{|M_2|^2}{|M_1|^2}}{\frac{i|M_2|}{|M_1|}(\frac{i|M_2|}{|M_1|}+1)\frac{(i+1)|M_2|}{|M_1|}(\frac{(i+1)|M_2|}{|M_1|}+1)} \right] \\
&= O(p^{-2}|M_1|^2 i^{-3}).
\end{aligned}$$

For $\frac{2np^2 i}{\tilde{y}_j^2 j}$,

$$\sum_{j\in J_i^x}\frac{(-1)^j}{j\tilde{y}_j^2}=\sum_{j\in J_i^x}(-1)^j\left[\frac{Aj+B}{\tilde{y}_j^2}+\frac{C}{j}\right]$$

where A, B, C satisfy

$$C\tilde{y}_j^2+Aj^2+Bj=1 \implies A=\frac{-p^2}{|M_2|^2n^2},\quad B=\frac{2p}{n|M_2|},\quad C=\frac{1}{n^2}.$$

Thus, we have

$$\begin{aligned}
\sum_j\frac{(-1)^j}{j\tilde{y}_j^2}&=\sum_j(-1)^j\left[\frac{\frac{-p}{|M_2|n^2}\tilde{y}_j+\frac{p}{n|M_2|}}{\tilde{y}_j^2}+\frac{1}{n^2j}\right]\\
&=\sum_j(-1)^j\left[\frac{-p}{|M_2|n^2}\frac{1}{\tilde{y}_j}+\frac{1}{n^2j}+\frac{p}{n|M_2|}\frac{1}{\tilde{y}_j^2}\right]\\
&\sim\frac{|M_2|}{p}\frac{-p}{|M_2|n^2}\left[\log(\frac{i+1}{i})-\log(\frac{\frac{(i+1)|M_2|}{|M_1|}+1}{\frac{i|M_2|}{|M_1|}+1})\right]\\
&+\frac{1}{n^2}\left[\log(\frac{\frac{(i+1)|M_2|}{|M_1|}+\frac{|M_2|n}{p}}{\frac{i|M_2|}{|M_1|}+\frac{|M_2|n}{p}})-\log(\frac{\frac{(i+1)|M_2|}{|M_1|}+\frac{|M_2|n}{p}+1}{\frac{i|M_2|}{|M_1|}+\frac{|M_2|n}{p}+1})\right]+O(p^{-1}|M_1|^2|M_2|^{-1}n^{-1}i^{-3})\\
&\sim\frac{1}{n^2}\left[-\frac{1}{i}+\frac{1}{i+\frac{|M_1|}{|M_2|}}+\frac{1}{i+\frac{|M_1|n}{p}}-\frac{1}{i+\frac{|M_1|n}{p}+\frac{|M_1|}{|M_2|}}\right]+O(p^{-1}|M_1|^2|M_2|^{-1}n^{-1}i^{-3})\\
&=\frac{|M_1|}{n^2|M_2|}\left[\frac{(2i+\frac{|M_1|}{|M_2|})\frac{|M_1|n}{p}+\frac{|M_1|^2n^2}{p^2}}{i(i+\frac{|M_1|}{|M_2|})(i+\frac{|M_1|n}{p})(i+\frac{|M_1|n}{p}+\frac{|M_1|}{|M_2|})}\right]\\
&=\frac{|M_1|}{|M_2|n^2}\frac{O(\frac{i|M_1|n}{p})+O(\frac{|M_1|^2n^2}{p^2})}{O(i^4)+O(i^2\frac{|M_1|^2n^2}{p^2})}.
\end{aligned}$$

For $\frac{2p|M_1|}{\pi^2j}$, by letting $2a=\frac{i|M_2|}{|M_1|}+\frac{|M_2|n}{p}$, $2b=\frac{(i+1)|M_2|}{|M_1|}+\frac{|M_2|n}{p}$, and $t=1/2$, we

have

$$\begin{aligned}\sum_{j\in J_i^x}\frac{(-1)^j}{j} &= \sum_{(2k)\in J_i^x}(\frac{1}{2k}-\frac{1}{2k+1})\\ &= \sum_{(2k)\in J_i^x}\frac{1}{2k(2k+1)}\\ &= \frac{\frac{1}{i+\frac{|M_1|n}{p}}\left(\frac{1}{i+\frac{|M_1|n}{p}}+\frac{1}{\frac{i|M_2|}{|M_1|}+\frac{|M_2|n}{p}}\right)}{1+\frac{1}{i+\frac{|M_1|n}{p}}+\frac{1}{\frac{i|M_2|}{|M_1|}+\frac{|M_2|n}{p}}}+O(r^2)\\ &\leq \frac{1}{(i+\frac{|M_1|n}{p})^2}+O(r^2)=O(\frac{1}{i^2}).\end{aligned}$$

Combining the three terms, we see that

$$\begin{aligned}&\sum_{j\in J_i^x}(-1)^j\left(\frac{-2p^3i}{\pi^2|M_2|\tilde{y}_j^2}+\frac{2np^2i}{\tilde{y}_j^2j}+\frac{2p|M_1|}{\pi^2j}\right)\\ &= O(p|M_1|^2|M_2|^{-1}i^{-2})+\min\{O(i^{-2}p^{3/2-\sigma}),O(p^{3/2-2\delta+\sigma})\}\\ &+\min\{O(i^{-3}p^{3/2-2\sigma+\delta}),O(p^{3/2-\delta}i^{-1})\}+O(p|M_1|^2|M_2|^{-1}i^{-2})\\ &= O(p|M_1|i^{-2})+O(p^{3/2-\delta}i^{-1}).\end{aligned}$$

□

3.6.3 Proof of Proposition 3.4.1 (c)

Now, we are prepared to prove Proposition 3.4.1 (c).

Proof. of Proposition 3.4.1 (c):

For $|s| \geq p^{1/2+\epsilon+\sigma} \implies x_i \geq p^{\epsilon}$, we have

$$\sum_{x_i \geq p^{\epsilon}} O(p|M_1|x_i^{-2}) + O(p^{3/2-\delta}x_i^{-1}) + E(x_i) = O(p^{3/2-\sigma-\epsilon}) + O(p^{3/2-\delta}\log p).$$

Thus, adding the two parts, we get

$$O(p^{3/2-\sigma-\epsilon}) + O(p^{3/2-\delta}\log p) + O(p^{1/2-\delta+\epsilon}) = O(p^{3/2-\alpha_\epsilon}),$$

where $\alpha_\epsilon = \min\{\epsilon+\sigma, \delta-\epsilon\}$. Now, since ϵ is arbitrary, we can optimize α to be $\sigma + (\delta-\sigma)/2$.

For different components of $p_1^l(s), p_2^l(s)$, the same arguments work verbatim by re-enumerate the x_i and y_j-intervals, so the same estimate holds. Note that $n \sim p^{1/2+\delta}$ where $\delta \in (0, 1/2)$, so $p - n \sim p$.

□

3.7 Proof of Proposition 3.6.2

In this section, we show that the contributions from f_1, f_2, f_3, f_4, f_4' are all negligible. In increasing order of difficulty, we shall start with f_4' and end with f_1. The remaining error terms can be summed trivially over J_i^x and $\{i : i \geq p^{\epsilon}\}$, and the proof will be omitted. .

3.7.1 Estimates for f_4'

First, we note that

$$\begin{aligned}\frac{4p^2ij}{3n\pi^2}f_4'(j) &= \frac{4p^2ij}{3n\pi^2}\left(\frac{p^3}{|M_2|^3}\frac{1}{\tilde{y}_j^3}-\frac{1}{j^3}\right)\\ &= \frac{4p^2ij}{3n\pi^2}\left(\frac{3\frac{p^2}{|M_2|^2}j^2n-3\frac{p}{|M_2|}jn^2+n^3}{\tilde{y}_j^3j^3}\right)\\ &= O(p^4|M_2|^{-2}ij^{-1}\tilde{y}_j^{-3})+O(p^3|M_2|^{-1}nij^{-2}\tilde{y}_j^{-3})+O(p^2n^2ij^{-3}\tilde{y}_j^{-3}).\end{aligned} \tag{3.20}$$

Lemma 3.7.1. *For fixed integers $l,k>0$, one has*

$$\sum_{j\in J_i^x}\frac{1}{j^l\tilde{y}_j^k}=O(\min_{0\le s\le k}\{(|M_2|^{-s}|M_1|^si^{-s})\big(\frac{p}{|M_2|n}\big)^{l-s}\}p^{-k}i^{-k}|M_2|^1|M_1|^{k-1}),$$

where the constant depends on l,k.

With Lemma 3.7.1, we can prove Proposition 3.6.2 (a).

Proof. of Proposition 3.6.2 (a):

From (3.20), we can use Lemma 3.7.1, choosing the parameter s to be $0,1,2$ respectively for the three terms. Noting that $|M_1|\le|M_2|$, we get the desired estimate bound.

□

Proof. of Lemma 3.7.1:

$$\sum_{j\in J_i^x}\frac{1}{\tilde{y}_j^k}\sim\frac{|M_2|}{p}(\frac{|M_1|}{p})^{k-1}\left[\frac{1}{i^{k-1}}-\frac{1}{(i+1)^{k-1}}\right]$$
$$=\frac{|M_2||M_1|^{k-1}}{p^k}\frac{(i+1)^{k-1}-i^{k-1}}{(i(i+1))^{k-1}}$$
$$\sim\frac{|M_2||M_1|^{k-1}}{x_i^k p^k},$$

where we note that $x_{i+1}=x_i+1$. For the second equation, denoting $\frac{i|M_2|}{|M_1|}+\frac{|M_2|n}{p}$ by $\tilde{x}_i$, we have

$$\sum_{j\in J_i^x}\frac{1}{j^k}\sim\frac{(\tilde{x}_i+1)^{k-1}-\tilde{x}_i^{k-1}}{(\tilde{x}_i\tilde{x}_{i+1})^{k-1}}$$
$$\sim O(\min_{0\le s\le k}\{(|M_2|^{-s}|M_1|^s i^{-s})\big(\frac{p}{|M_2|n}\big)^{k-s}\}),$$

where we note that

$$\frac{1}{\tilde{x}_i}=O(\min\{\frac{|M_1|}{|M_2|i},|\frac{p}{|M_2|n}|\}).$$

Now, by Hölder's inequality, we can derive the result.

□

3.7.2 Estimates for f_3 and f_4

We are going to use the comparison lemma: If $f(x)$ is monotone, then

$$|\sum_{x=a}^{b}f(x)-\int_{a-1}^{b}f(t)\,dt|\le|\int_{a-1}^{b}f(t)\,dt-\int_{a}^{b+1}f(t)\,dt|.$$

Lemma 3.7.2. *The following statements are true:*

- $f_3(j) = O(|M_2|p^{-1}j^{-2})$,

- $f_4(j) = O(|M_2|^2p^{-2}j^{-3})$.

The constant of the big-O notation is independent of $|M_2|$ and p.

Proof. For f_3, we have that

$$\begin{aligned}
|f_3(j)| &\leq |\log(\frac{\tilde{y}_{j+1}}{\tilde{y}_j}) - \log(\frac{\tilde{y}_{j+1}+1}{\tilde{y}_j+1})| \\
&= |\log(\frac{\tilde{y}_{j+1}(\tilde{y}_{j+1}+1)}{\tilde{y}_j(\tilde{y}_{j+1}+1)})| \\
&= |\log(1+\frac{\tilde{y}_{j+1}-\tilde{y}_j}{\tilde{y}_j(\tilde{y}_{j+1}+1)})| \\
&= |\log(1+\frac{p}{|M_2|}\frac{1}{\frac{p^2j^2}{|M_2|^2}-2\frac{pjn}{|M_2|}+n^2+\frac{pj}{|M_2|}(\frac{p}{|M_2|}+1)})| \\
&= O(|M_2|p^{-1}j^{-2}).
\end{aligned}$$

For f_4, note that $\frac{1}{s(s+n)} = \frac{1}{n}(\frac{1}{s} - \frac{1}{s+n})$ is monotone.

$$
\begin{aligned}
|f_4(y_j)| &\le |\log(\frac{\tilde{y}_{j+1}(\tilde{y}_j+n)}{\tilde{y}_j(\tilde{y}_{j+1}+n)}) - \log(\frac{(\tilde{y}_{j+1}+1)(\tilde{y}_j+n+1)}{(\tilde{y}_j+1)(\tilde{y}_{j+1}+n+1)})| \\
&= |\log(1-\frac{1}{\tilde{y}_{j+1}+1}) + \log(1-\frac{1}{\tilde{y}_j+(n+1)}) - \log(1-\frac{1}{\tilde{y}_j+1}) - \log(1-\frac{1}{\tilde{y}_{j+1}+(n+1)})| \\
&= |(-\frac{1}{\tilde{y}_{j+1}+1} - \frac{1}{\tilde{y}_j+(n+1)} + \frac{1}{\tilde{y}_j+1} + \frac{1}{\tilde{y}_{j+1}+(n+1)}) \\
&+ \frac{1}{2}(-\frac{1}{(\tilde{y}_{j=1}+1)^2} - \frac{1}{(\tilde{y}_j+(n+1))^2} + \frac{1}{(\tilde{y}_j+1)^2} + \frac{1}{(\tilde{y}_{j+1}+(n+1))^2})| + O(\frac{1}{\tilde{y}_j^3}) \\
&= |\left(\frac{p/|M_2|}{(\tilde{y}_{j+1}+1)(\tilde{y}_j+1)} - \frac{p/|M_2|}{(\tilde{y}_j+(n+1))(\tilde{y}_{j+1}+(n+1))}\right) \\
&+ \frac{1}{2}\left(\frac{2\frac{p}{|M_2|}(\frac{pj}{|M_2|}+1-n)+\frac{p^2}{|M_2|^2}}{(\tilde{y}_j+1)^2(\tilde{y}_{j+1}+1)^2} - \frac{1}{2}\left(\frac{2\frac{p}{|M_2|}(\frac{pj}{|M_2|}+1)+\frac{p^2}{|M_2|^2}}{(\tilde{y}_j+1+n)^2(\tilde{y}_{j+1}+1+n)^2}\right)| + O(\tilde{y}_j^3)\right. \\
&= |\frac{p}{|M_2|}\frac{n(\tilde{y}_j+\tilde{y}_{j+1})+(n+1)^2-1}{(\tilde{y}_{j+1}+1)(\tilde{y}_j+1)(\tilde{y}_j+(n+1))(\tilde{y}_{j+1}+(n+1))}| \\
&\quad + O(|M_2|^2p^{-2}j^{-3}) + O(|M_2|^3p^{-3}j^{-3}) \\
&= O(|M_2|^2p^{-2}j^{-3}).
\end{aligned}
$$

□

Proof. of Proposition 3.6.2 (b):

Note that, by Lemma 3.7.2,

$$-\frac{4p|M_2|i}{\pi^2}f_3(j) + \frac{4p^2ij}{n\pi^2}f_4(j) = O(|M_2|^2ij^{-2}) + O(|M_2|^2n^{-1}ij^{-2}) = O(|M_2|^2ij^{-2}).$$

Now,

$$\begin{aligned}\sum_{j\in J_i^x}\frac{1}{j^2} &\sim \frac{1}{\frac{i|M_2|}{|M_1|}+\frac{|M_1|n}{p}}-\frac{1}{\frac{(i+1)|M_2|}{|M_1|}+\frac{|M_1|n}{p}}\\ &= \frac{|M_2|/|M_1|}{(\frac{i|M_2|}{|M_1|}+\frac{|M_1|n}{p})(\frac{(i+1)|M_2|}{|M_1|}+\frac{|M_1|n}{p})}\\ &= O(\frac{|M_1|}{|M_2|i^2}).\end{aligned} \tag{3.21}$$

Then, summing over all possible x_i, we see that

$$\sum_{x=1}^{|M_1|}|M_1||M_2|\frac{x}{x^2}\sim |M_1||M_2|\log p,$$

which concludes the proof. □

3.7.3 Estimates for f_2

Proof. of Proposition 3.6.2 (c):

Suppose $\{\frac{p}{|M_2|}\}=\delta$, where $\{x\}=x-\lfloor x\rfloor$. Let $\{\frac{py_j}{|M_2|}\}=1-\epsilon=1-\epsilon_j$, then for a given $t\in\mathbb{Z}$,

$$\begin{aligned}\frac{j}{t}-j\log(\frac{t+1-\epsilon}{t-\epsilon}) &= \frac{j}{t}-y_j\log(1+\frac{1}{t-\epsilon})\\ &= \frac{j}{t}-j(\frac{1}{t-\epsilon}-\frac{1}{2(t-\epsilon)^2}+O(t^{-3}))\\ &= \frac{-\epsilon j}{t(t-\epsilon)}+\frac{j}{2(t-\epsilon)^2}+O(t^{-3}j).\end{aligned}$$

Summing over t from $\lceil \frac{pj}{|M_2|} \rceil$ to $\lfloor \frac{py(j+1)}{|M_2|} \rfloor$, we have

$$\begin{aligned}
\sum_{t=\lceil \frac{pj}{|M_2|} \rceil}^{\lfloor \frac{p(j+1)}{|M_2|} \rfloor} \frac{-j\epsilon}{t(t-\epsilon)} &= \sum_{t=\lceil \frac{pj}{|M_2|} \rceil}^{\lfloor \frac{p(j+1)}{|M_2|} \rfloor} y_j \left(\frac{1}{t} - \frac{1}{t-\epsilon}\right) \\
&\sim j\left(\log \frac{\lceil \frac{p(j+1)}{|M_2|} \rceil}{\lceil \frac{pj}{|M_2|} \rceil} - \log \frac{\frac{p(j+1)}{|M_2|} + (1-2\epsilon-\delta)}{\frac{pj}{|M_2|}} \right) \\
&= j\left(\log \frac{(\frac{p(j+1)}{|M_2|} + (1-\epsilon-\delta))\frac{pj}{|M_2|}}{(\frac{pj}{|M_2|}+\epsilon)(\frac{p(j+1)}{|M_2|} + (1-2\epsilon-\delta))} \right) \\
&= j\left(\log \left(1 + \frac{\epsilon \frac{pj}{|M_2|} - \epsilon \frac{p(j+1)}{|M_2|} + O(1)}{(\frac{pj}{|M_2|}-\epsilon)(\frac{p(j+1)}{|M_2|} + (1-2\epsilon-\delta))} \right) \right) \\
&= j \frac{\epsilon \frac{pj}{|M_2|} - \epsilon \frac{p(j+1)}{|M_2|}}{(\frac{pj}{|M_2|}+\epsilon)(\frac{p(j+1)}{|M_2|} + (1-2\epsilon-\delta))} + O(\frac{|M_2|^2}{p^2} j^{-1}) \\
&= \frac{-\epsilon \frac{pj}{|M_2|}}{(\frac{pj}{|M_2|}+\epsilon)(\frac{p(j+1)}{|M_2|} + (1-2\epsilon-\delta))} + O(\frac{|M_2|^2}{p^2} j^{-1}) \\
&= O(|M_2| p^{-1} j^{-1}) + O(|M_2|^2 p^{-2} j^{-1}).
\end{aligned}$$

The other term can be obtained similarly. Now,

$$\sum_{j=1}^{|M_2|} \frac{4p|M_1|j}{\pi^2} f_2(j) = \sum_{j=1}^{|M_2|} O(|M_1||M_2| j^{-1}) = O(|M_1||M_2| \log |M_2|).$$

□

3.7.4 Estimates for f_1

Proof. of Proposition 3.6.2 (d):

Since $(p, |M_2|) = 1$, we see that the fractional part of $\{py_j/|M_2|\}_{j=1}^{|M_2|}$ runs through $\{k/|M_2|\}_{k=0}^{|M_2|-1}$.

We denote the fractional part of a number x by $\{x\} = x - \lfloor x \rfloor$. Let $\{p/|M_2|\} = \delta$, and $\{py_j/|M_2|\} = \epsilon_j$, then

$$f_1(y_j) = -(1-\epsilon_j) - \{\epsilon_j + \delta\} = \begin{cases} -1-\delta & \text{if} \quad \epsilon + \delta < 1 \\ -\delta & \text{if} \quad \epsilon + \delta \geq 1 \end{cases}$$

Without loss of generality, we may assume that $\delta \leq 1/2$. Since f_1 changes signs from one y_j-interval to another, it is important to identify where $|f_1|$ attains δ.

