ABSTRACT

Title of Dissertation: U(R) PHASE RETRIEVAL,
LOCAL NORMALIZING FLOWs,
AND HIGHER ORDER FOURIER TRANSFORMS

Christopher B. Dock
Doctor of Philosophy, 2022

Dissertation Directed by: Professor Radu Balan
Department of Mathematics

The classical phase retrieval problem arises in contexts ranging from speech recognition to x-ray crystallography and quantum state tomography. The generalization to matrix frames is natural in the sense that it corresponds to quantum tomography of impure states. Chapter 1 provides computable global stability bounds for the quasi-linear analysis map β and a path forward for understanding related problems in terms of the differential geometry of key spaces. In particular, Chapter 1 manifests a Whitney stratification of the positive semidefinite matrices of low rank which allows us to "stratify" the computation of the global stability bound. We show that for the impure state case no such global stability bounds can be obtained for the non-linear analysis map α with respect to certain natural distance metrics. Finally, our computation of the global lower Lipschitz constant for the β analysis map provides novel conditions for a frame to be generalized phase retrievable.

In Chapter 2 we develop the concept of local normalizing flows. Normalizing flows provide

an elegant approach to generative modeling that allows for efficient sampling and exact density evaluation of unknown data distributions. However, current techniques have significant limitations in their expressivity when the data distribution is supported on a low-dimensional manifold or has a non-trivial topology. We introduce a novel statistical framework for learning a mixture of local normalizing flows as "chart maps" over the data manifold. Our framework augments the expressivity of recent approaches while preserving the signature property of normalizing flows, that they admit exact density evaluation. We learn a suitable atlas of charts for the data manifold via a vector quantized auto-encoder (VQ-AE) and the distributions over them using a conditional flow. We validate experimentally that our probabilistic framework enables existing approaches to better model data distributions over complex manifolds.

In Chapter 3 we examine higher order Fourier transforms in both discrete and continuous contexts. We demonstrate a connection to a matrix time variant of the free Schrödinger equation, as well as a potential application to magnetic resonance imaging. In the discrete case we show that the reconstruction properties of higher order Fourier frames are intricately related to quadratic Gauss sums.

U(R) PHASE RETRIEVAL, LOCAL NORMALIZING FLOWS, AND HIGHER ORDER FOURIER TRANSFORMS

by

Christopher B. Dock

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
2022

Advisory Committee:
Professor Radu Balan, Chair/Advisor
Professor Vince Lyzinski Co-Chair
Professor John Benedetto
Professor Behtash Bebadi
Professor Maria Cameron
Professor Doron Levy

Foreword

The structure of this thesis is divided into three chapters examining distinct mathematical problems: the $U(r)$ phase retrieval problem, the problem of normalizing flow expressivity, and the problem of higher order (and redundant information) Fourier transforms. While these topics are distinct, they possess some striking mathematical connections. For instance Chapters 1 and 3 both examine what one might call "matrix frame theory," albeit from different perspectives: In Chapter 1 we ask the question of which generalized frames for $\mathbb{C}^{n\times r}$ of the form $(A_j)_{j=1}^m \subset \text{Sym}(\mathbb{C}^n)$ are sufficient to recover an arbitrary matrix $z \in \mathbb{C}^{n\times r}$ from measurements of the form $(\langle zz^*, A_j\rangle)_{j=1}^m$, up to its orbit under right multiplication by $U(r)$. In doing so, we also give a measure, in terms of the lower Lipschitz constant of a particular analysis map, of how good one can expect said recovery to be in the presence of noise for a given choice of generalized frame. In Chapter 3, meanwhile, we ask the question of what additional information is gained by extending the discrete Fourier basis $(e_k)_{k=0}^{d-1}$ with $(e_k)_j := e^{2\pi ijk/d}$ for $\mathbb{C}^d$ to a quadratic Fourier frame of the form $(q_{k,l})_{k,l=1}^d$ with $(q_{k,l})_j := e^{2\pi i(lj^2+kj)/d}$ (noting that the d^2 quadratic Fourier coefficients of $v \in \mathbb{C}^d$ comprise the matrix $[\langle q_{k,l}, v\rangle]_{k,l=1}^d \in \mathbb{C}^{d\times d}$). In this context we show that there exist sub-sampling schemes in which the loss of linear frequency information can be compensated for by information from quadratic frequencies, and as in Chapter 1 provide an estimate of the reconstruction error in the presence of noise.

The mathematical thread connecting Chapters 1 and 2 is not frame theory but differential geometry. In Chapter 1, computation of the lower Lipschitz bounds of the analysis maps α and

β necessitates a foray into the theory of Whitney stratification of semi-algebraic varieties (sets defined by finite Boolean combinations of polynomial inequalities). This allows computation of the relevant lower Lipschitz constants to be "stratified" over sets that are manifolds rather than semi-algebraic varieties, which in turn allows the problem to be fully linearized. In somewhat of a happy accident, we were also able to show that the family of Riemannian metrics giving rise to one of the distance metrics of interest was "compatible" across the stratification of $\mathbb{C}^{n\times r}/U(r)$ – in a sense defining a Riemannian geometry on the entire semi-algebraic variety. Meanwhile in Chapter 2 we lean heavily on the basic machinery of differential geometry, employing normalizing flows not as global diffeomorphisms but instead as chart maps in a suitably chosen atlas. Moreover, we similarly localize the technique used in [1] of post-composing normalizing flows with conformal transformations in order to handle low dimensional manifolds. We show that doing so is natural by appealing to the theory of locally conformally flat manifolds.

Finally one would be remiss not to note that both frame theory and more generally harmonic analysis (Chapters 1 and 3) and generative machine learning (Chapter 2) are different philosophical approaches to the same problem, namely to provide representations of functions that are both sufficiently expressive (can accurately represent a rich class of functions) and have nice properties (parameter efficiency, robustness to noise or partial loss of the representation, etc).

This thesis includes ongoing work and work already submitted for publication. In particular:

1. Chapter 1 was a project with Radu V. Balan. An abridged version of Chapter 1 was submitted for publication to the SIAM Journal of Matrix Analysis and has passed the first round of revisions. The full paper can be found on arχiv at https://arxiv.org/abs/2109.14522v2.

2. Some of the results in Chapter 1 related to Lipschitz analysis were presented at the Approximation Theory 16 conference at Vanderbilt University in May 2019. The full presentation can be found at https://cbartondock.github.io/plain-academic/slides/AT16.pdf.

3. Results from Chapter 1 related to differential geometry of $S^{r,0}(\mathbb{C