In order to do that, we first introduce the notion of the critical zone.

Definition 3.7.3. Given $\delta \leq 1/2$, the critical zone $\bar{A} \subset S^1$, the unit circle, is defined as $\bar{A} = [1-\delta, 1)$. The discrete counterpart $A \subset \mathbb{Z}/|M_2|\mathbb{Z}$ is $A = \{x \in \mathbb{Z}/|M_2|\mathbb{Z} : \frac{x}{|M_2|} \in \bar{A}\}$.

We should note that $\{pj/|M_2|\} \in A$ if and only if $|f_1(j)| = \delta$. Thus, the problem now depends on when $\{pj/|M_2|\}$ lies in A so as to account for cancellation.

Now, we note that there are effectively $|M_2|/|M_1|$ y_j-intervals within one x_i-interval. Also, the corresponding y_j-intervals in consecutive x_i-intervals have different signs. In particular, $y_{j+2|M_2|/|M_1|}$-interval and y_j-interval have the same sign. Since we assume that $|M_1|$ is even, $\frac{2|M_2|}{|M_1|}\mathbb{Z}/|M_2|\mathbb{Z}$ is an additive subgroup of order $|M_1|/2$. Also, for any given j, $\{[\{py_k/|M_2|\}]\}_{k=j}^{j+2|M_2|/|M_1|-1}$ are distinct representatives of the coset.

As p is a unit in $\mathbb{Z}/|M_2|\mathbb{Z}$, we can replace the representatives by $\{-k\}_{k=1}^{|M_1|/2}$. Also, we see that between each coset, the number of elements inside the critical zone

A differs by at most 1. Thus, the excessive parts that are not cancelled contribute at most $|M_2|/|M_1|$.

For the boundary contribution of one x_i-interval, we see that the incomplete sums on both sides combine to represent the coset $|M_2|/|M_1|$.

The argument above applies for summation over the whole group, but in our case we need to avoid the singularity at $-n$, which splits the summation range into 2 parts. Nonetheless, we shall show that the intuition still holds true even with segmented sums.

If $\gamma = \{p/|M_2|\} < p^{-1/2+\sigma}$, then $|f_1(j)| = \delta$ for at most p^σ times, so the contribution is $\sqrt{p}|M_2| = O(p)$.

First, when we split the summation range into 2 parts, note that since the complete summation gives at most the order of $|M_2|/|M_1|$, it suffices to estimate for one part and get the estimate of the other part by subtraction.

As it suffices to estimate for the range $-n \leq s \leq p/2$, we are looking at the following quantity

$$I = \sum_{a \leq t \leq b} \sum_{j=0}^{\frac{2|M_2|}{|M_1|}-1} (-1)^j g\left[pj + \frac{tp|M_2|}{|M_1|}\right],$$

where $|b-a| = O(|M_1|)$, and $g = \mathbb{1}_A : \mathbb{Z}/|M_2|\mathbb{Z} \to \mathbb{R}$ is the characteristic function of A. Moreover, $|A| \sim \delta|M_2|$.

Now,

$$
\begin{aligned}
I &= \sum_{a\le t\le b}\sum_{j=0}^{\frac{2|M_2|}{|M_1|}-1}(-1)^j g[pj+\frac{tp|M_2|}{|M_1|}] \\
&= \frac{1}{\sqrt{|M_2|}}\sum_j(-1)^j\sum_{a\le t\le b}\sum_{k\in\mathbb{Z}/|M_2|\mathbb{Z}}\hat{g}[k]e^{-2\pi\imath tpk/|M_1|}e^{2\pi\imath kpy_j/|M_2|} \\
&= \frac{1}{\sqrt{|M_2|}}\sum_k\hat{g}[k]\Big(\sum_{a\le t\le b}e^{-2\pi\imath tpk/|M_1|}\Big)\Big(\sum_{y_j}(-1)^{y_j}e^{2\pi\imath kpy_j/|M_2|}\Big) \\
&= \frac{1}{|M_2|}\sum_k\bar{C}_k\frac{\sin(\pi k|A|/|M_2|)}{\sin(\pi k/|M_2|)}\frac{\sin(\pi k(b-a+1)p/|M_1|)}{\sin(\pi pk/|M_1|)}\frac{\sin(2\pi kp/|M_1|)}{\sin(2\pi kp/|M_2|)}\sin(\pi kp/|M_2|),
\end{aligned}
$$

where $|\bar{C}_k|=1$ for all k. Thus, by Hölder's inequality, the identity formula of the Fejér kernel, and change of variables ($kp\mapsto l$), we see that

$$
\begin{aligned}
|I| &\le \frac{1}{|M_2|}\Big(\sum_k|\frac{\sin(\pi k|A|/|M_2|)}{\sin(\pi k/|M_2|)}|^2\Big)^{1/2}\cdot \\
&\quad\Big(\sum_{l\in\mathbb{Z}/|M_2|\mathbb{Z}}\Big|\frac{\sin(\pi l(b-a+1)/|M_1|)}{\sin(\pi l/|M_1|)}\Big|^2\Big|\frac{\sin(2\pi l/|M_1|)}{\sin(2\pi l/|M_2|)}\sin(\pi l/|M_1|)\Big|^2\Big)^{1/2} \\
&\le \frac{1}{|M_2|}\sqrt{|A|}\sqrt{(b-a+1)\frac{|M_2|}{|M_1|}}\frac{|M_2|}{|M_1|} \\
&= O(\frac{|M_2|}{|M_1|}).
\end{aligned}
$$

As a result, the contribution from each ends is at most $|M_2|/|M_1|$, which concludes our proof. □

3.8 Extension to general cases

The results we presented above only apply for very specific cases, and it is significantly harder to extend the result to the general cases with the same technique. The approaches we introduce below are still work in progress and need to be further refined.

For the Legendre symbol χ, it has intimate connection to the Kloosterman sums.

Definition 3.8.1. Let p be a prime. For any integers a, b, the quantity

$$K[a, b; p] = \sum_{x \in (\mathbb{Z}/p\mathbb{Z})^\times} \exp(2\pi\imath(ax + bx^{-1})/p),$$

where x^{-1} denotes the multiplicative inverse of x in $\mathbb{Z}/p\mathbb{Z}$, is called a Kloosterman sum.

In connecting χ with the Kloosterman sum, Ernst Jacobsthal wrote down a formula in the footnote on page 239 of [43], while referring the readers to his Ph.D thesis. In [7], one of the authors derived the proof again.

Lemma 3.8.2 (Lemma 3.3 in [7])**.** *Let a be an integer not divisible by p and $F : \mathbb{Z}/p\mathbb{Z} \to \mathbb{C}$ be any function. Then*

$$\sum_{x \in (\mathbb{Z}/p\mathbb{Z})^\times} F[x + ax^{-1}] = \sum_{x \in \mathbb{Z}/p\mathbb{Z}} F[x] + \sum_{x \in \mathbb{Z}/p\mathbb{Z}} \chi[x^2 - 4a]F[x].$$

From Lemma 3.8.2, we can derive the following equality:

Lemma 3.8.3. *For any $m_1 \neq m_2$, one has*

$$\sum_{k\in\mathbb{Z}/p\mathbb{Z}} \chi[k+m_1-m_2]\chi[k] = -1.$$

Proof. Since χ is a character on the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$,

$$\begin{aligned}\sum_{k\in\mathbb{Z}/p\mathbb{Z}} \chi[k+m_1-m_2]\chi[k] &= \sum_{k\in\mathbb{Z}/p\mathbb{Z}} \chi[k(k+m_1-m_2)] \\ &= \sum_{k\in\mathbb{Z}/p\mathbb{Z}} \chi[(k+(m_1-m_2)/2)^2-(m_1-m_2)^2/4] \\ &= \sum_{k\in\mathbb{Z}/p\mathbb{Z}} \chi[k^2-(m_1-m_2)^2/4],\end{aligned}$$

by shifting the indices in the last line. Choosing $a=(m_1-m_2)^2/16$ and let F be the constant function taking value 1, one has

$$p-1 = p+\sum_{x\in\mathbb{Z}/p\mathbb{Z}} \chi[x^2-4a] = p+\sum_{k\in\mathbb{Z}/p\mathbb{Z}} \chi[k+m_1-m_2]\chi[k].$$

□

Let $\tau_m\chi[k] := \chi[k+m]$. Then,

Corollary 3.8.4. *Given a fixed m_2, and $\Omega_1(n_1) \subset \mathbb{Z}/p\mathbb{Z}$,*

$$\| \sum_{m_1\in\Omega_1(n_1)} \tau_{m_1-m_2}\chi\|^2 = |\Omega_1(n_1)|(p-1) - |\Omega_1(n_1)|(|\Omega_1(n_1)|-1).$$

Proof.

$$
\begin{aligned}
&< \sum_{m_1 \in \Omega_1(n_1)} \tau_{m_1-m_2}\chi, \sum_{m_1 \in \Omega_1(n_1)} \tau_{m_1-m_2}\chi> \\
&= \sum_{m_1 \in \Omega_1(m_1)} <\tau_{m_1-m_2}\chi, \tau_{m_1-m_2}\chi> + \sum_{m_1 \neq m_1'} <\tau_{m_1-m_2}\chi, \tau_{m_1'-m_2}\chi> \\
&= |\Omega_1(n_1)| \sum_{k \in \mathbb{Z}/p\mathbb{Z}} |\chi[k]|^2 - |\Omega_1(n_1)|(|\Omega_1(n_1)| - 1) \\
&= |\Omega_1(n_1)|(p-1) - |\Omega_1(n_1)|(|\Omega_1(n_1)| - 1).
\end{aligned}
$$

□

To estimate $|\sum_{m_1 \in \Omega_1(n_1)} A_p(\chi)[m_1 - m_2, n_1 - n_2]|$, note that

$$
\begin{aligned}
\sum_{m_1 \in \Omega_1(n_1)} A_p(\chi)[m_1 - m_2, n_1 - n_2] &= \frac{1}{p} \sum_{k \in \mathbb{Z}/p\mathbb{Z}} \sum_{m_n \in \Omega_1(n_1)} \tau_{m_1-m_2}\chi[k]\chi[k] e_p(k(n_1 - n_2)) \\
&= \frac{1}{\sqrt{p}} (\mathcal{F}v)[n_1 - n_2],
\end{aligned}
$$

where $\mathcal{F} = \frac{1}{\sqrt{p}}(e_p(kl))_{k,l \in \mathbb{Z}/p\mathbb{Z}}$ is the Discrete Fourier Transform (DFT) matrix and

$$
v = \left(\sum_{m_1 \in \Omega_1(n_1)} (\tau_{m_1-m_2}\chi)[k]\chi[k] \right)^t_{k \in \mathbb{Z}/p\mathbb{Z}},
$$

with the slight abuse of notation $v = v(\Omega_1(n_1), m_2)$. From Corollary 3.8.4, we see that

$$
\|v\|_2^2 = |\Omega_1(n_1)|(p-1) - |\Omega_1(n_1)|(|\Omega_1(n_1)| - 1) - |\sum_{m_1 \in \Omega_1(n_1)} \tau_{m_1-m_2}\chi[0]|^2.
$$

Thus,

$$|\sum_{m_1\in\Omega_1(n_1)} A_p(\chi)[m_1-m_2,n_1-n_2]| \le \frac{1}{\sqrt{p}}\|\mathcal{F}v\|_\infty.$$

Remark 3.8.5. For $n_1 = n_2$, we have

$$\hat{v}[0] = (\mathcal{F}v)[0] = \sum_{k\in\mathbb{Z}/p\mathbb{Z}}\sum_{m_1\in\Omega_1(n_1)} \tau_{m_1-m_2}\chi[k]\chi[k] = -|\Omega_1(n_1)|.$$

Since $\mathcal{F}$ is a unitary transform, we see that

$$\sum_{x\in\mathbb{Z}/p\mathbb{Z}} |\mathcal{F}v[x]|^2 = |\Omega_1(n_1)|(p-1) - |\Omega_1(n_1)|(|\Omega_1(n_1)|-1) - |\sum_{m_1\in\Omega_1(n_1)} \tau_{m_1-m_2}\chi[0]|^2 \le |\Omega_1(n_1)|p.$$

Thus, the average entry-wise magnitude of $\mathcal{F}v$ will be

$$\frac{1}{\sqrt{p}}\|\mathcal{F}v\|_2 \le \sqrt{|\Omega|_1(n_1)}.$$

If we can replace $\frac{1}{\sqrt{p}}(\mathcal{F}v)[n_1-n_2]$ by $\frac{\sqrt{|\Omega_1(n_1)|}}{\sqrt{p}}$, then assuming $|\mathcal{N}| \le p^c$, we will have

$$\sum_{n_1,n_2\in\mathcal{N}}\sum_{m_2\in\Omega_2(n_2)} |\sum_{m_1\in\Omega_1(n_1)} A_p(\chi)[m_1-m_2,n_1-n_2]| \le \sum_{n_1\in\mathcal{N}}\sqrt{|\Omega_1(n_1)|} \le p^{c/2}p^{1/4}.$$

It indicates that if $\mathcal{F}v$ is of constant amplitude (CA), then we would have proven Theorem 3.3.1 with even smaller exponents. The following lemma related $\mathcal{F}v$ being CA with v being of zero auto-correlation (ZAC):

Lemma 3.8.6. *Given a sequence $x \in \mathbb{C}^N$, $\hat{x}$ is CA if and only if x is ZAC.*

It leads us to examine the auto-correlation of v. However, as we will see below,

v is not ZAC, and the estimation on its auto-correlation is not enough to provide a meaningful bound. Instead, we will employ the power method to examine the high moments of $\hat{v}$.

3.8.1 Weil's Exponential Sums and the Power Method

Mentioned in [5] and proven in [52], we have the following Weil's exponential sum estimate:

Lemma 3.8.7 (Theorem 3.1 in [5], Theorem 2C of Chapter 2 in [52]). *Given $k \in \mathbb{N}$, there exists p_0 such that for any $0 < t_1 < \cdots < t_k < p$, one has*

$$|\sum_{x \in \mathbb{Z}/p\mathbb{Z}} \chi(x+t_1)\cdots\chi(x+t_k)| \leq kp^{1/2}.$$

Remark 3.8.8. For any vector v,

$$\|v\|_\infty \leq \inf_{s \geq 1} \|v\|_s.$$

Now, we are able to give the following estimate on $\|\mathcal{F}v\|_\infty$:

Theorem 3.8.9 (This theorem is wrong!). *Given fixed $m_2, \Omega_1(n_1)$, one has*

$$\|v\|_{16} \leq |\Omega_1(n_1)|^{3/4} p^{1/16}.$$

In fact, for any integer $s = 2^n$ for some $n \in \mathbb{N}$, one has

$$\|v\|_s \leq |\Omega_1(n_1)|^{3/4} p^{1/s}.$$

Before proving Theorem 3.8.9, we shall see how we can use it to prove RIP:

Proof. of RIP:

Given a fixed $\mathcal{N} \subset \mathbb{Z}/p\mathbb{Z}$ and disjoint $\Omega_1, \Omega_2 \subset \mathbb{Z}/p\mathbb{Z} \times \mathcal{N}$, we have

$$\begin{aligned}
\left| \sum_{(m_1,n_1)\in\Omega_1} \sum_{(m_2,n_2)\in\Omega_2} <u_{m_1,n_1}, u_{m_2,n_2}> \right| &\leq \sum_{n_1,n_2\in\mathcal{N}} \sum_{m_2\in\Omega_2(n_2)} \left| \sum_{m_1\in\Omega_1(n_1)} A_p(\chi)[m_1 - m_2, n_1 - n_2] \right| + \frac{8}{9} \\
&\leq \sum_{(m_2,n_2)\in\Omega_2} \sum_{n_1\in\mathcal{N}} \frac{1}{\sqrt{p}} \|\hat{v}\|_\infty + \frac{8}{9} \\
&\leq \frac{\sqrt{p}}{3} \frac{1}{\sqrt{p}} \sum_{n_1\in\mathcal{N}} |\Omega_1(n_1)|^{3/4} \liminf_{s>1} p^{1/s} + \frac{8}{9} \\
&\leq \frac{1}{3} |\mathcal{N}| \left(\frac{|\Omega_1|}{|\mathcal{N}|} \right)^{3/4} + \frac{8}{9} \\
&\leq \frac{1}{3} \frac{1}{3^{3/4}} p^{\frac{0.09c}{4} + \frac{3}{8}} + \frac{8}{9} \\
&\leq p^{\frac{0.09c}{4} - \frac{1}{8}} p^{1/2},
\end{aligned}$$

where the fourth inequality follows from the fact that $f(t) = t^{3/4}$ is concave. □

Before proving Theorem 3.8.9, we recall the following facts:

Remark 3.8.10. The following statements are true:

- For any two vectors $f, g \in \mathbb{C}^p$, $\widehat{f * g}[t] = \sqrt{p} \hat{f}[t] \hat{g}[t]$ for any $t \in \mathbb{Z}/p\mathbb{Z}$,
- $\|f * g\|_1 \leq \|f\|_1 \|g\|_1$, and

- $\|f * g\|_2 \leq \|f\|_1 \|g\|_2$ by Minkowski's inequality.

Proof. of Theorem 3.8.9: First, consider

$$\begin{aligned} |p^3 \sum_t |\hat{v}[t]|^8 e_p(tl)| &= |<\tau_{-l}(v*v*v*v), v*v*v*v>| \\ &\leq \|v*v*v*v\|_2^2 \\ &\leq \|v*v*v\|_1^2 \|v\|_2^2 \\ &\leq \|v\|_1^6 \|v\|_2^2 \\ &\leq p^3 \|v\|_2^6 \|v\|_2^2 \\ &\leq p^3 \cdot |\Omega_1(n_1)|^4 p^4 = |\Omega_1(n_1)|^4 p^7. \end{aligned} \tag{3.22}$$

On the other hand,

$$\begin{aligned} &<\tau_{-l}(v*v*v*v), v*v*v*v> \\ &= \sum_{m_1^1,\ldots,m_1^8} \sum_k \Big(\sum_{s_1,s_2,s_3} \chi[l+k+(m_1^1-m_2)-s_1]\chi[l+k-s_1]\chi[s_1+(m_1^2-m_2)-s_2]\chi[s_1-s_2]\cdots \Big) \\ &\cdot \Big(\sum_{t_1,t_2,t_3} \chi[k+(m_1^5-m_2)-t_1]\chi[k-t_1]\chi[t_1+(m_1^6-m_2)-t_2]\chi[t_1-t_2]\cdots \Big). \end{aligned}$$

Thus,

$$
\begin{aligned}
&|<\tau_{-l}(v*v*v*v), v*v*v*v>| \\
&\leq \sum_{m_1^1,\ldots,m_1^8} \sum_k \sum_{s_2,s_3} \left| \sum_{s_1} \chi[l+k+(m_1^1-m_2)-s_1]\chi[l+k-s_1]\chi[s_1+(m_1^2-m_2)-s_2]\chi[s_1-s_2] \right| \\
&\cdot \sum_{t_2,t_3} \left| \sum_{t_1} \chi[k+(m_1^5-m_2)-t_1]\chi[k-t_1]\chi[t_1+(m_1^6-m_2)-t_2]\chi[t_1-t_2] \right| \\
&= \sum_{m_1^1,\ldots,m_1^8} \sum_k \sum_{s_2,s_3} S(l,k,m_1^1,m_1^2,s_2)T(k,m_1^5,m_1^6,t_2).
\end{aligned}
$$

By multiplying the each of the first two terms by -1, we are ready to apply Lemma 3.8.7:

1. Suppose
$$l+k+(m_1^1-m_2), l+k-s_1, (m_1^2-m_2)-s_2, s_2$$
are distinct, then by Lemma 3.8.7, $|S(l,k,m_1^1,m_1^2,s_2)| \leq 4p^{1/2}$.

2. Suppose two of the four terms above are identical, then by Lemma 3.8.3, $|S(l,k,m_1^1,m_1^2,s_2)| = 1$.

3. Suppose there are two identical pairs, then
$$\begin{cases} l+k+m_1^1 = s_2 - m_1^2 \\ l+k = s_2 \end{cases},$$
which implies $m_1^1 = m_1^2$, and $k = s_2 - l$. For a fixed l, there are $|\Omega_1(n_1)|p$ such solutions.

Taking all the situations into account, we have

$$|<\tau_{-l}(v*v*v*v), v*v*v*v>| \leq 16|\Omega_1(n_1)|^8 p^6 + 2|\Omega_1(n_1)|^7 \cdot 4p^{5.5} - |\Omega_1(n_1)|^6 p^5$$
$$\leq 17|\Omega_1(n_1)|^8 p^6. \tag{3.23}$$

Combining the estimates in (3.22) and (3.23), we see that

$$|p^3 \sum_t |\hat{v}[t]|^8 e_p(tl)| \leq \min\{|\Omega_1(n_1)|^4 p^7, 17|\Omega_1(n_1)|^8 p^6\} \leq \sqrt{17}|\Omega_1(n_1)|^6 p^{6.5}.$$

Let $y[t] := |\hat{v}[t]|^8$, then

$$|\hat{y}[l]| \leq \sqrt{17}\Omega_1(n_1)^6 p^3.$$

By Parsevel's identity,

$$\sum_{t\in\mathbb{Z}/p\mathbb{Z}} |y[t]|^2 = \sum_l |\hat{y}[l]|^2 \leq 17\Omega_1(n_1)^{12} p^7.$$

Taking the 16-th root on both sides, we get

$$\|\hat{v}\|_{16} \leq 17^{1/16}\Omega_1(n_1)^{3/4} p^{7/16}.$$

□

3.8.2 Numerical Experiments

Since we want to show that

$$\|\hat{v}\|_\infty \leq |\Omega_1(n_1)|^{1-\epsilon}$$

for some $\epsilon > 0$, where

$$\hat{v}[s] = \widehat{v_{\Omega_1(n_1)}}[s] = \frac{1}{\sqrt{p}} \sum_{k \in \mathbb{Z}/p\mathbb{Z}} \sum_{m \in \Omega_1(n_1)} \tau_m \chi[k] \chi[k] e_p(kn),$$

it is perhaps beneficial to examine whether the claim has some substance in it.

We have the following two algorithms that attempt to verify our conjecture.

Algorithm 1. This algorithm seeks to simulate $\sup_{M \subset \mathbb{Z}/p\mathbb{Z}: |M| \leq \sqrt{p}} \|\widehat{v_M}\|_\infty$:

1. A fixed prime $p \in \mathbb{N}$ and number of iterations $iter \in \mathbb{N}$ are given.
2. for i in range($\lfloor \sqrt{p} \rfloor$):
3. for j in range($iter$):
4. $M_j \leftarrow$ random sample from $\mathbb{Z}/p\mathbb{Z} \backslash \{0\}$ with size j.
5. $Value[i] \leftarrow \sup_j \|\widehat{v_{M_j}}\|_\infty$.
6. Plot $(\{1, \ldots, \lfloor \sqrt{p} \rfloor\}, Value)$.

Algorithm 2. This algorithm seeks to simulate $\sup_{M \subset \mathbb{Z}/p\mathbb{Z}: |M| \leq \sqrt{p}} \sup_{t \in \mathbb{Z}/p\mathbb{Z}} |<\tau_t v_M, v_M>|$.

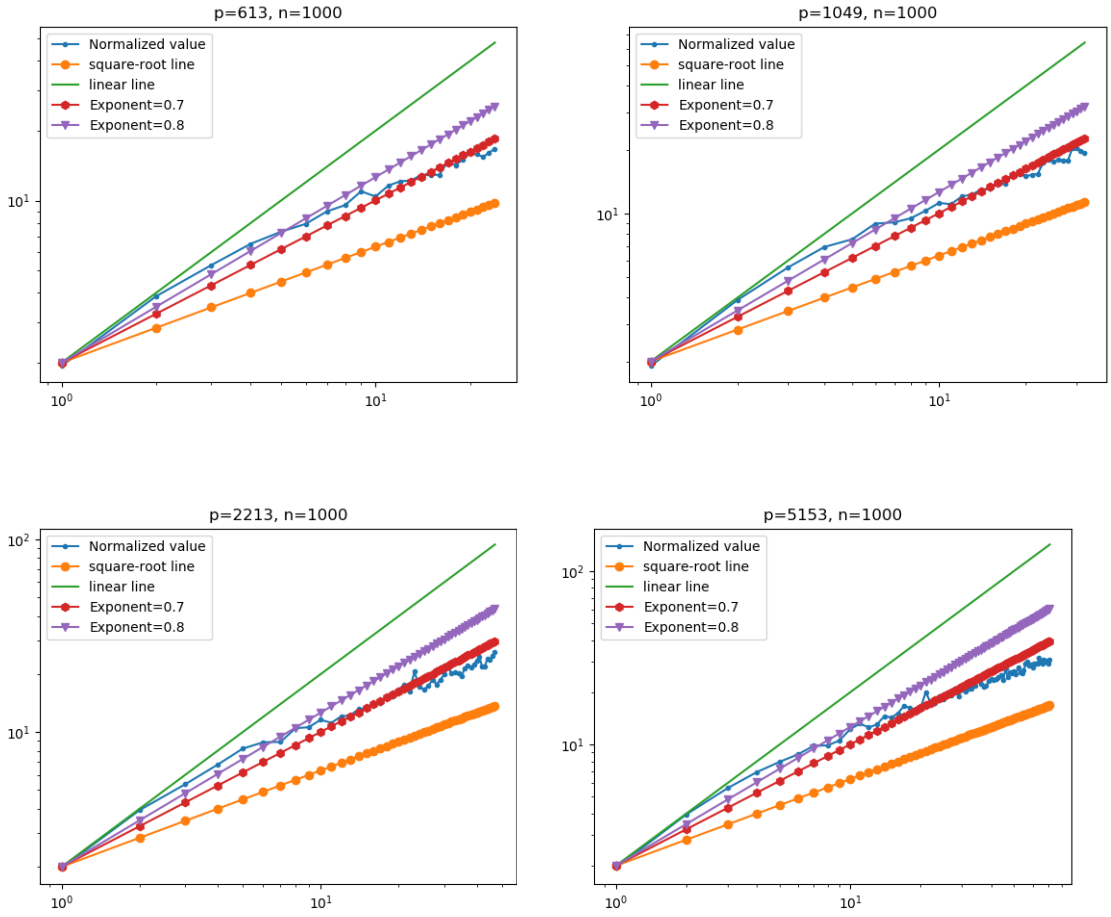


Figure 3.1: Illustration of the value $\sup_{M:|M|=j} \|\hat{v}\|_\infty$ with respect to j in log-log plot. Data are normalized by $2/\sqrt{p}$

1. A fixed prime $p \in \mathbb{N}$ and number of iterations $iter \in \mathbb{N}$ are given.

2. for i in range($\lfloor\sqrt{p}\rfloor$):

3. for j in range($iter$):

4. $M_j \leftarrow$ random sample from $\mathbb{Z}/p\mathbb{Z}\backslash\{0\}$ with size j.

5. $Value[i] \leftarrow \sup_j \sup_t |<\tau_t v_{M_j}, v_{M_j}>|$.

6. Plot $(\{1, \ldots, \lfloor\sqrt{p}\rfloor\}, Value)$.

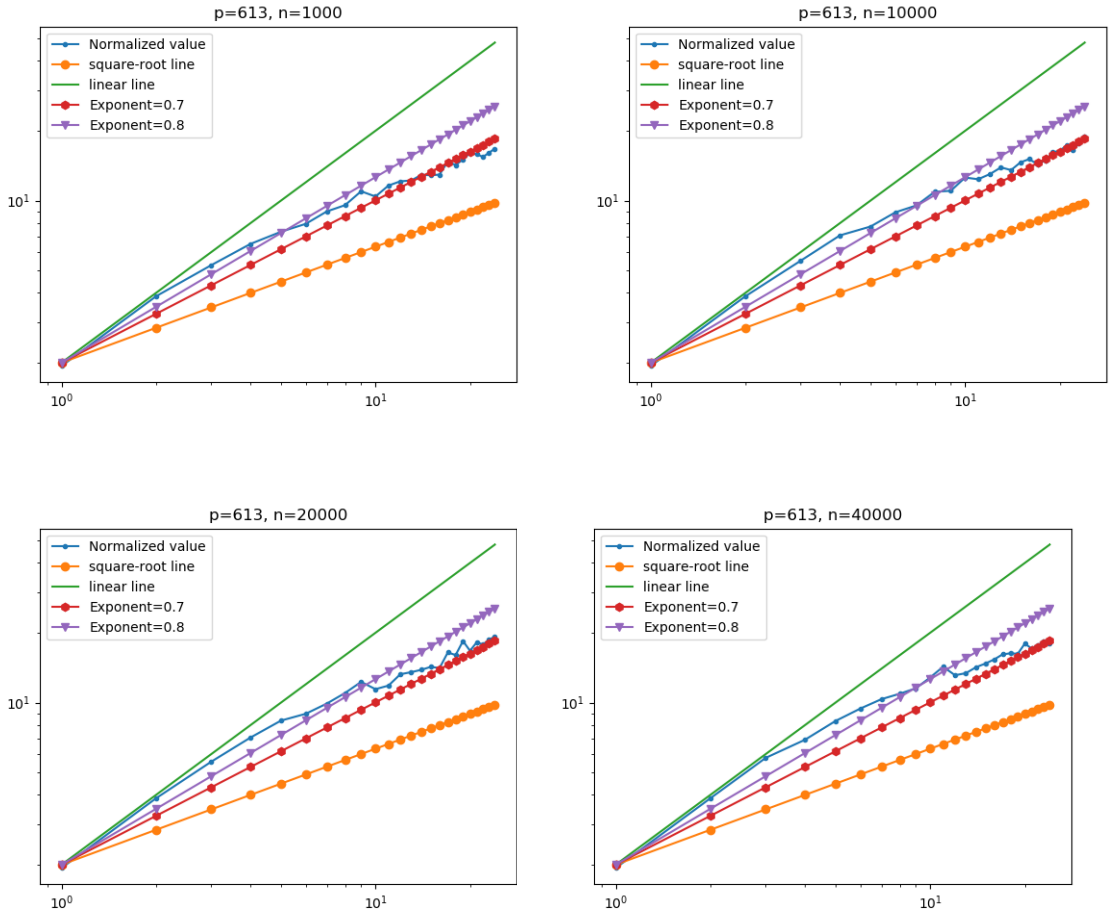


Figure 3.2: Given a fixed p, we examine the asymptotic behavior of the graph as the number of iterations increases. Illustration of the value $\sup_{M:|M|=j} \|\hat{v}\|_\infty$ with respect to j in log-log plot. Data are normalized by $2/\sqrt{p}$

3.8.3 Premature Ideas

To tackle the problems posed in Section 3.8, we made many unsuccessful attempts. Below we recount some of them, and even though they do not seem to lead to anywhere, they may be of use after some modification.

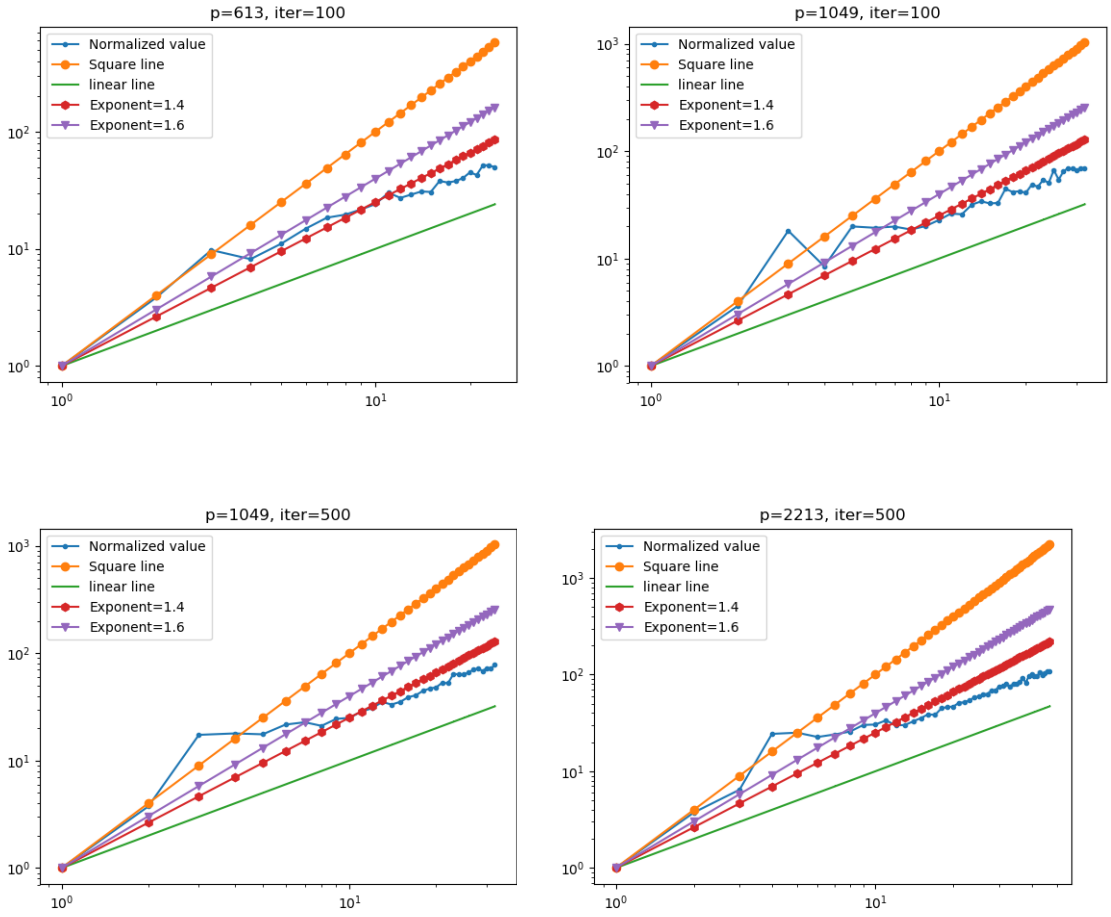


Figure 3.3: Illustration of $\sup_{M:|M|=j} \sup_{t\in\mathbb{Z}/p\mathbb{Z}} |<\tau_t v_M, v_M>|$ with respect to j in log-log plot. Data are normalized by the first entry.

3.8.3.1 Exact Counting

The idea is to derive the exact distribution of signs of $\{\chi[k+m_1]\chi[k+m_2]\chi[k+m_3]\chi[k]\} + k$. Incidentally, it is quite accurate up to counting $\{\chi[k+m_1]\chi[k+m_2]\chi[k]\}_k$, but things break down in the case with four elements, and one needs to resort to Weil's exponential sum estimate, which defeats the purpose of exact counting.

- $n = 2$:

First, one should note that for any $m \neq 0$, $\sum_k \chi[k+m]\chi[k] = -1$. Thus, $\{\chi[k+m]\chi[k]\}_k$ has 2 zeros, $(p-1)/2$ negative ones, and $(p-3)/2$ positive ones. Thus, looking at $w[k] := \chi[k+m] + \chi[k]$, we see that w must satisfy

$$w \to \begin{cases} \pm 2 & : \ \frac{p-3}{2} \\ 0 & : \ \frac{p-1}{2} \\ \pm 1 & : \ 2 \end{cases}$$

by viewing the positive sign in $\chi[k+m]\chi[k]$ as amplification and negative sign as cancellation.

- $n = 3$:

  We know that $\sum_k (\chi[k+m_1] + \chi[k+m_2])\chi[k] = -2$. Suppose that there are $\tilde{j}$ of 2's, and $\tilde{i}$ of 1's, then we have the following situations:

  1. $\chi[m_1] + \chi[m_2] = \pm 2 \implies \tilde{i} = 0, 2$:
     - $\tilde{i} = 0$: $-2 = 2\tilde{j} - 2(\frac{p-3}{2} - 1 - \tilde{j}) - 2 \implies \tilde{j} = \frac{p-5}{4}$.
     - $\tilde{i} = 2$: $-2 = 2\tilde{j} - 2(\frac{p-3}{2} - 1 - \tilde{j}) + 2 \implies \tilde{j} = \frac{p-9}{4}$.
  2. $\chi[m_1] + \chi[m_2] = 0 \implies \tilde{i} = 1$:
     - $\tilde{i} = 1$: $-2 = 2\tilde{j} - 2(\frac{p-3}{2} - \tilde{j}) \implies \tilde{j} = \frac{p-5}{4}$.

  Note that $\tilde{i}$ depends on the sign distribution of $\chi[m_1]$ and $\chi[m_2]$.

  As the number of $\tilde{j}$ represents amplification, we can derive that for $z[k] = \chi[k+m_1] + \chi[k+m_2] + \chi[k]$, one has

$$z \to \begin{cases} \pm 3 & : \quad \tilde{j} \\ \pm 2 & : \quad \tilde{i} + 1 \\ 0 & : \quad 2 - \tilde{i} \\ \pm 1 & : \quad (\frac{p-3}{2} - \tilde{j}) + \frac{p-1}{2} - 1 \end{cases} .$$

Up to this point, we are able to fairly accurately capture the exact distributions. We shall see the difficulty arising when $n = 4$.

- $n = 4$:

  Again, we have $\sum_k (\chi[k+m_1]+\chi[k+m_2]+\chi[k+m_3])\chi[k] = -3$. We assume that m_1, m_2, m_3 are distinct nonzero residues. We look at $s = \chi[m_1]+\chi[m_2]+\chi[m_3]$. Since none of them are 0, it can only be $\pm 1, \pm 3$.

  Let's talk about the case when $(s, \tilde{i}, \tilde{j}) = (\pm 3, 0, \frac{p-5}{2})$. If there are $\tilde{a}$ of 3's, $\tilde{b}$ of 2's, and $\tilde{c}$ of 1's, then

$$\begin{aligned} -3 &= 3\tilde{a} - 3(\frac{p-5}{4} - 1 - \tilde{a}) + 2\tilde{b} - 2(1 - \tilde{b}) + \tilde{c} - (\frac{p-5}{4} + \frac{p-1}{2} - \tilde{c}) \\ &= 6\tilde{a} + 4\tilde{b} + 2\tilde{c} - \frac{3}{2}(p-5) - 1 \\ &\implies 3\tilde{a} + 2\tilde{b} + \tilde{c} = \frac{3}{4}(p-5) - 1. \end{aligned}$$

  On the other hand, we see that what we really want is

$$\begin{aligned} \sum_k \chi[k+m_1]\chi[k+m_2]\chi[k+m_3]\chi[k] &= \tilde{a} - \tilde{c} - (\tilde{j} - \tilde{a}) + ((\frac{p-3}{2} - \tilde{j}) + \frac{p-1}{2} - 1 - \tilde{c}) \\ &= 2\tilde{a} - 2\tilde{c} + \frac{p-5}{2} + 2 \\ &= 8\tilde{a} + 4\tilde{b} - (p-9), \end{aligned}$$

where $\tilde{b} = 0$ or 1.

Note that $\tilde{a}$ represents the number such that all four elements have the same sign, i.e.,

$$\tilde{a} = |\{k \in \mathbb{Z}/p\mathbb{Z} : \chi[k+m_1] = \chi[k+m_2] = \chi[k+m_3] = \chi[k]\}|.$$

From the pseudo-randomness of Legendre symbols, we see that $|\tilde{a} - \frac{p}{8}| = O(p^{1/2})$. However, how the distribution of (m_1, m_2, m_3) affects the distance, we are not sure yet. Knowing this more thoroughly will help us pin down the exact value, and in turn it may help us crack $<\tau_t v, v>$ problem.

Algorithm 3. This algorithm simulates the distribution of $\{\sum_k \chi[k+m_1]\chi[k+m_2]\chi[k+m_3]\chi[k]\}_{(m_1,m_2,m_3)}$.

1. A fixed prime $p \in \mathbb{N}$ and number of iterations $iter \in \mathbb{N}$ are given.
2. for j in range($iter$):
3. $M_j \leftarrow$ random sample (m_1^j, m_2^j, m_3^j) from $\mathbb{Z}/p\mathbb{Z}\backslash\{0\}$ with size 4.
4. $Value[i] \leftarrow \sum_k \chi[k+m_1^j]\chi[k+m_2^j]\chi[k+m_3^j]\chi[k]$.
5. Plot histogram of $Value$.

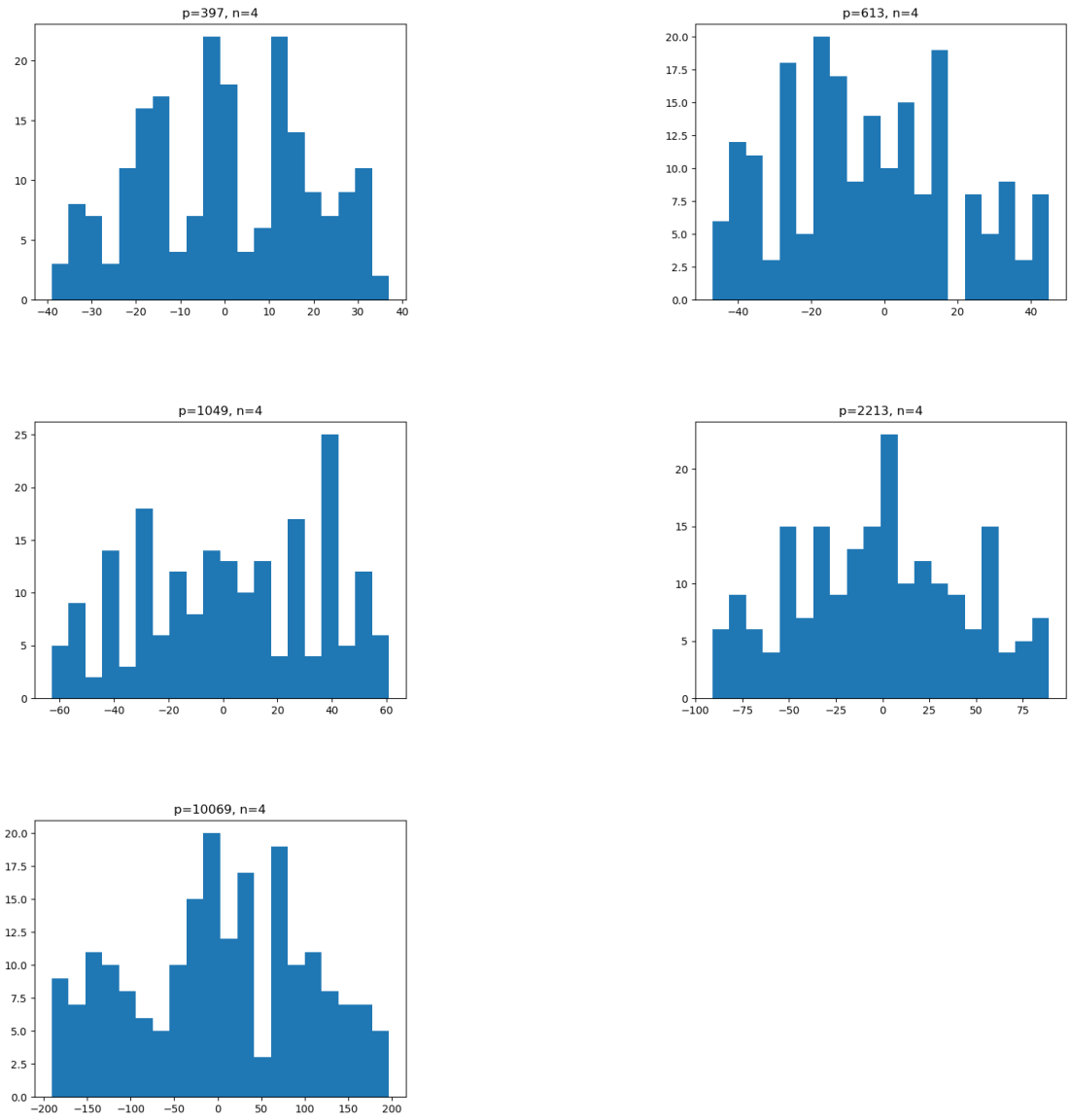


Figure 3.4: Distribution of $\sum_k \chi[k+m_1]\chi[k+m_2]\chi[k+m_3]\chi[k]$ for 200 random samples of (m_1, m_2, m_3).

3.8.3.2 Splitting Convolutions into Shifted Products

We will illustrate with the case $v * v * v * v$.

$$\begin{aligned} v * v * v * v[k] &= \sum_{s_1} v[s_1] \sum_{s_2} v[k - s_1 - s_2] \sum_{s_3} v[s_3] v[s_2 - s_3] \\ &= \sum_{s_2} \left(\sum_{s_1} v[s_1] v[k - s_1 - s_2] \right) \left(\sum_{s_3} v[s_3] v[s_2 - s_3] \right) \\ &= \sum_{s_2} \left(\sum_{s_1} v[s_1] v[s_1 + s_2 - k] \right) \left(\sum_{s_3} v[s_3] v[s_3 - s_2] \right) \\ &= \sum_{s_2} \left(<\tau_{k-s_2} v, v><\tau_{s_2} v, v> \right). \end{aligned}$$

On the other hand, $\|v * v * v * v\|_2^2 = p^{3/2} \||\hat{v}|^4\|_2^2$. For n layers of convolution, the shifted products absorbs about $(n-1)/2$ layers of whole sums. If we can show that $<\tau_t v, v> = O(M^{2-\epsilon})$, independent of p, then we can estimate the remaining sums trivially and still arrive at the desired bound. The problem is, is it possible to bound $<\tau_t v, v>$ solely with, say, $M^{1.6}$? It seems unlikely, as the simulation shows that the quantity also grows with $\sqrt{p}$, which means we might need better estimations.

3.8.3.3 Randomness of Legendre Symbol

The pseudo-randomness of Legendre symbols may also help us in estimating $\|\hat{v}\|_\infty$.

By viewing $\mathbb{Z}/p\mathbb{Z}$ as a sample space with the normalized counting measure as

the probability measure, we see that each $\tau_t\chi$ is almost Bernoulli with

$$\mathbb{P}(\chi = 1) = \mathbb{P}(\chi = -1) = \frac{p-1}{2p}.$$

Also, they are almost mutually independent, as $Cov(\tau_t\chi, \chi) = \frac{-1}{p}$. Thus, it might be expected that $\sum_{m\in M} \tau_m\chi$ may weakly converge to a normal random variable $X \sim \mathcal{N}(0, M\frac{p-1}{p})$.

Now, $v_M = X\chi$ is a perturbed version of a normal random variable, and $<\tau_t v_M, v_M> = p\mathbb{E}(\tau_t v_M \cdot v_M)$. If we naively believe that is indeed the case, and suppose the index shift t is such that $\tau_t M \cap M = \phi$, then we may invoke

$$<\tau_t v_M, v_M> = p\mathbb{E}(\tau_t v_M \cdot v_M) = p\mathbb{E}(\tau_t v_M)\mathbb{E}(v_M) = \frac{1}{p}\Big(\sum_k v_M[k+t]\Big)\Big(\sum_k v_M[k]\Big) = \frac{|M|^2}{p}, \tag{Naive}$$

by noting that $\sum_{m\in M}\sum_k \chi[k+m]\chi[k] = -|M|$.

If that is the end, then everything will be solved. However, there is bound to be some error term. Thus, it will become

$$<\tau_t v_M, v_M> = \frac{|M|^2}{p} + Error. \tag{Realistic}$$

The first term is inconsequential, while we would have hoped the error term to be of the form $C|M|^{1-\epsilon}$, which is not entirely outrageous if we think of v_M as some sort of perturbed normal random variable. On the other hand, the left-hand side seems to also grow with $p^{1/2}$. If that is the case, then the end estimate of $\|\hat{v}\|_\infty$ will take

on an additional $p^{1/4}$, which does not bode well at all, given that we are not able to accommodate this factor with the gained space from $|M|^{1-\epsilon}$.

Chapter 4: Weight Identification for ReLu Neural Networks

In this chapter, we discuss whether it is possible to recover the weights in neural networks with rectified linear units (ReLU) as activation functions up to permutation and re-scaling with positive factors. In particular, we extend the approach employed in [33, 34] to neural networks with non-smooth activation functions, specifically leaky ReLU neural networks. A more precise description of the problems is given in Section 4.1.3. We first investigate 1-layer networks, which yield similar results to the ones discussed in [33]. Then, we outline our work on 2-layer networks. This part is not yet finished, and research is still ongoing.

First, we start with the definition of a (leaky) ReLU function:

Definition 4.0.1. A *Rectified Linear Unit* (ReLU) σ is defined to be $\sigma(x) = x$ if $x > 0$, and $\sigma(x) = 0$ if $x \leq 0$.

A leaky ReLU σ_η with parameter $\eta \in [0, 1)$ satisfies $\sigma_\eta(x) = x$ if $x > 0$, and $\sigma_\eta(x) = \eta x$ if $x \leq 0$.

We shall cover the preliminaries in Section 4.1 before stating and proving our results in Sections 4.2 and 4.3.

4.1 Preliminaries

4.1.1 Tensor Decomposition and Its Motivation

A neural network $\tilde{f}(x) : \mathbb{R}^d \to \mathbb{R}$ is an alternating composition of affine transformations and non-linear activation functions:

$$\tilde{f}(x) = A_L \circ \sigma \circ A_{L-1} \circ \cdots \circ \sigma \circ A_1(x),$$

where $A_j : \mathbb{R}^{d_{j-1}} \to \mathbb{R}^{d_j}$ are affine transformations and σ is a non-linear activation function applied entry-wise.

For exposition, we first restrict ourselves on the case where $\tilde{f}$ has only one hidden layer:

$$f(x) = \sum_{i=1}^{m} \alpha_i \sigma(\sum_{j=1}^{d} w_{ij} x_j + \theta_i) = \sum_{i=1}^{m} g_i(a_i \cdot x), \quad x \in \mathbb{R}^d.$$

Our task is to recover all $\{\alpha_i\}, \{w_{ij}\}, \{\theta_i\}$, or equivalently, $\{g_i, a_i\}$. Note that f is a sum of ridge functions with unknown ridge directions. Estimation of the sum of ridge functions has been extensively studied in approximation theory [11, 22, 23, 28]. In particular, in [10] it was pointed out that higher order differentiation can be used to obtain information on the ridge directions $\{a_j\}$ by

$$D_{c_1} \cdots D_{c_k} f(x) = \sum_i g_i^{(k)}(a_i \cdot x)(a_i \cdot c_1) \cdots (a_i \cdot c_k).$$

This provides a connection between identification of $\{a_i\}$ and tensor decomposition of the derivatives of f. In particular, for all $1 \leq \ell, s \leq d$,

$$\partial_\ell f(x) = \sum_{i=1}^m g_i'(a_i \cdot x)(a_i)_\ell \implies \nabla f(x) = \sum_{i=1}^m g_i'(a_i \cdot x) a_i.$$
$$\partial_\ell \partial_s f(x) = \sum_{i=1}^m g_i''(a_i \cdot x)(a_i)_\ell (a_i)_s \implies \nabla^2 f(x) = \sum_{i=1}^m g_i''(a_i \cdot x) a_i \otimes a_i.$$

However, for some cases it is not feasible to obtain derivatives of f due to limited access to the samples. It is then necessary to employ weak differentiation: given random samples $x_k \sim \mu$ for some distribution μ with density $p(x)$, one can calculate

$$\begin{aligned}
\Delta_N^\ell(f) = \frac{1}{N}\sum_{k=1}^N (-1)^\ell f(x_k) \frac{\nabla^\ell p(x_k)}{p(x_k)} &\approx \int_{\mathbb{R}^d} (-1)^\ell f(x) \frac{\nabla^\ell p(x)}{p(x)} p(x)\, dx \\
&= \int_{\mathbb{R}^d} (-1)^\ell f(x) \nabla^\ell p(x) dx \\
&= \int_{\mathbb{R}^d} \nabla^k f(x) d\mu(x) \\
&= \sum_{i=1}^m \left(\int_{\mathbb{R}^d} g_i^{(k)}(a_i \cdot x) d\mu(x) \right) a_i \otimes \cdots \otimes a_i.
\end{aligned} \tag{4.1}$$

In the case when strong derivatives can be obtained, for any x,

$$\nabla f(x) \in A_1 = span\{a_1, \ldots, a_m\} \subset \mathbb{R}^d,$$
$$\nabla^2 f(x) \in \mathscr{A} = span\{a_1 \otimes a_1, \ldots, a_m \otimes a_m\} \subset \mathbb{R}^{d \times d},$$

where $a_j \otimes a_j = a_j a_j^T$. Suppose that $\{a_j\}_j$ are orthonormal, then we can exactly

recover $\{a_j\}$ from the eigen-decomposition of $\nabla^2 f(x)$ from just one sample. However, there are some difficulties in practice: (i) $\{a_j\}_j$ are generally not orthonormal, and (ii) the stability of eigen-decomposition relies on spectral gaps, which we do not know *a priori*.

To deal with the difficulties above, [33, 34] sample many approximations of $\{\nabla^2 f(x_i)\}_i$ so as to approximate $\mathscr{A}$ with $\tilde{\mathscr{A}}$. Using a technique called *whitening*, the weights $\{a_j\}_j$ can be assumed to be nearly orthogonal. Then, it was shown that unit-norm approximations of $\{\pm a_j\}_j$ can be recovered by searching for rank-1 matrices in $\tilde{\mathscr{A}}$ with the following optimization scheme

$$\arg\max \|M\|_\infty, \quad \|M\|_F \le 1, M \in \tilde{\mathscr{A}},$$

where $\|\cdot\|_\infty$ is the ℓ^2-to-ℓ^2 operator norm.

For neural networks with two layers, they are no longer sums of ridge functions. Instead, they take the following form

$$f(x) = \sum_{\ell=1}^{m_1} \alpha_\ell \sigma(b_\ell \cdot g(x) + d_\ell) = h \circ g(x), \quad (g(x))_j = \sigma(a_j \cdot x + c_j),\ j = 1, \dots, m_0.$$

Its second derivative has the following form

$$\begin{aligned}\nabla^2 f(x) &= \sum_{\ell=1}^{m_1} h_\ell'(b_\ell^T g(A^T x)) \sum_{j=1}^{m_0} b_{j\ell} g_j''(a_j \cdot x) a_j \otimes a_j \\ &+ \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m_0} h_\ell''(b_\ell^T g(A^T x)) b_{i\ell} b_{j\ell} g_i'(a_i^T) g_j'(a_j^T x)(a_i \otimes a_j + a_j \otimes a_i) \\ &= \sum_{\ell=1}^{m_1} h_\ell'(b_\ell^T g(A^T x)) \sum_{j=1}^{m_0} b_{j\ell} g_j''(a_j \cdot x) a_j \otimes a_j \\ &+ \sum_{\ell=1}^{m_1} h_\ell''(b_\ell^T g(A^T x)) \left[\sum_{i=1}^{m_0} b_{i\ell} g_i'(a_i^T x) a_i\right] \otimes \left[\sum_{j=1}^{m_0} b_{j\ell} g_j'(a_j^T x) a_j\right],\end{aligned}$$

where $A = (a_1|\cdots|a_{m_0})$.

While the first m_0 terms of $\nabla^2 f$ are $\{a_j \otimes a_j\}_j$ which do not depend on $x \in \mathbb{R}^d$, $\sum_i = 1^{m_0} b_{i\ell} g_i'(a_i^T x) a_i$ does change for different x. It is particularly problematic as now $\{\nabla^2 f(x_k)\}_k$ do not belong in the same (m_0+m_1)-dimensional subspace of $\mathbb{R}^{d\times d}$. However, assuming $\max_i \|g_i''\|_\infty$ can be well-controlled, each $\nabla^2 f(x)$ is close to the space $span\{a_j \otimes a_j, v_\ell \otimes v_\ell\}_{j,\ell}$, where $v_\ell = \sum_{i=1}^{m_0} b_{i\ell} g_i'(0) a_i$.

4.1.2 Perspective and Prior Works

In [4], it is shown that the third order tensor decomposition can be done by alternative rank-1 updates even for overcomplete cases. [44] analyzed a two-layer feedforward neural network with linear activation function on the second layer using (4.1) and the same techniques of [4].

In [33], the authors recovered weights for neural networks with one hidden layer using multiple instances of principal Hessian directions of the network, making it more robust than [44, 46] where only one instance is utilized. [34] extended the

same method to neural networks with two layers. For two layers, the principal Hessian direction belongs in different linear subspaces at each sample point due to the non-linearity of activation functions. A reference subspace was set up, and the error estimate relies on the bound of derivatives of the activation function. In both [33, 34], the activation function is assumed to be smooth, which excludes popular activation functions such as (leaky) Rectified Linear Unit (ReLU). In [34], it is more complicated to use weak differentiation due to the variability of principal Hessian directions, and recovery of bias terms is less explicit.

4.1.3 Problem Description

Our main focus here is on neural networks of one and two layers with ReLU as the activation function. The goal is to identify the weights of the neural networks up to permutation and re-scaling.

We can show that the ReLU neural networks with one layer can be written as $f(x) = \sum_{\ell=1}^{m} \sigma(a_\ell \cdot x + c_\ell)$, and the ones with two layers can be written as $f(x) = \sum_{\ell=1}^{m_1} \alpha_\ell \sigma(b_\ell \cdot g(x) + d_\ell)$, where $(g(x))_j = \sigma(a_j \cdot x + c_j)$ for all j. Note that there may be multiple neural networks corresponding to the same f by permutation and re-scaling with positive factors. However, up to those trivial transformations, we aim to reconstruct the parameters $\{a_j, c_j\}, \{b_\ell, d_\ell, \alpha_\ell\}$.

For simplicity, we shall assume that we have full access to the values of f, that is, we may sample however many points at wherever we desire. This is to investigate whether the algorithms to recover weights of f by from its second order

weak derivatives do exist. Nonetheless, our ultimate goal is to provide constructive algorithms and give guidelines on the selection of points during the process.

4.1.4 Results

This chapter follows similar techniques as [33, 34], except for the principal Hessian direction step: due to the difference between smooth activation functions and leaky ReLU, there is a different outlook than result in [34]. In their case, there would be *entangled weights* that vary for different sampling points. As for ours, the number of entangled weights grows exponentially with the dimension of the first layer. We propose a technique called *net-spreading*, performing weak differentiation locally for ReLU neural networks. Leveraging the properties of ReLU, a multi-scale version of the original algorithm used by Fornasier et al. can be employed. Favorable properties of leaky ReLUs also allow for reliable function recovery.

4.2 One-Layer Case

Leaky ReLU neural networks with one hidden layer have the following form

$$f(x) = \sum_{\ell=1}^{m} \sigma(a_\ell \cdot x + c_\ell),$$

where $\sigma(x) = \sigma_\eta(x) = \eta x + (1-\eta)x\mathbb{1}_{x>0}$ for $\eta \in [0,1)$ is the leaky ReLU function. Note that when $\eta = 0$, σ_0 is the usual ReLU function. In this section, we follow almost verbatim on [33].

Definition 4.2.1. Given $\phi \in C_c^\infty(\mathbb{R}^d)$, a compactly supported smooth function, and $\{x_k\}_{k=1}^N$ a set of N realizations of a random variable with probability density function $p(x)$, the ℓ-th order weak derivative of f $\nabla_{w,\phi}^\ell(f)$ is defined to be

$$\nabla_{w,\phi}^\ell(f) := \int_{\mathbb{R}^d} (-1)^\ell f(x) \nabla^\ell \phi(x)\, dx.$$

Its approximation $\Delta_{N,\phi}^\ell(f)$ is defined as

$$\Delta_{N,\phi}^\ell(f) := \frac{1}{N} \sum_{k=1}^N (-1)^\ell f(x_k) \frac{\nabla^\ell \phi(x_k)}{p(x_k)}.$$

Algorithm 4. Given $x \in \mathbb{R}^d$, a shallow neural network $f : \mathbb{R}^d \to \mathbb{R}$ with layer width $m < d$:

- Construct $Y_1 = (\Delta_{N,\phi_1}(f) | \cdots | \Delta_{N,\phi_{m_\chi}}(f))$. Compute the SVD

$$Y_1^T = \begin{pmatrix} \bar{U}_1 & \bar{U}_2 \end{pmatrix} \begin{pmatrix} \bar{\Sigma}_1 & \\ & \bar{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \bar{V}_1^T \\ \bar{V}_2^T \end{pmatrix}.$$

  Let A_1 be the row space of $\bar{V}_1^T$, where $\bar{V}_1^T \in \mathbb{R}^{m \times d}$.

- Dimension reduction using A_1. Let $P_{A_1} = BB^T : \mathbb{R}^d \to \mathbb{R}^d$ be the orthogonal projection onto A_1 and define $\bar{f} : \mathbb{R}^m \to \mathbb{R}$ by $\bar{f}(y) := \sum_i g_i(\alpha_i \cdot y)$ where $\alpha_i = B^T a_i$. Then $f(x) = \bar{f}(B^T x)$ for all $x \in \mathbb{R}^d$. In particular, we may assume $d = m$.

- Construct $Y_2 = (vec(\Delta^2_{N,\phi_1}(f))|\cdots|vec(\Delta^2_{N,\phi_{m_\chi}}(f)))$. Compute the SVD

$$Y_2^T = \begin{pmatrix} \tilde{U}_1 & \tilde{U}_2 \end{pmatrix} \begin{pmatrix} \tilde{\Sigma}_1 & \\ & \tilde{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \tilde{V}_1^T \\ \tilde{V}_2^T \end{pmatrix}.$$

  Let $\tilde{\mathscr{A}}$ be the row space of $\tilde{V}_1^T \in \mathbb{R}^{m\times m^2}$.

- Whitening: recursively making $\{a_i\}_i$ more and more orthonormal by finding positive definite matrices in $\tilde{\mathscr{A}}$.

- Optimize $\max \|M\|_\infty$ such that $\|M\|_F \le 1, M \in \tilde{\mathscr{A}}$. If the top eigenvalue of M is not $\|M\|_\infty$, replace M by $-M$. Take and record the top eigenvector, and repeat the same process many times. Use k-mean to recover $\{a_i\}_i$.

- Recover f: Let $\{\hat{a}_j\}_j$ be the normalized approximation to $\{a_j\}_j$. Let $\{\hat{b}_j\}_j$ be the dual basis to $\{\hat{a}_j\}_j$. Set $\hat{g}_j(t) = f(t\hat{b}_j)$ and $\hat{f}(x) = \sum_j \hat{g}_j(\hat{a}_j \cdot x)$.

In order to identify $\{a_\ell\}_\ell \subset \mathbb{R}^d$, we follow the same process in [33, 34] and recover them through the tensor decomposition of $\nabla^2 f$, the Hessian of f.

Due to the non-differentiability of σ at 0, we discuss the Hessian of f in the weak sense, i.e. , we would like to evaluate the following expressions: for any i, j,

$$\int_{\mathbb{R}^d} f(x)\partial_i\phi\, dx$$

and

$$\int_{\mathbb{R}^d} f(x)\partial_i\partial_j\phi\, dx.$$

4.2.1 Auxiliary Theorems

In this section, we list some important results that we will refer to in the following sections.

Theorem 4.2.2 (Chernoff Bound, [56]). *Let $X_1, \ldots, X_n$ be independent random, positive semi-definite matrices of dimension $m \times m$. Moreover, suppose that*

$$\sigma_1(X_j) \leq C$$

almost surely for all $j = 1, \ldots, n$. Let

$$\mu_{min} = \sigma_m(\sum_{j=1}^{n} \mathbb{E}X_j)$$

be the smallest singular value of the sum of the expectations. Then

$$\mathbb{P}\left\{\sigma_m(\sum_{j=1}^{n} X_j) - \mu_{min} \leq -s\mu_{min}\right\} \leq m \exp(\frac{-\mu_{min}s^2}{2C})$$

for all $s \in (0, 1)$.

Theorem 4.2.3 (Wedin's Bound, [57]). *Given two matrices B and $\tilde{B}$ with corresponding singular value decomposition*

$$B = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix},$$

and

$$\tilde{B} = \begin{pmatrix} \tilde{U}_1 & \tilde{U}_2 \end{pmatrix} \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & \tilde{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \tilde{V}_1^T \\ \tilde{V}_2^T \end{pmatrix}.$$

If there exists an $\bar{\alpha} > 0$ *such that*

$$\begin{cases} \min_{\ell,\tilde{\ell}} |\sigma_{\tilde{\ell}}(\tilde{\Sigma}_1) - \sigma_{\ell}(\Sigma_2)| \geq \bar{\alpha} \\ \min_{\tilde{\ell}} |\sigma_{\tilde{\ell}}(\tilde{\Sigma}_1)| \geq \bar{\alpha}, \end{cases}$$

then

$$\max\{\|U_1 U_1^T - \tilde{U}_1 \tilde{U}_1^T\|_F, \|V_1 V_1^T - \tilde{V}_1 \tilde{V}_1^T\|_F\} \leq \frac{\sqrt{2}}{\bar{\alpha}} \|B - \tilde{B}\|_F.$$

Corollary 4.2.4. *Let* $X_1, \ldots, X_N$ *be independent zero-mean* $d_1 \times d_2$*-random matrices. Assume that*

$$\|X_j\| \leq K$$

almost surely for all $1 \leq j \leq N$*, and denote*

$$\sigma^2 = \max\left(\|\sum_{j=1}^{N} \mathbb{E}(X_j X_j^T)\|, \|\sum_{j=1}^{N} (X_j^T X_j)\| \right),$$

then

$$\mathbb{P}\left(\|\sum_{j=1}^{N} X_j\| > \eta \right) \leq (d_1 + d_2)] exp(\frac{-\eta^2}{2(\sigma^2 + K\eta/3)}).$$

4.2.2 Weak Differentiation of Leaky ReLU Neural Networks

Proposition 4.2.5. *For any $1 \leq i \leq m$, $\phi \in C_c^\infty(\mathbb{R}^d)$, $\boldsymbol{a} \in \mathbb{R}^d$, and $c \in \mathbb{R}$,*

$$\int_{\mathbb{R}^d} \sigma(\boldsymbol{a} \cdot x + c) \partial_i \phi \, dx = -\int_{\{\boldsymbol{a} \cdot x + c \geq 0\}} \boldsymbol{a}_i \phi \, dx - \eta \int_{\{\boldsymbol{a} \cdot x + c \leq 0\}} \boldsymbol{a}_i \phi \, dx.$$

In particular, the weak derivative of $\sigma(\boldsymbol{a} \cdot x + c)$ is $\boldsymbol{a}_i(\eta + (1-\eta)\mathbb{1}_{\{\boldsymbol{a} \cdot x + c \geq 0\}})$ and $\nabla f(x) = \sum_{\ell=1}^m a_\ell(\eta + (1-\eta)\mathbb{1}_{A_\ell})$ where $A_\ell = \{x \in \mathbb{R}^d : a_\ell \cdot x + c_\ell \geq 0\}$ in distribution sense.

Proof. Note that $\sigma_\eta(t) = \eta t + (1-\eta)\sigma_0(t)$, and the strong derivative of $\mathbf{a} \cdot x + c$ is $\mathbf{a}$. Thus, it suffices to evaluate the weak derivative of $\sigma_0(\mathbf{a} \cdot x + c)$. We start by assuming that $\mathbf{a} = a\ \mathbf{e}_1$, where $\mathbf{e}_1$ is the first canonical coordinate in $\mathbb{R}^d$. Once we deduced the formula, we may then generalize to arbitrary $\mathbf{a}$ via change of variable.

Now,

$$\int_{\mathbb{R}^d} \sigma_0(ax_1 + c) \partial_i \phi \, dx = 0$$

for all $i > 1$, since $\sigma_0(ax_1 + c)$ does not depend on x_i and ϕ is compactly supported.

For $i = 1$,

$$\begin{aligned}
\int_{\mathbb{R}^d} \sigma_0(ax_1 + c) \partial_1 \phi \, dx &= \int_{\mathbb{R}^{d-1}} \int_{\{ax_1 + c \geq 0\}} (ax_1 + b) \partial_1 \phi \, dx_1 \, d\hat{x}_1 \\
&= -\int_{\mathbb{R}^{d-1}} \int_{\{ax_1 + c \geq 0\}} a\phi \, dx_1 \, d\hat{x}_1 \\
&= -\int_{\{\mathbf{a} \cdot x + c \geq 0\}} \mathbf{a}_1 \phi \, dx,
\end{aligned}$$

where $\hat{x}_1 = (x_2, \ldots, x_d)^t$.

Now, for general $\mathbf{a}$, construct an orthogonal matrix $Q = (q_1|q_2|\dots|q_d)$, where $q_1 = \frac{\mathbf{a}}{\|\mathbf{a}\_2|}$. Then,

$$\begin{aligned}\int_{\mathbb{R}^d} \sigma_0(\mathbf{a}\cdot x + c)\partial_i\phi(x)\,dx &= \int_{\mathbb{R}^d} \sigma_0(\|\mathbf{a}\|_2 y_1 + c)\sum_\ell (q_\ell)_i \partial_\ell\tilde{\phi}(y)\,dy \\ &= \int_{\mathbb{R}^d} \sigma_0(\|\mathbf{a}\|_2 y_1 + c)\frac{\mathbf{a}_i}{\|\mathbf{a}\|_2}\partial_1\tilde{\phi}(y)\,dy \\ &= -\int_{\{\|\mathbf{a}\|_2 y_1 + c\geq 0\}} \mathbf{a}_i\tilde{\phi}(y)\,dy \\ &= -\int_{\{\mathbf{a}\cdot x + c\geq 0\}} \mathbf{a}_i\phi(x)\,dx,\end{aligned}$$

where we set $y = Q^t x$ and $\tilde{\phi}(y) := \phi(Q^t x)$. Furthermore, we note that $\partial_i\phi = \sum_\ell (q_\ell)_i\partial_\ell\tilde{\phi}$.

As for the last statement, we see that

$$\int_{\mathbb{R}^d} f(x)\partial_i\phi\,dx = \sum_{\ell=1}^m \int_{\mathbb{R}^d} \sigma(a_\ell\cdot x + c_\ell)\partial_i\phi(x)\,dx = -\sum_{\ell=1}^m \left[\int_{a_\ell\cdot x + c_\ell\geq 0} \phi\,dx + \eta\int_{\{a_\ell\cdot x + c_\ell<0\}}\phi\,dx\right](a_\ell)_i.$$

□

Proposition 4.2.6. *For any $1\leq i,j\leq m$, $\phi\in C_c^\infty(\mathbb{R}^d)$, $\boldsymbol{a}\in\mathbb{R}^d$, and $c\in\mathbb{R}$,*

$$\int_{\mathbb{R}^d}\sigma(\boldsymbol{a}\cdot x + c)\partial_i\partial_j\phi\,dx = (1-\eta)\frac{\boldsymbol{a}_i\boldsymbol{a}_j}{\|\boldsymbol{a}\|_2}\int_{\{\boldsymbol{a}\cdot x + c = 0\}}\phi\,dx.$$

In particular, the second weak derivative of f satisfies $\nabla^2 f(x) = (1-\eta)\sum_{\ell=1}^m \frac{1}{\|a_\ell\|_2}\mathbb{1}_{\partial A_\ell} a_\ell\otimes a_\ell$.

Proof. Again, we note that $\sigma(a_\ell\cdot x + c_\ell) = \eta(a_\ell\cdot x + c_\ell) + (1-\eta)\sigma_0(a_\ell\cdot x + c_\ell)$, where the first term has zero second derivative. Thus, it suffices to consider the second

weak derivative of σ_0.

First, we start with the canonical case where $\mathbf{a} = a\ \mathbf{e}_1$. Then, other than the case $i = j = 1$,

$$\int_{\mathbb{R}^d} \sigma_0(ax_1 + c)\partial_i\partial_j\phi\, dx = 0.$$

For $i = j = 1$,

$$\begin{aligned}\int_{\mathbb{R}^d} \sigma_0(ax_1 + c)\partial_1^2\phi\, dx &= -a\int_{\mathbb{R}^{d-1}}\int_{ax_1+c\geq 0} \partial_1\phi\, dx_1 d\hat{x}_1 \\ &= sgn(a)a\int_{\mathbb{R}^{d-1}} \phi(\frac{-c}{a}, \hat{x}_1)\, d\hat{x}_1 \\ &= |a|\int_{ax_1+c=0} \phi(x)\, dx.\end{aligned}$$

Again, for general $\mathbf{a}$, we consider the orthogonal matrix $Q = (q_1|\dots|q_d)$ with $q_1 = \frac{\mathbf{a}}{\|\mathbf{a}\|_2}$. Noting that $\partial_i\partial_j\phi = \sum_\ell\sum_s (q_\ell)_i(q_s)_j\partial_\ell\partial_s\tilde{\phi}$, we have

$$\begin{aligned}\int_{\mathbb{R}^d} \sigma_0(\mathbf{a}\cdot x + c)\partial_i\partial_j\phi(x)\, dx &= \int_{\mathbb{R}^d} \sigma_0(\|\mathbf{a}\|_2 y_1 + c)\sum_\ell\sum_s (q_\ell)_i(q_s)_j\partial_\ell\partial_s\tilde{\phi}(y)\, dx \\ &= \|\mathbf{a}\|_2\frac{(\mathbf{a})_i(\mathbf{a})_j}{\|\mathbf{a}\|_2^2}\int_{\{\|\mathbf{a}\|_2 y_1+c=0\}} \tilde{\phi}(y)\, dy \\ &= \frac{(\mathbf{a})_i(\mathbf{a})_j}{\|\mathbf{a}\|_2}\int_{\mathbf{a}\cdot x+c=0} \phi(x)\, dx.\end{aligned}$$

□

4.2.3 Dimension Reduction

To recover the weights, we follow the procedures in Algorithm 4. In view of Proposition 4.2.5, we can create many instances of vectors in $span\{a_j\}_j$ by changing

the test function ϕ. Consider the following matrix

$$J_{\mathcal{V}}[f] = \int_{\mathcal{V}} (\int_{\mathbb{R}^d} f(x) \nabla \phi_\nu \, dx)(\int_{\mathbb{R}^d} f(x) \nabla \phi_\nu \, dx)^T d\pi(\nu),$$

where $(\mathcal{V}, \pi)$ is a probability space of test functions. To ensure non-degeneracy of $J_{\mathcal{V}}[f]$, we may let $\{\phi_\nu\}_{\nu \in \mathcal{V}}$ be a bounded resolution of identity, i.e. , a set of non-negative smooth and compactly supported functions ϕ_ν such that $\int_{\mathcal{V}} \phi_\nu(x) d\pi(\nu) = 1$ for all x in some pre-determined bounded set. For a discrete set $\mathcal{V}$, $\{\phi_\nu\}$ is a bounded partition of unity.

Note that if one draws the test functions ϕ_j of $Y_1 = (\Delta_{N,\phi_1}(f)| \cdots |\Delta_{N,\phi_{m_\chi}}(f))$ from $(\mathcal{V}, \pi)$, then $Y_1 \sim X_1 = (\nabla_{w,\phi_1}(f)| \cdots |\nabla_{w,\phi_{m_\chi}}(f))$. We shall show that if the m-th singular value $\sigma_m(J_{\mathcal{V}}[f]) \geq \alpha > 0$, then $\sigma_m(X_1)$ will be bounded from below with high probability.

Let $d\mu(x) = p(x)dx$ be a compactly supported probability measure on $\mathbb{R}^d$. Define, for $j = 1, 2$,

$$C_{\mathcal{V}} = \sup_{\nu \in \mathcal{V}} \sup_{x \in supp(\mu)} \frac{\phi_\nu(x)}{p(x)}, \quad C_{\nu,j} = \sup_{\nu \in \mathcal{V}} \sup_{x \in supp(\mu)} \| \frac{\nabla^2 \phi_\nu(x)}{p(x)} \|_2.$$

Lemma 4.2.7. *Assume that $\{a_j\}_j$ are linearly independent with $\max_j \|a_j\|^2 = C_0$ and $\sigma(J_{\mathcal{V}}[f]) \geq \alpha > 0$. Then, for any $s \in (0, 1)$ we have that*

$$\sigma_m(X_1) \geq \sqrt{m_\chi \alpha (1 - s)}$$

with probability at least $1 - m \exp(\frac{-m_\chi \alpha s^2}{2(C_0 C_{\mathcal{V}} m)^2})$.

Proof. Let $P^A : \mathbb{R}^d \to \mathbb{R}^m$ be a compression onto $A = span\{a_j\}$. In particular, we set the rows of P^A to be an orthonormal basis of A. Note that $\sigma_j(X_1) = \sigma_j(P^A X_1) = \sqrt{\sigma_j(P^A X_1 X_1^T (P^A)^T)}$, where

$$X_1 X_1^T = \sum_{k=1}^{m_\chi} \nabla_{w,\phi_k}(f) \nabla_{w,\phi_k}(f)^T.$$

Now, for each k,

$$\begin{aligned}
\sigma_1(P^A \nabla_{w,\phi_k}(f) \nabla_{w,\phi_k}(f)^T (P^A)^T) &= \|\nabla_{w,\phi_k}(f) \nabla_{w,\phi_k}(f)^T\|_F \\
&= \|\nabla_{w,\phi_k}(f)\|_2^2 \\
&= \|\sum_{j=1}^{m} (\int_{\{a_j \cdot x + c_j \geq 0\}} \phi_k \, dx + \eta \int_{\{a_j \cdot + c_j \leq 0\}} \phi \, dx) a_j\|_2^2 \\
&\leq (\sum_{j=1}^{m} C_\mathcal{V} \|a_j\|^2)^2 \leq m C_\nu^2 C_0^2.
\end{aligned}$$

Note that $\mathbb{E}(P^A \nabla_{w,\phi_k}(f) \nabla_{w,\phi_k}(f)^T (P^A)^T) = P^A J_\mathcal{V}[f] (P^A)^T$, and $\sigma_m(P^A J_\mathcal{V}[f] (P^A)^T) = \sigma(J_\mathcal{V}[f]) = \alpha$. Then, by Theorem 4.2.2, we see that

$$\mathbb{P}\left\{\sigma_m(P^A X_1 X_1^T (P^A)^T) \geq (1-s) m_\chi \alpha\right\} \leq m \exp(\frac{-m_\chi \alpha s^2}{2 m C_\mathcal{V}^2 C_0^2}).$$

Taking square-root, we have proven the claim. □

Now, we know that $Y_1 \sim X_1$, and the row space of X_1^T is $A = span\{a_j\}$ as long as $\sigma_m(X_1) > 0$. Let the row space of $\tilde{V_1}^T$ from Algorithm 4 be $\tilde{A}$. We shall compare the distance between the two spaces A and $\tilde{A}$. Thus, we have the following proposition:

Proposition 4.2.8. *Let $C_f = \sup_{x \in supp(\mu)} |f(x)|$. Suppose we sample N points $\{x_k\}$ according to μ and obtain $\{f(x_k)+n_k\}$, where n_k are independent zero-mean random variables with $|n_k| \leq C_{\mathcal{N}}$ almost surely. Define $P_A, P_{\tilde{A}}$ be the orthogonal projection to $A, \tilde{A}$, respectively. Then,*

$$\|P_A - P_{\tilde{A}}\|_F \leq \frac{2\eta}{\sqrt{(1-s)\alpha} - \eta}$$

with probability at least $1 - m\exp(\frac{-m_\chi \alpha s^2}{2(C_0 C_{\mathcal{V}} m)^2}) - (d+1)m_\chi \exp(\frac{-\eta^2 N}{2(K^2 + K\eta/3)})$, where $K = C_{\mathcal{V},1}(C_f + C_{\mathcal{N}}) + mC_{\mathcal{V},1}C_0$.

Proof. Let

$$Y_1^T = \begin{pmatrix} \bar{U}_1 & \bar{U}_2 \end{pmatrix} \begin{pmatrix} \bar{\Sigma}_1 & \\ & \bar{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \bar{V}_1^T \\ \bar{V}_2^T \end{pmatrix}, \quad X_1^T = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix}$$

where $\tilde{\Sigma}_1, \Sigma_1 \in \mathbb{R}^{m \times m}$. By construction, $\Sigma_2 = 0$, so we may choose $\bar{\alpha} = \sigma_m(Y_1)$, and by Theorem 4.2.3, we have

$$\|P_A - P_{\tilde{A}}\|_F = \|\tilde{V}_1\tilde{V}_1^T - V_1 V_1^T\|_F \leq \frac{2\|Y_1 - X_1\|_F}{\sigma_m(Y_1)} \leq \frac{2\|Y_1 - X_1\|_F}{\sigma_m(X_1) - \|Y_1 - X_1\|_F}, \quad (4.2)$$

where we used Weyl's estimate on $\sigma_m(Y_1)$.

From Lemma 4.2.7, we see that $\sigma_m(X_1) \geq \sqrt{m_\chi \alpha(1-s)}$ with probability at least $1 - m\exp(\frac{-m_\chi \alpha s^2}{2(C_0 C_{\mathcal{V}} m)^2})$. Now, we shall estimate $\|X_1 - Y_1\|_F$.

Now, $\|X_1 - Y_1\|_F^2 = \sum_{j=1}^{m_\chi} \|(X_1)_j - (Y_1)_j\|_2^2$, where

$$\|X_j - Y_j\|_2 = \|\frac{1}{N}\sum_{k=1}^{N}(f(x_k)+n_k)\frac{\nabla\phi_j(x_k)}{p(x_k)} - \int_{\mathbb{R}^d} f(x)\frac{\nabla\phi_j(x)}{p(x)}p(x)\,dx\|_2 = \|\frac{1}{N}\sum_{k=1}^{N}\chi_{k,j}\|,$$

where

$$\chi_{k,j} = (f(x_k)+n_k)\frac{\nabla\phi_j(x_k)}{p(x_k)} - \int_{\mathbb{R}^d} f(x)\frac{\nabla\phi_j(x)}{p(x)}p(x)\,dx.$$

Now,

$$\begin{aligned}
\|\chi_{k,j}\|_2 &= \|(f(x_k)+n_k)\frac{\nabla\phi_j(x_k)}{p(x_k)} - \int_{\mathbb{R}^d} f(x)\frac{\nabla\phi_j(x)}{p(x)}p(x)\,dx\|_2 \\
&= \|(f(x_k)+n_k)\frac{\nabla\phi_j(x_k)}{p(x_k)} - \sum_{j=1}^{m}(\int_{\{a_j\cdot x+c_j\geq 0\}}\phi_k\,dx + \eta\int_{\{a_j\cdot+c_j\leq 0\}}\phi\,dx)a_j\|_2 \\
&\leq (C_f + C_{\mathcal{N}})C_{\mathcal{V},1} + mC_0C_{\mathcal{V}} =: K,
\end{aligned}$$

and

$$\max\{\|\sum_{k=1}^{N}\mathbb{E}(\chi_{k,j}\chi_{k,j}^T)\|, \|\sum_{k=1}^{N}\mathbb{E}(\chi_{k,j}^T\chi_{k,j})\|\} \leq \sum_{k=1}^{N}\mathbb{E}\|\chi_{k,j}\|^2 \leq NK^2 =: \sigma^2.$$

Thus, using Corollary 4.2.4, one has

$$\mathbb{P}(\|\frac{1}{N}\sum_{k=1}^{N}\chi_{k,j}\|_2 > \eta) \leq (d+1)\exp(\frac{-\eta^2 N}{2(K^2+K\eta/3)}),$$

and $\|X_1 - Y_1\|_F^2 = \sum_{j=1}^{m_\chi}\|(X_1)_j - (Y_1)_j\|_2^2 \leq m_\chi\eta^2$ with probability at least $1 - m_\chi(d+1)\exp(\frac{-\eta^2 N}{2(K^2+K\eta/3)})$. Plugging it back in (4.2), we can prove the claim.

□

After obtaining an approximation $\tilde{A}$ to $A = span\{a_1, \ldots, a_m\}$, we are now able to reduce our ambient space $\mathbb{R}^d$ to $\mathbb{R}^m$ if $d > m$:

Proposition 4.2.9 ([33] Theorem 1.1)**.** *Consider*

$$f(x) = \sum_{i=1}^{m} g_i(a_i \cdot x), \quad x \in B_1^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\},$$

where $m \leq d$. Let $\tilde{A} = span\{b_1, \ldots, b_m\}$ where $\{b_j\}$ form an orthonormal basis in $\tilde{A}$. Let $B = (b_1|\cdots|b_m)$. Then, one can construct

$$\tilde{f} = \sum_{i=1}^{m} \tilde{g}_i(\alpha_i \cdot y), \quad y \in B_1^m \subset \mathbb{R}^m,$$

with $\alpha_i = B^T a_i$. Then, for all $\hat{f} : \mathbb{R}^m \to \mathbb{R}$, the following estimate holds

$$\|f - \hat{f}(B^T \cdot)\|_\infty \leq \|f\|_{Lip} \|P_A - P_{\tilde{A}}\|_F + \|\tilde{f} - \hat{f}\|_\infty.$$

Moreover, for any other set of vectors $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_m\} \subset \mathbb{R}^m$,

$$\|\alpha_i - B\hat{\alpha}_i\|_2 \leq \|P_A - P_{\tilde{A}}\|_F + \|\alpha_i - \hat{\alpha}_i\|_2.$$

In particular, from now on we may assume that $d = m$.

4.2.4 Second Order Derivative

Again, we draw randomly $\{\phi_j\}$ from $(\mathcal{V}, \pi)$ and $\{x_k\}_{k=1}^N$ from $B_1^d(0)$. Then, we define $Y_2 = (vec(\Delta^2_{N,\phi_j}(f)))_{j=1}^{m_\chi}$ to approximate $X_2 = (vec(\nabla^2_{w,\phi_j}(f)))_{j=1}^{m_\chi}$. Suppose

that

$$H[f] = \int_{\mathcal{V}} \int_{\mathbb{R}^d} f(x) \nabla^2 \phi_\nu \, dx \, d\pi(\nu)$$

satisfies $\sigma_m(H[f]) = \alpha_2 > 0$. Then, we have the following results:

Lemma 4.2.10. *Suppose $\sigma_m(H[f]) = \alpha_2 > 0$, and $d\mu(x) = p(x)dx$ is the uniform measure on $B_1^m(0)$, then for any $s \in (0,1)$, X_2 satisfies*

$$\sigma_m(X_2) \geq \sqrt{m_\chi \alpha_2 (1-s)}$$

with probability at least $1 - m \exp(\frac{-m_\chi \alpha s^2}{2\bar{C}})$, where $\bar{C} = (1-\eta)^2 m^2 C_0 C_{\mathcal{V}}^2 \frac{\Gamma(m/2+1)}{\sqrt{\pi}\Gamma((m-1)/2+1)}$.

Proof. The proof follows almost the same derivation of Lemma 4.2.7. The only difference is to estimate

$$\begin{aligned}
\sigma_1(vec(\nabla^2_{w,\phi_j}(f))vec(\nabla^2_{w,\phi_j}(f))^T) &= \| \int_{\mathbb{R}^d} f(x) \nabla_j^2(x) \, dx \|_2^2 \\
&= (1-\eta)^2 \| \sum_{i=1}^m \int_{\{a_i \cdot x + c_i = 0\}} \frac{\phi_j(x)}{p(x)} p(x) \, dx \frac{a_i a_i^T}{\|a_i\|_2} \|_2^2 \\
&\leq (1-\eta)^2 m^2 C_0^2 C_{\mathcal{V}}^2 \frac{\lambda(B_1^{(m-1)}(0))}{\lambda(B_1^m(0))} \\
&= (1-\eta)^2 m^2 C_0 C_{\mathcal{V}}^2 \frac{\Gamma(m/2+1)}{\sqrt{\pi}\Gamma((m-1)/2+1)},
\end{aligned}$$

where we note that $\{a_i \cdot x + c_i = 0\} \cap B_1^m(0)$ is a cross-section of $B_1^m(0)$ and has volume at most $\lambda(B_1^{m-1}(0))$. □

Then, we can quantify the goodness of approximation Y_2 to X_2:

Proposition 4.2.11. *Suppose we sample N points $\{x_k\}$ according to the uniform distribution μ and obtain $\{f(x_k) + n_k\}$, where n_k are independent zero-mean random*

variables with $|n_k| \leq C_{\mathcal{N}}$ *almost surely. Define* $P_{\mathscr{A}}, P_{\tilde{\mathscr{A}}}$ *be the orthogonal projection to* $\mathscr{A}, \tilde{\mathscr{A}}$*, respectively. Then,*

$$\|P_{\mathscr{A}} - P_{\tilde{\mathscr{A}}}\|_F \leq \frac{2\eta}{\sqrt{(1-s)\alpha} - \eta}$$

with probability at least $1 - m\exp(\frac{-m_\chi \alpha s^2}{2\bar{C}}) - 2mm_\chi \exp(\frac{-\eta^2 N}{2(\bar{K}^2 + \bar{K}\eta/3)})$*, where* $\bar{K} = C_{\mathcal{V},2}(C_f + C_{\mathcal{N}}) + \sqrt{\bar{C}}$*.*

4.2.5 Whitening Process

Before we recover $\{a_j\}_j$ from $\tilde{\mathscr{A}}$, it is possible to use the following whitening process such that $\{a_j\}$ be comes nearly orthonormal. This step is important in the sense that the weight recovery is guaranteed when $\{a_j\}$ is nearly orthonormal.

We first introduce the following definition:

Definition 4.2.12. Let $a_1, \ldots, a_m$ be unit vectors. Then, we define

$$S(a_1, \ldots, a_m) = \inf\{(\sum_{i=1}^{m} \|a_i - w_i\|_2^2)^{1/2} : \{w_1, \ldots, w_m\} \text{ forms an orthonormal basis in } \mathbb{R}^m\}.$$

The following theorem gives a guideline on how to make $\{a_j\}_j$ nearly orthonormal.

Proposition 4.2.13 ([33], Theorem 3.7)**.** *Let* $\gamma, \eta > 0$ *be positive real numbers. Let* $\|P_{\mathscr{A}} - P_{\tilde{\mathscr{A}}}\|_F \leq \eta$ *and let* $\tilde{G} \in \tilde{\mathscr{A}}$ *be positive definite with* $\tilde{G} \geq \gamma I_m$*. Consider the*

spectral decomposition $\tilde{G} = UDU^T$ *and let* $W = D^{-1/2}U^T$. *Then,*

$$S(\frac{Wa_1}{\|Wa_1\|}, \dots, \frac{Wa_m}{\|Wa_m\|}) \leq \frac{\sqrt{2}\eta\|\tilde{G}\|_F}{\gamma}.$$

To use Proposition 4.2.13, one needs to search for positive definite matrices in $\tilde{\mathscr{A}}$ with preferably large γ. It can be satisfied with the following optimization scheme

$$\max_{\tilde{A}\in\tilde{\mathscr{A}},\|\tilde{A}\|_F=1} \min_{x\in\mathbb{R}^m,\|x\|_2=1} x^T\tilde{A}x. \tag{4.3}$$

For the sake of completeness, we detail the whitening algorithm and demonstrate how we can manipulate $\{a_j\}$ without knowing them *a priori*.

Algorithm 5 ([33], Algorithm 3.2)**.**

- Fix $\eta > 0$ and let $f^{(0)}(x) = \sum_{i=1}^m g_i^{(0)}(a_i^{(0)} \cdot x)$.
- Compute $\tilde{\mathscr{A}}^{(k+1)}$ by using Proposition 4.2.11 with accuracy $\eta > 0$ from point values of $f^{(k)}$.
- Define $W^{(k+1)}$ as the whitening matrix of the vectors $\{a_i^{(k)}\}_{i=1}^m$ by using $\tilde{\mathscr{A}}^{(k+1)}$ and (4.3).
- Define $a_i^{(k+1)} = W^{(k+1)}a_i^{(k)}/\|W^{(k+1)}a_i^{(k)}\|_2$ for all i.
- Denote $f^{(k+1)}(x) = \sum_{i=1}^m g_i^{(k+1)}(a_i^{(k+1)} \cdot x) = f^{(k)}((W^{(k+1)})^T x)$.

4.2.6 Weight Recovery

After whitening process, we may assume that $\{a_j\}$ are nearly orthonormal. Here, we shall introduce the following optimization scheme:

$$\arg\max \|M\|_\infty, \quad \|M\|_F \leq 1,\ M \in \tilde{\mathscr{A}}, \tag{4.4}$$

where $\|M\|_\infty := \max_{x:\,\|x\|_2=1} \|Mx\|_2$. We can assume that

- There exists an orthonormal basis $\{w_1, \ldots, w_m\}$ such that $(\sum_{j=1}^m \|a_j - w_j\|_2^2)^{1/2} = \epsilon > 0$.
- $\hat{\mathscr{A}} = span\{w_j \otimes w_j\}_j$.
- $\|P_{\mathscr{A}} - P_{\tilde{\mathscr{A}}}\|_F \leq \eta$.
- $\|P_{\hat{\mathscr{A}}} - P_{\tilde{\mathscr{A}}}\|_F \leq 4\epsilon + \eta =: \nu$.

Then, we recover the weights $\{a_j\}$ by the following algorithm:

Algorithm 6 ([33], Algorithm 3.3)**.**

- Let M be a local maximizer of (4.4).
- If $\|M\|_\infty$ is not an eigenvalue of M, replace M by $-M$.
- Denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ the eigenvalues of M arranged in decreasing order.

- Take the eigenvalue decomposition of M, i.e. $M = \sum_{j=1}^{m} \lambda_j u_j \otimes u_j$.
- Put $\hat{a} := u_1$.

The performance of Algorithm 6 is guaranteed by the following theorem.

Theorem 4.2.14 ([33], Theorem 3.12)**.** *If $0 < \nu < 1/(cm)$ for a suitable constant $c > 6$, then there exists $j_0 \in \{1, \ldots, m\}$ such that the vector $\hat{a}$ found by Algorithm 6 satisfies $\min_{s \in \{-1,1\}} \|s\hat{a} - a_{j_0}\|_2 \leq 5\nu$.*

4.2.7 Function Recovery

Now, it is possible to recovery f:

Algorithm 7. ·

- Let $\hat{a}_j$ be the normalized approximations of a_j, $j = 1, \ldots, m$.
- Let $\{\hat{b}_j\}_j$ be the dual basis to $\{\hat{a}_j\}_j$.
- Put $\hat{g}_j(t) := f(t\hat{b}_j)$, $t \in (-1/\|\hat{b}_j\|_2, 1/\|\hat{b}_j\|_2)$.
- Put $\hat{f}(x) := \sum_{j=1}^{m} \hat{g}_j(\hat{a}_j \cdot x)$, $\|x\|_2 \leq 1$.

Theorem 4.2.15 ([33], Theorem 4.1)**.** *Let $S(a_1, \ldots, a_m) \leq \epsilon$, $S(\hat{a}_1, \ldots, \hat{a}_m) \leq \epsilon'$, and $(\sum_{j=1}^{m} \|a_j - \hat{a}_j\|_2^2)^{1/2} \leq \eta$. Then $\hat{f}$ constructed in Algorithm 7 satisfies*

$$\|f - \hat{f}\|_\infty \leq 5C_2(1 + \xi(\epsilon, \epsilon')) \max(\eta, \eta^2),$$

where $\xi(\epsilon, \epsilon') \to 0$ as $(\epsilon, \epsilon') \to (0, 0)$.

4.3 Two-Layer Case

For neural networks with two hidden layers, functions take the following form

$$f(x) = \sum_{\ell=1}^{m_1} \alpha_\ell \sigma(b_\ell \cdot g(x) + d_\ell) = h(g(x)), \quad (g(x))_j = \sigma(a_j \cdot x + c_j),\ j = 1, \ldots, m_0.$$

We shall show that we may assume that $\{a_j\}, \{b_\ell\}$ are all unit vectors.

Proposition 4.3.1. *Given any two-layer neural network $f : \mathbb{R}^d \to \mathbb{R}$, we have*

$$f(x) = \sum_{\ell=1}^{m_1} \alpha_i \sigma(b_\ell \cdot g(x) + d_\ell) = \sum_{\ell=1}^{m_1} \tilde{\alpha}_i \sigma(\tilde{b}_\ell \cdot \tilde{g}(x) + \tilde{d}_\ell),$$

where $(g(x))_j = \sigma(a_j \cdot + c_j)$, $(\tilde{g}(x))_j = \sigma(\frac{a_j}{\|a_j\|_2} x + \frac{c_j}{\|a_j\|})$, $(\tilde{b}_\ell)_j = \frac{(b_\ell)_j \|a_j\|_2}{C_\ell}$, $\tilde{d}_\ell = \frac{d_\ell}{C_\ell}$, $\tilde{\alpha}_\ell = C_\ell \alpha_\ell$, *and* $C_\ell = \sqrt{\sum_j (b_\ell)_j^2 \|a_j\|^2}$.

Proof. It follows from the fact that when σ is a leaky ReLU function, $\sigma(cx) = c\sigma(x)$ for any $c \geq 0$. □

To evaluate its first order derivative, we refer to the following result on the composition of distributions:

Theorem 4.3.2 ([3] Corollary 3.1)**.** *Given an open set $\Omega \subset \mathbb{R}^d$, let $g \in BV(\Omega)$ be of bounded variation. If h is Lipschitz continuous, then $f = h \circ g$ satisfies*

$$\nabla f = \nabla h \cdot \nabla g.$$

Using Theorem 4.3.2 and Proposition 4.2.5, we have the following proposition:

Proposition 4.3.3. *The weak derivative $\partial_j f$ satisfies*

$$\partial_j f(x) = \sum_{\ell=1}^{m_0}\sum_{s=1}^{m_0}(b_\ell)_s(a_s)_j(\eta+(1-\eta)\mathbb{1}_{\{b_\ell\cdot g(x)+d_\ell\geq 0\}})(\eta+(1-\eta)\mathbb{1}_{\{a_s\cdot x+c_s\geq 0\}}).$$

Proof. It suffices to note that $\nabla h = \sum_{\ell=1}^{m_0} b_\ell^t(\eta+(1-\eta)\mathbb{1}_{\{b_\ell\cdot g(x)+d_\ell\geq 0\}})$ and the s-th row of ∇g is $(\nabla g)_s = a_s^t(\eta+(1-\eta)\mathbb{1}_{\{a_s\cdot x+c_s\geq 0\}})$. □

The Hessian of f is arguably the most interesting part. In [34], the *entangled* weights appear, which involve both $\{a_s\}_s$ and $\{b_\ell\}_\ell$ and varies for different x. The same thing happens more or less, but currently it is not yet enough for direct application of their methods.

Proposition 4.3.4. *For any $1\leq i,j\leq d$ and $\phi\in C_c^\infty(\mathbb{R}^d)$,*

$$\begin{aligned}\int_{\mathbb{R}^d} f(x)\partial_i\partial_j\phi\,dx = \sum_{s=1}^{m_0}\Bigg[\sum_{\ell=1}^{m_0}(b_\ell)_s\Bigg((1-\eta)\int_{\{a_s\cdot x+c_s=0\}\cap\{b_\ell\cdot g(x)+d_\ell\geq 0\}}\phi\,dx \\ +\eta\int_{\{a_s\cdot x+c_s=0\}}\phi\,dx\Bigg)\Bigg]\frac{(a_s)_i(a_s)_j}{\|a_s\|_2} \\ +\sum_{\alpha\in\{0,1\}^m}\sum_{\ell=1}^{m_0}C_\alpha^\ell\frac{(v_\alpha^\ell)_i(v_\alpha^\ell)_j}{\|v_\alpha^\ell\|_2},\end{aligned}$$

where $C_\alpha^\ell = \int_{D_\alpha^\ell}\phi\,dx$, $v_\alpha^\ell = \sum_t(\eta+(1-\eta)\alpha_t)(b_\ell)_t a_t$, and $D_\alpha^\ell = \{x\in\mathbb{R}^d : b_\ell\cdot g(x)+d_\ell = 0,\ g_t(x) = (a_s\cdot x+c_s)(\eta+(1-\eta)\alpha_t)\}$.

In particular,

$$\int_{\mathbb{R}^d} f(x)\nabla^2\phi(x)\,dx = \sum_{s=1}^{m_0}\frac{\gamma_s}{\|a_s\|_2}a_s\otimes a_s + \sum_{\ell=1}^{m_0}\sum_{\alpha\in\{0,1\}^{m_0}}\frac{C_\alpha^\ell}{\|v_\alpha^\ell\|_2}v_\alpha^\ell\otimes v_\alpha^\ell.$$

Proof. The proof involves partition of $\mathbb{R}^d$ according to the sets $A_s = \{a_s\cdot+c_s\geq 0\}$

and $B^\ell = \{B_\ell \cdot g(x) + d_\ell \geq 0\}$. To simplify the proof, we shall implicitly partition ϕ with partition of unity to avoid the artificial boundaries created during the process.

By Proposition 4.3.3, we know that

$$\begin{aligned}
&\int_{\mathbb{R}^d} f(x)\partial_i\partial_j\phi(x)\,dx \\
&= -\sum_{\ell=1}^{m_0}\sum_{s=1}^{m_0}\int_{\mathbb{R}^d}(b_\ell)_s(a_s)_j(\eta\mathbb{1}_{\{b_\ell\cdot g(x)+d_\ell\leq 0\}}+\mathbb{1}_{\{b_\ell\cdot g(x)+d_\ell\geq 0\}})(\eta\mathbb{1}_{\{a_s\cdot x+c_s\leq 0\}}+\mathbb{1}_{\{a_s\cdot x+c_s\geq 0\}})\partial_i\phi(x)\,dx.
\end{aligned} \tag{4.5}$$

Since $u_1(x) = (\eta+(1-\eta)\mathbb{1}_{\{b_\ell\cdot g(x)+d_\ell\geq 0\}})$ and $u_2(x) = (\eta+(1-\eta)\mathbb{1}_{\{a_s\cdot x+c_s\geq 0\}})$ are of bounded variation locally, we may compose $(u_1, u_2)^t$ with multiplication function $M(x, y) = xy$. Since M is again Lipschitz, we may use Theorem 4.3.2 to deduce the product rule for $u_1(x)u_2(x)$ by

$$\begin{aligned}
&-\sum_{\ell=1}^{m_0}\sum_{s=1}^{m_0}(b_\ell)_s(a_s)_j\int_{\mathbb{R}^d}u_1(x)u_2(x)\partial_i\phi\,dx \\
&=\sum_{\ell=1}^{m_0}\sum_{s=1}^{m_0}(b_\ell)_s(a_s)_j\int_{\mathbb{R}^d}(1-\eta)u_1(x)\partial_i\mathbb{1}_{\{a_s\cdot x+c_s\geq 0\}}\phi\,dx \\
&\quad+\sum_{\ell=1}^{m_0}\sum_{s=1}^{m_0}(b_\ell)_s(a_s)_j\int_{\mathbb{R}^d}(1-\eta)\partial_i\mathbb{1}_{\{b_\ell\cdot g(x)+d_\ell\geq 0\}}u_2(x)\phi\,dx \\
&=: I_1 + I_2,
\end{aligned}$$

where we note that $\partial_i u_1 = (1-\eta)\partial_i\mathbb{1}_{\{b_\ell\cdot g(x)+d_\ell\geq 0\}}$, and $\partial_i u_2 = (1-\eta)\partial_i\mathbb{1}_{\{a_s\cdot x+c_s\geq 0\}}$.

The first term I_1 is rather straightforward, and we can see that $I_1 = (1-\eta)\sum_s \frac{(a_s)_i(a_s)_j}{\|a_s\|_2}\sum_\ell(b_\ell)_s((1-\eta)\int_{\partial A_s\cap B^\ell}\phi\,dx + \eta\int_{\partial A_s\cap(B^\ell)^c}\phi\,dx$.

For I_2, it is a little more complicated as B_ℓ is not a half space. In fact, it may not even be convex if b_ℓ is not positive. However, locally B_ℓ may still resemble a

half space.

Define, for $\alpha \in \{0,1\}^{m_0}$, $A_\alpha = \{x \in \mathbb{R}^d : \ g_t(x) = (a_t \cdot x + c_t)(\eta + (1-\eta)\alpha_t) \, \forall t\}$. In particular, on A_α, $a_s \cdot x + c_s \geq 0$ if and only if $\alpha_s = 1$. Then,

$$\begin{aligned} I_2 &= \int_{\mathbb{R}^d} \partial_i \mathbb{1}_{\{b_\ell \cdot g(x) + d_\ell \geq 0\}} \left(\sum_{s=1}^{m_0} (b_\ell)_s (a_s)_j (\eta + (1-\eta) \mathbb{1}_{\{a_s \cdot x + c_s \geq 0\}}) \right) \phi(x) dx \\ &= \sum_{\alpha \in \{0,1\}^{m_0}} \int_{A_\alpha} \partial_i \mathbb{1}_{\{b_\ell \cdot g(x) + d_\ell \geq 0\}} \left(\sum_{s=1}^{m_0} (b_\ell)_s (a_s)_j (\eta + (1-\eta) \mathbb{1}_{a_s \cdot x + c_s \geq 0}) \right) \phi(x) dx \\ &= \sum_{\alpha \in \{0,1\}^{m_0}} \int_{A_\alpha} \partial_i \mathbb{1}_{\{b_\ell \cdot g(x) + d_\ell \geq 0\}} \left(\sum_{t} (b_\ell)_t (a_t)_j (\eta + (1-\eta)\alpha_t) \right) \phi(x) dx. \end{aligned}$$

On A_α, $b_\ell \cdot g(x) + d_\ell = v_\alpha^\ell \cdot x + d_{\alpha,\ell}$, where $d_{\alpha,\ell} = \sum_t (b_\ell)_t c_t (\eta + (1-\eta)\alpha_t) + d_\ell$. Thus,

$$\begin{aligned} I_2 &= \sum_{\alpha \in \{0,1\}^{m_0}} \int_{A_\alpha} \partial_i \mathbb{1}_{\{b_\ell \cdot g(x) + d_\ell \geq 0\}} (v_\alpha^\ell)_j \phi(x) dx \\ &= \sum_{\alpha \in \{0,1\}^{m_0}} (v_\alpha^\ell)_j \frac{(v_\alpha^\ell)_i}{\|v_\alpha^\ell\|_2} \int_{A_\alpha \cap \partial B^\ell} \phi(x) dx. \end{aligned}$$

□

4.3.1 Some Auxiliary Results

As we mentioned above, $B^\ell = \{b_\ell \cdot g(x) + d_\ell \geq 0\}$ is not a half plane in general. Here, we prove some properties of B^ℓ.

Proposition 4.3.5. *Given $\boldsymbol{b} \in \mathbb{R}^{m_0}$ and $d \in \mathbb{R}$, if there exists i_0 such that $\boldsymbol{b}_{i_0} > 0$, then $\mathcal{B} = \{ \ \boldsymbol{b} \cdot g(x) + d \geq 0\}$ is path connected and unbounded in $\mathbb{R}^{m_0}$ where*

$g(x)_s = \sigma(a_s \cdot x + c_s)$ *with* $\{a_s\}_s$ *linearly independent.*

Proof. Without loss of generality, we may assume that $i_0 = 1$. Since $\{a_s\}_s$ are linearly independent, the subspace $V = span\{a_s : s \geq 2\}$ is a proper subspace of $\mathbb{R}^d$ and $a_1 \notin V$. Then, consider $\tilde{a} = a_1 - P_V(a_1)$ where P_V is the orthogonal projection onto V. For large enough $C > 0$, $C\tilde{a} \in \mathcal{B}$ since $\mathbf{b}_1 g_1(C\tilde{a}) = \mathbf{b}_1 \sigma(Ca_1 \cdot \tilde{a} + c_1) = \mathbf{b}_1 \sigma(C\|\tilde{a}\|^2 + c_1)$ will be very large while the remaining terms are unchanged as $C \to \infty$, due to the fact that $\tilde{a} \in V^\perp$. Thus, $\mathcal{B}$ is unbounded.

Given $x_1, x_2 \in \mathcal{B}$, choose $\alpha > 0$ large enough so that $\min\{\ \mathbf{b}_1 g_1(x_1 + \alpha\tilde{a}),\ \mathbf{b}_1 g_1(x_2 + \alpha\tilde{a})\} \geq \max\{\sum_{s\geq 2} |\ \mathbf{b}_s| g_s(x_1), \sum_{s\geq 2} |\ \mathbf{b}_2| g_s(x_2)\} + d$. It is possible since for any $x \in \mathbb{R}^d$, $g_1(x + \alpha\tilde{a}) \to \infty$ as $\alpha \to \infty$. Then, the line segment connecting $\tilde{x}_i := x_i + \alpha\tilde{a}$ for $i = 1, 2$ lies completely inside $\mathcal{B}$. In particular, for $t \in (0, 1)$,

$$\begin{aligned}
\mathbf{b} \cdot g\left((t\tilde{x}_1 + (1-t)\tilde{x}_2)\right) &= \mathbf{b}_1 g_1(t\tilde{x}_1 + (1-t)\tilde{x}_2) + \sum_{s\geq 2} \mathbf{b}_s g_s(t\tilde{x}_1 + (1-t)\tilde{x}_2) \\
&\geq \mathbf{b}_1 g_1(t\tilde{x}_1 + (1-t)\tilde{x}_2) - \sum_{s\geq 2} |\ \mathbf{b}_s| g_s(t\tilde{x}_1 + (1-t)\tilde{x}_2) \\
&\geq t\left(\ \mathbf{b}_1 g_1(\tilde{x}_1) - \sum_{s\geq 2} |\ \mathbf{b}_s| g_s(\tilde{x_1})\right) \\
&\quad + (1-t)\left(\ \mathbf{b}_1 g_1(\tilde{x}_2) - \sum_{s\geq 2} |\ \mathbf{b}_s| g_s(\tilde{x_2})\right) \\
&> 0,
\end{aligned}$$

where we use the fact in the second inequality that g_s are convex and g_1 is affine on the segment from $\tilde{x}_1$ to $\tilde{x}_2$. Thus, the line segment from x_1 to $\tilde{x}_1$ to $\tilde{x}_2$ to x_2 lies entirely in $\mathcal{B}$.

□

To prove Proposition 4.3.4, we need to understand the weak derivative of the product of two indicator functions of half planes.

Lemma 4.3.6. *Given $a_1, a_2 \in \mathbb{R}^d$, $c_1, c_2 \in \mathbb{R}$, and $\phi \in C_c^\infty(\mathbb{R}^d)$,*

$$\int_{\mathbb{R}^d} \mathbb{1}_{\{a_1 \cdot x + c_1 \geq 0\}} \mathbb{1}_{a_2 \cdot x + c_2 \geq 0} \partial_i \phi \, dx = C_1 a_1 + C_2 a_2,$$

for some $C_1, C_2 \in \mathbb{R}$ independent of i.

Proof. As before, we first examine a simpler case before generalizing the result. Given $a_1, c_1, a_{21}, a_{22}, c_2 \in \mathbb{R}$ and $i > 2$,

$$\int_{\mathbb{R}^d} \mathbb{1}_{\{a_1 x_1 + c_1 > 0\}} \mathbb{1}_{\{a_{21} x_1 + a_{22} x_2 + c_2 > 0\}} \partial_i \phi \, dx = 0.$$

For $i = 1$,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \mathbb{1}_{a_1 x_1 + c_1 > 0} \mathbb{1}_{a_{21} x_1 + a_{22} x_2 + c_2 > 0} \partial_1 \phi \, dx \\
&= \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} \mathbb{1}_{a_1 x_1 + c_1 > 0} \mathbb{1}_{a_{21} x_1 + a_{22} x_2 + c_2 > 0} \partial_1 \phi \, dx_1 d\hat{x}_1 \\
&= \int_{\mathbb{R}^{d-1}} \left(\phi(\frac{-c_1}{a_1}, \hat{x}_1) - \phi(\frac{-(a_{22} x_2 + c_2)}{a_{21}}, \hat{x}_1) \right) \mathbb{1}_{\{\frac{-(a_{22} x_2 + c_2)}{a_{21}} > \frac{-c_1}{a_1}\}} d\hat{x}_1 \\
&= sgn(a_1) \int_{\{a_1 x_1 + c_1 = 0\} \cap \{a_{21} x_1 + a_{22} x_2 + c_2 > 0\}} \phi \, dx \\
&\quad + sgn(a_{21}) \int_{\{a_1 x_1 + c_1 > 0\} \cap \{a_{21} x_1 + a_{22} x_2 + c_2 = 0\}} \phi \, dx.
\end{aligned}$$

For $i = 2$,

$$\int_{\mathbb{R}^d} \mathbb{1}_{a_1 x_1 + c_1 > 0} \mathbb{1}_{a_{21} x_1 + a_{22} x_2 + c_2 > 0} \partial_2 \phi \, dx = sgn(a_{22}) \int_{\{a_1 x_1 + c_1 > 0\} \cap \{a_{21} x_2 + a_{22} x_2 + c_2 = 0\}} \phi \, dx.$$

Thus, for general $a_1, a_2 \in \mathbb{R}^d$, $c_1, c_2 \in \mathbb{R}$, we have, by change of variable $y = Q^t x$ with $Q = (q_1 \mid \cdots \mid q_d)$ and $q_1 = a_1/\|a_1\|_2$, $q_2 = \frac{a_2 - \frac{a_1 \cdot a_2}{\|a_1\|_2^2} a_1}{\|a_2 - \frac{a_1 \cdot a_2}{\|a_1\|_2^2} a_1\|_2}$,

$$
\begin{aligned}
&\int_{\mathbb{R}^d} \mathbb{1}_{\{a_1 \cdot x + c_1 > 0\}} \mathbb{1}_{\{a_2 \cdot x + c_2 > 0\}} \partial_i \phi \, dx \\
&= \int_{\mathbb{R}^d} \mathbb{1}_{\{\|a_1\|_1 y_1 + c_1 > 0\}} \mathbb{1}_{\{\frac{a_1 \cdot a_2}{\|a_1\|_2} y_1 + a_2 \cdot q_2 y_2 + c_2 > 0\}} \sum_s (q_s)_i \partial_s \tilde{\phi}(y) \, dy \\
&= (q_1)_i \left(\int_{\{a_1 \cdot x + c_1 = 0\} \cap \{a_2 \cdot x + c_2 > 0\}} \phi \, dx + sgn(a_1 \cdot a_2) \int_{\{a_1 \cdot x + c_1 > 0\} \cap \{a_2 \cdot x + c_2 = 0\}} \phi \, dx \right) \\
&\quad + (q_2)_i \int_{\{a_1 \cdot x + c_1 > 0\} \cap \{a_2 \cdot x + c_2 = 0\}} \phi \, dx \\
&= C_1 (a_1)_i + C_2 (a_2)_i,
\end{aligned}
$$

by splitting q_1, q_2 into linear combinations of a_1, a_2 and noting that $a_2 \cdot q_2 > 0$ by construction.

□

4.3.2 Net Spreading

As we can see in Proposition 4.3.4, the number of terms can be potentially up to $m_0 + m_1 2^{m_0} \gg m_0^2$, whose span may be the whole space. We will need to find a way to make the number of terms smaller, or we will not be able to reconstruct the terms. Fornasier et al. would have run into the same problem if they used passive sampling, but the nature of ReLU actually helps us circumvent the difficulty. We shall name the following tactic *net spreading*:

1. We first obtain many samples, say uniformly.

2. We compute weak differentiation with test functions of large support.

3. Partition the space sequentially into small cubes. Each time, we produce test functions with support about the size of the cube.

4. Search for active cubes: that is, the cubes that return non-trivial weak differentiation.

5. For active cubes, we choose many different test functions with roughly the same support. After getting many "weak samples", we perform PCA on it to find the dominating terms.

6. Partition the active cubes and perform yet again the searching.

7. Stop partitioning when an active cube returns only one dimension via PCA.

In order to perform the procedure, we need to answer the following questions first:

1. Determine the number of terms we are getting from each cube.

2. Estimate the size of approximation error.

3. Estimate how often we get too many ($\geq m_0^2$) terms.

Lemma 4.3.7. *Given* $\{\phi_l\}_{l=1}^{m_\chi} \subset C_c^\infty(\mathbb{R}^{m_0})$, $\{X_k\}_{k=1}^N$ *be i.i.d. uniform from* $\Omega \subset \mathbb{R}^{m_0}$. *Let* $\hat{W} \in \mathbb{R}^{m_0^2 \times L}$ *be the matrix whose columns are the vectorization of*

$$\hat{W}_l = \frac{1}{N} \sum_k f(X_k) \nabla^2 \phi_l(X_k),$$

and $W \in \mathbb{R}^{m_0^2 \times L}$ *be the matrix whose columns are the vectorization of*

$$W_l = \int_\Omega f(x)\nabla^2\phi_l(x)dx.$$

Then,

$$\|W - \hat{W}\|_F \leq \sqrt{m_\chi}\eta$$

with probability greater than $1 - 2m_0 m_\chi \exp(\frac{-\eta^2 N}{2(K^2+K\eta/3)})$*, where* $K = (2\|f\|_\infty + C_\mathcal{N}) \max_x \left\| \frac{\nabla^2\phi_l(x)}{p(x)} \right\|$.

Proof. First, we note that $\|W - \hat{W}\|_F = \sum_{l=1}^{m_\chi} \|vec(W_l) - vec(\hat{W}_l)\|_2^2 = \sum_{l=1}^{m_\chi} \|W_l - \hat{W}_l\|_F^2$, where

$$W_l - \hat{W}_l = \frac{1}{N}\sum_{k=1}^{N}(f(X_k) + n_k)\frac{\nabla^2\phi_l(X_k)}{p(X_k)} - \int_{\mathbb{R}^{m_0}} f(x)\nabla^2\phi_l(x)dx,$$

which is a $m_0 \times m_0$ matrix. Letting $\chi_{k,l} = (f(X_k)+n_k)\frac{\nabla^2\phi_l(X_k)}{p(X_k)} - \int_{\mathbb{R}^{m_0}} f(x)\nabla^2\phi_l(x)dx$, $\|W_l - \hat{W}_l\|_F = \|\frac{1}{N}\sum_{k=1}^{N}\chi_{k,l}\|_F$. The spectral norm can be bounded by $\|\chi_{k,l}\| \leq (2\|f\|_\infty + C_\mathcal{N}) \max_x \left\| \frac{\nabla^2\phi_l(x)}{p(x)} \right\| =: K$. Moreover, $\sum_{k=1}^{N} \mathbb{E}\|\chi_{k,l}\|^2 \leq NK^2 =: \sigma^2$. Then,

$$\mathbb{P}\left(\left\| \frac{1}{N}\sum_{k=1}^{N}\chi_{k,l} \right\| > \eta \right) < 2m_0 \exp\left(\frac{-\eta^2 N}{2(K^2 + K\eta/3)}\right).$$

Thus, $\|W - \hat{W}\|_F \leq \sqrt{m_\chi}\eta$ with probability larger than $1 - 2m_0 m_\chi \exp(\frac{-\eta^2 N}{2(K^2+K\eta/3)})$.

□

Lemma 4.3.8. *Compute the SVD of* $W = U\Sigma V^T, \hat{W} = \hat{U}\hat{\Sigma}\hat{V}^T$ *and take the first* q

columns of $U, \hat{U}$ *to create* $P_W = U|_q(U|_q)^T$, $P_{\hat{W}} = \hat{U}|_q(\hat{U}|_q)^T$. *Then,*

$$\|P_W - P_{\hat{W}}\|_F \leq \frac{2\eta}{\sqrt{\alpha(1-s)} - \eta}$$

with probability greater than $1 - 2m_0 m_\chi \exp(\frac{-\eta^2 N}{2(K^2 + K\eta/3)}) - q\exp(-\frac{m_\chi \alpha s^2}{2(\|f\|_\infty C_{\nu,2})^2})$, *where* $C_{\nu,2} = \max_{l\in\nu}\max_x \|\frac{\nabla^2\phi_l(x)}{p(x)}\|$.

Lemma 4.3.9 (Singular value gap between meaningful and meaningless terms)**.** *Suppose the test function* ϕ_l *intersects with* q *edges. Then, with probability greater than* $1 - (m_0+q)\exp(-\frac{m_\chi \alpha s^2}{2(C_0 C_\nu m_0)^2}) - 2m_0\exp(-\frac{\eta^2 N}{4[2(m_0 Q)^2 + m_0 Q\eta/3]})$, *where* $Q = (C_0 + C_{\mathcal{N}}/m_0)C_\nu$, $\sigma_{m_0+q}(W_l) \geq \sqrt{m_\chi \alpha(1-s)}$, $\sigma_{m_0+q}(\hat{W}_l) \geq \sqrt{m_\chi}(\sqrt{\alpha(1-s)} - \eta)$, *and* $\sigma_{m_0+q+1}(\hat{W}_l) \leq \sqrt{m_\chi}\eta$.

Proof. Using Weyl's estimate, we see that $\sigma_{m_0+q}(\hat{W}_l) \geq \sigma_{m_0+q}(W_l) - \|W - \hat{W}\|_F$, and $\sigma_{m_0+q+1}(\hat{W}_l) \leq \sigma_{m_0+q+1}(W_l) + \|W_l - \hat{W}_l\|_F = \|W_l - \hat{W}_l\|_F$ since W_l consists of only $m_0 + q$ symmetric rank-1 matrices.

□

Lemma 4.3.10 ("Lebesgue number" of edges)**.** *Suppose that* $\sup_\ell \|b_\ell\|_\infty \leq C_b$, $\sup_\ell d_\ell \leq C_d$ *then in a cube* $[-R, R]^{m_0}$, *the measure of the set*

$$\Gamma_{\delta,k} = \{x \in \mathbb{R}^{m_0} : \textit{a cube with center } x \textit{ and length } \delta \textit{ intersect more than } k \textit{ entangled boundaries}\}$$

satisfies $\mu(\Gamma_{\delta,k}) \leq (2RC_b)^{m_0} m_1 \left(\frac{m_0\delta}{2R}\right)^{-\log k}$.

Proof. For a fixed ℓ, we consider $b_\ell \cdot g(x) + d_\ell = \sum_t (b_\ell)_t (a_t \cdot x + c_t)_+ + d_\ell$, where $x_+ =$

$\max\{x, 0\}$. We shall make successively the following three change of coordinates to simplify the expression:

1. $y = A^t x$,

2. $\bar{y}_t = y_t + c_t$,

3. $\tilde{y}_t = |(b_\ell)_t| \bar{y}_t$.

Note that we assumed that A is orthogonal by whitening. Then, the expression becomes $\sum_t sgn((b_\ell)_t)(\tilde{y}_t)_+ + d_\ell$.

Now, the boundaries $\{a_t \cdot x + c_t = 0\}$ is transformed into $\tilde{y}_t = 0$ for all t, and $\{b_\ell \cdot g(x) + d_\ell = 0 : (g(x))_t = (a_t \cdot x + c_t)\alpha_t, \alpha \in \{0,1\}^{m_0}\}$ becomes $\{\sum_{t:\alpha_t=1}(b_\ell)_t \tilde{y}_t + d_\ell = 0 : \tilde{y}_t \geq 0 \iff \alpha_t = 1\}$. Each of the entangled boundaries is positioned in one 2^{m_0}-drant. Thus, if a cube of length δ intersects with multiple boundaries, it must be near the lower-dimensional faces where some entries are 0.

Around the face with s entries being 0 and others away from 0 by more than δ, cubes with center at those points intersect with at most 2^s entangled boundaries. The set of points near these intersections has volume not greater than $(2\delta)^s(2d_\ell)^{m_0-s}$. There are in total $\binom{m_0}{s}$ choices, so the total volume is $\binom{m_0}{s}(2\delta)^s(2d_\ell)^{m_0-s} \leq (2d_\ell)^{m_0}(\frac{m_0\delta}{d_\ell})^s$.Also, a point intersects one or more fixed boundaries if and only if some entries are small.

Let the change of coordinate be denoted as $\mathcal{A}$. Thus, given k, $\mathcal{A}\Gamma_{\delta,k}$ has volume

$$\mu(\mathcal{A}\Gamma_{\delta,k,\ell}) \leq \sum_{s \geq \log_2 k} (2d_\ell)^{m_0}(\frac{m_0\delta}{d_\ell})^s \leq \frac{1}{1-\epsilon}(2d_\ell)^{m_0}(\frac{m_0\delta}{d_\ell})^{\log_2 k},$$

where $\epsilon = m_0\delta/d_\ell$. $\mu(\mathcal{A}(\cup_\ell \Gamma_{\delta,k/m_1,\ell})) \leq \frac{m_1}{1-\epsilon}(2C_d)^{m_0}(\frac{m_0\delta}{C_d})^{\log_2 k - \log_2 m_1}$. Reverting back

to the original space, the volume of $\Gamma_{\delta,k} \subset \cup_\ell \Gamma_{\delta,k/m_1,\ell}$ is less than $\frac{m_1}{1-\epsilon}(2C_bC_d)^{m_0}(\frac{m_0\delta}{C_d})^{\log_2 k - \log_2 m_1}$.

□

4.3.3 Function Recovery

As we recover $\{\pm a_j\}, \{\pm Ab_\ell\}_\ell$, we are still far from recovering f. In particular, we need to address the following problems:

- Assigning $\{\pm a_j\}$ to the first layer.
- Determine the correct orientation of $\{a_j\}$.
- Determine the correct orientation of $\{b_\ell\}$.
- Determine $\{\alpha_\ell, d_\ell\}, \{c_j\}$.

Once we are able to carry out all actions above, we will be able to recover f fully.

4.3.3.1 Assigning $\{a_j\}$ and $\{b_\ell\}$ to their respective layers

Note that we have recovered $\{\pm a_j\} \cup \{v_\alpha^\ell\}$. Now, for any $\alpha_1, \alpha_2 \in \{0,1\}^{m_0}$, $<v_{\alpha_1}^\ell, v_{\alpha_2}^\ell> \geq \eta^2 \|b_\ell\|_2^2$. Thus, we may use this to distinguish $\{a_j\}$ from $\{b_\ell\}$.

Proposition 4.3.11. *If $S(a_1, \ldots, a_{m_0}) < \delta$, then for any $\ell \in \{1, \ldots, m_1\}$, $\alpha_1, \alpha_2 \in \{0,1\}^{m_1}$, one has*

$$|<v_{\alpha_1}^\ell, v_{\alpha_2}^\ell>| \geq \frac{\eta^2}{(1+\delta)^2} - \frac{2\delta + \delta^2}{(1-\delta)^2}.$$

Proof. Since $S(a_1, \ldots, a_{m_1}) < \delta$, there exists an orthonormal basis $\{w_1, \ldots, w_{m_0}\}$ such that $(\sum \|a_i - w_i\|^2)^{1/2} < \delta$. Let $A = (a_1|\ldots|a_{m_0}), W = (w_1|\ldots|w_{m_0})$, then

$\|A - W\|_F < \delta$ by construction. Then, for any $b \in \mathbb{R}^{m_0}$, $|\|Ab\|_2 - \|b\|| = |\|Ab\|_2 - \|Wb\|_2| \leq \|(A-W)b\|_2 < \delta\|b\|_2$.

Now, for any ℓ, α_1, α_2, $v^\ell_{\alpha_j} = \frac{1}{\|Ab^{\alpha_j}_\ell\|} Ab^{\alpha_j}_\ell$, where $(b^{\alpha_j}_\ell)_t = (b_\ell)_t(\eta + (1-\eta)(\alpha_j)_t)$, where $j = 1, 2$. Since $v^\ell_{\alpha_j}$ is of unit-norm, we have that $\|\frac{1}{\|Ab^{\alpha_j}_\ell\|} b^{\alpha_j}_\ell\|_2 \in [\frac{1}{1+\delta}, \frac{1}{1-\delta}]$. Moreover, as $\|b^{\alpha_j}_\ell\|_2 \leq \|b_\ell\|_2 = 1$, we have that $\|Ab^{\alpha_j}_\ell\|_2 \leq 1 + \delta$. Then,

$$
\begin{aligned}
|<v^\ell_{\alpha_1}, v^\ell_{\alpha_2}>| &= |(\frac{1}{\|Ab^{\alpha_2}_\ell\|} b^{\alpha_2}_\ell)^T A^T A (\frac{1}{\|Ab^{\alpha_1}_\ell\|_2} b^{\alpha_1}_\ell)| \\
&\geq |\frac{1}{\|Ab^{\alpha_1}_\ell\| \|Ab^{\alpha_2}_\ell\|} (b^{\alpha_2}_\ell)^T W^T W b^{\alpha_1}_\ell| \\
&\quad - |(\frac{1}{\|Ab^{\alpha_2}_\ell\|} b^{\alpha_2}_\ell)^T \left[(A-W)^T W + W^T(A-W) + (A-W)^T(A-W)\right] (\frac{1}{\|Ab^{\alpha_1}_\ell\|_2} b^{\alpha_1}_\ell)| \\
&\geq \frac{1}{(1+\delta)^2} |<b^{\alpha_1}_\ell, b^{\alpha_2}_\ell>| \\
&\quad - \frac{1}{(1-\delta)^2} \left(\|A-W)^T W\|_\infty - \|W^T(A-W)\|_\infty - \|(A-W)^T(A-W)\|_\infty\right) \\
&\geq \frac{\eta^2}{(1+\delta)^2} - \frac{1}{(1-\delta)^2}(2\delta + \delta^2),
\end{aligned}
$$

where we note that the operator norm is dominated by the Frobenius norm and that $<b^{\alpha_1}_\ell, b^{\alpha_2}_\ell> \leq \sum_t (b_\ell)^2_t \eta^2 = \eta^2$. □

4.3.3.2 Orienting $\{a_j\}$ and obtaining bias $\{c_j\}$

We orient $\{a_j\}$ by looking at the first weak derivative of f. Let $\{\hat{c}_j\}$ be the dual basis of an arbitrarily oriented version of $\{a_j\}$, i.e. , $\hat{c}_i \cdot a_j = (-1)^{n_i}\delta_{i,j}$ where $n_i \in \{0, 1\}$. Then, we plot out $f_j(t) = f(t\hat{c}_j)$ for $t \in \mathbb{R}$. Then, f_j is a piece-wise linear function with a finite amount of break-points. We examine the slope change around

each break-points. Suppose that around break-point t_0, we have $f'(t_0^+) = \frac{1}{c}f'(t_0^-)$, then we know that $n_j = 0$, and $c_j = -t_0$. If $f'(t_0^-) = \frac{1}{c}f'(t_0^+)$, then $n_j = 1$ and $c_j = t_0$.

4.3.3.3 Determining the orientation of $\{b_\ell\}$ and $\{\alpha_\ell\}$

Once we have determined the orientation of $\{a_j\}_j$, we are now able to assign $\{b_\ell\}$ to the second layer with possibly wrong orientations. Then, as $\{b_\ell\}$ are linearly independent, we may construct $\{\beta_\ell\}$ such that $b_\ell \cdot \beta_s = (-1)^{n_\ell}\delta_{\ell,s}$, where $n_\ell \in \{0, 1\}$.

Now, since $\{a_j\}_j$ form a basis in $\mathbb{R}^{m_0}$, there exists $\{x_\ell\}$ such that $a_j \cdot x_\ell = \gamma_j^\ell(\beta_\ell)_j$ for all j, where $\gamma_j^\ell = 1$ if $(\beta_\ell)_j \geq 0$, and $1/c$ if $(\beta_\ell)_j < 0$. Then, if we calculate the derivative of $f_\ell(t) = f(tx_\ell)$, then either $f'_\ell(t) = \alpha_\ell$ or $f'_\ell(-t) = \alpha_\ell$ for large t. the value with larger magnitude is α_ℓ, while we re-orientate our b_ℓ to $-b_\ell$ if $f'_\ell(-t) = \alpha_\ell$.

4.3.3.4 Determining $\{d_\ell\}$

Basically we follow the same process as the previous section while looking closely at the break-points of f^ℓ. It has only one break-point, and judging from the orientation of $\{b_\ell\}$, we can recover $\{d_\ell\}$ perfectly.

4.3.4 Future Works

For the two-layer case, the biggest remaining difficulty is the theoretical guarantee for net-spreading, as it is not yet clear how to determine the threshold for

cut-offs in the algorithm. The first step is to identify the cases when no components are activated in a region. This will help us reduce the chance of false positives. The second one is to estimate the gap between the smallest activated singular value and the largest inactive one.

Bibliography

[1] Akram Aldroubi, Jacqueline Davis, and Ilya Krishtal. Exact reconstruction of spatially undersampled signals in evolutionary systems. *arXiv preprint arXiv:1312.3203*, 2013.

[2] Akram Aldroubi, Jacqueline Davis, and Ilya Krishtal. Exact reconstruction of signals in evolutionary systems via spatiotemporal trade-off. *Journal of Fourier Analysis and Applications*, 21(1):11–31, 2015.

[3] Luigi Ambrosio and Gianni Dal Maso. A general chain rule for distributional derivatives. *Proceedings of the American Mathematical Society*, 108(3):691–702, 1990.

[4] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014.

[5] Afonso S Bandeira, Matthew Fickus, Dustin G Mixon, and Joel Moreira. Derandomizing restricted isometries via the legendre symbol. *Constructive Approximation*, 43(3):409–424, 2016.

[6] John J Benedetto, Alexander M Powell, and Ozgur Yilmaz. Sigma-delta quantization and finite frames. *IEEE Transactions on Information Theory*, 52(5): 1990–2005, 2006.

[7] John J Benedetto, Robert L Benedetto, and Joseph T Woodworth. Optimal ambiguity functions and weil's exponential sum bound. *Journal of Fourier Analysis and Applications*, 18(3):471–487, 2012.

[8] James Blum, Mark Lammers, Alexander M Powell, and Özgür Yılmaz. Sobolev duals in frame theory and sigma-delta quantization. *Journal of Fourier Analysis and Applications*, 16(3):365–381, 2010.

[9] Jean Bourgain, Stephen Dilworth, Kevin Ford, Sergei Konyagin, Denka Kutzarova, et al. Explicit constructions of rip matrices and related problems. *Duke Mathematical Journal*, 159(1):145–185, 2011.

[10] Martin D Buhmann and Allan Pinkus. Identifying linear combinations of ridge functions. *Advances in Applied Mathematics*, 22(1):103–118, 1999.

[11] Emmanuel J Candès. Ridgelets: estimating with ridge functions. *The Annals of Statistics*, 31(5):1561–1599, 2003.

[12] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[13] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

[14] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[15] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[16] James Candy. *Decimation for sigma delta modulation*, volume 34. IEEE transactions on communications, 1986.

[17] Evan Chou and C. Sinan Güntürk. Distributed noise-shaping quantization: Ii. classical frames. *Excursions in Harmonic Analysis, Volume 5: The February Fourier Talks at the Norbert Wiener Center*, (179-198), 2017.

[18] Evan Chou and C.Sinan Güntürk. Distributed noise-shaping quantization: I. beta duals of finite frames and near-optimal quantization of random measurements. *Constructive Approximation*, 44(1):1–22, 2016.

[19] Evan Chou, C. Sinan Güntürk, Felix Krahmer, Rayan Saab, and Özgür Yılmaz. *Noise-shaping quantization methods for frame-based and compressive sampling systems*. Number 157–184. Springer, 2015.

[20] Wu Chou, Ping Wah Wong, and Robert M Gray. Multistage sigma-delta modulation. *IEEE Transactions on Information theory*, 35(4):784–796, 1989.

[21] Fan RK Chung. Several generalizations of weil sums. *Journal of Number Theory*, 49(1):95–106, 1994.

[22] Albert Cohen, Ingrid Daubechies, Ronald DeVore, Gerard Kerkyacharian, and Dominique Picard. Capturing ridge functions in high dimensions from point queries. *Constructive Approximation*, 35(2):225–243, 2012.

[23] Paul G Constantine. *Active subspaces: Emerging ideas for dimension reduction in parameter studies*, volume 2. SIAM, 2015.

[24] Ingrid Daubechies and Ron DeVore. Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order. *Annals of mathematics*, 158(2):679–710, 2003.

[25] Ingrid Daubechies and Rayan Saab. A deterministic analysis of decimation for sigma-delta quantization of bandlimited functions. *IEEE Signal Processing Letters*, 22(11):2093–2096, 2015.

[26] Ingrid Daubechies, Ronald A DeVore, C Sinan Gunturk, and Vinay A Vaishampayan. A/d conversion with imperfect quantizers. *IEEE Transactions on Information Theory*, 52(3):874–885, 2006.

[27] Percy Deift, Felix Krahmer, and C Sınan Güntürk. An optimal family of exponentially accurate one-bit sigma-delta quantization schemes. *Communications on Pure and Applied Mathematics*, 64:883–919, 2011.

[28] Ronald A DeVore, Konstantin I Oskolkov, and Pencho P Petrushev. Approximation by feed-forward neural networks. *Annals of Numerical Mathematics*, 4: 261–288, 1996.

[29] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[30] Yonina C Eldar and Helmut Bolcskei. Geometrically uniform frames. *IEEE Transactions on Information Theory*, 49(4):993–1006, 2003.

[31] PF Ferguson, A Ganesan, and RW Adams. One bit higher order sigma-delta a/d converters. *IEEE International Symposium on Circuits and Systems*, pages 890–893, 1990.

[32] Massimo Fornasier, Karin Schnass, and Jan Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12(2):229–262, 2012.

[33] Massimo Fornasier, Jan Vybíral, and Ingrid Daubechies. Robust and resource efficient identification of shallow neural networks by fewest samples. *arXiv preprint arXiv:1804.01592*, 2018.

[34] Massimo Fornasier, Timo Klock, and Michael Rauchensteiner. Robust and resource efficient identification of two hidden layer neural networks. *arXiv preprint arXiv:1907.00485*, 2019.

[35] G David Forney. Geometrically uniform codes. *IEEE Transactions on Information Theory*, 37(5):1241–1260, 1991.

[36] John Friedlander and Henryk Iwaniec. Estimates for character sums. *Proceedings of the American Mathematical Society*, 119(2):365–372, 1993.

[37] Vivek K Goyal, Jelena Kovacevic, and Martin Vetterli. Quantized frame expansions as source-channel codes for erasure channels. *Proceedings DCC'99 Data Compression Conference (Cat. No. PR00096)*, pages 326–335, 1999.

[38] Vivek K Goyal, Jelena Kovačević, and Jonathan A Kelner. Quantized frame expansions with erasures. *Applied and Computational Harmonic Analysis*, 10 (3):203–233, 2001.

[39] C Sinan Güntürk. One-bit sigma-delta quantization with exponential accuracy. *Communications on Pure and Applied Mathematics*, 56(11):1608–1630, 2003.

[40] Thang Huynh and Rayan Saab. Fast binary embeddings, and quantized compressed sensing with structured matrices. *arXiv preprint arXiv:1801.08639*, 2018.

[41] Hiroshi Inose and Yasuhiko Yasuda. A unity bit coding method by negative feedback. *Proceedings of the IEEE*, 51:1524–1535, 1963.

[42] Mark Iwen and Rayan Saab. Near-optimal encoding for sigma-delta quantization of finite frame expansions. *Journal of Fourier Analysis and Applications*, 19(6):1255–1273, 2013.

[43] Ernst Jacobsthal. Über die darstellung der primzahlen der form 4n+ 1 als summe zweier quadrate. *Journal für die reine und angewandte Mathematik*, 132(238–245), 1907.

[44] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[45] Na Lei, Kehua Su, Li Cui, Shing-Tung Yau, and David Xianfeng Gu. A geometric view of optimal transportation and generative model. *Computer Aided Geometric Design*, 68:1–21, 2019.

[46] Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

[47] Kung-Ching Lin. Analysis of decimation on finite frames with sigma-delta quantization. *Constructive Approximation*, 50(3):507–542, 2019.

[48] Kung-Ching Lin. Three dimensional sums of character gabor systems. *arxiv preprint arXiv:1909.11561*, 2019. Ready for submission.

[49] Yue M Lu and Martin Vetterli. Distributed spatio-temporal sampling of diffusion fields from sparse instantaneous sources. *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2009 3rd IEEE International Workshop on*, pages 205–208, 2009.

[50] Yue M Lu and Martin Vetterli. Spatial super-resolution of a diffusion field by temporal oversampling in sensor networks. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processin*, (LCAV-CONF-2009-009): 2249–2252, 2009.

[51] Dustin G Mixon. Explicit matrices with the restricted isometry property: Breaking the square-root bottleneck. *Compressed sensing and its applications*, pages 389–417, 2015.

[52] Wolfgang M Schmidt. *Equations over finite fields: an elementary approach*, volume 536. Springer, 2006.

[53] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *arXiv preprint arXiv:1906.05497*, 2019.

[54] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. *arXiv preprint arXiv:1902.10170*, 2019.

[55] S Tewksbury and RW Hallock. Oversampled, linear predictive and noise-shaping coders of order n¿ 1. *IEEE Transactions on Circuits and Systems*, 25(7):436–447, 1978.

[56] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[57] Per-Ake Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[58] André Weil. On some exponential sums. *Proceedings of the National Academy of Sciences*, 34(5):204–207, 1948.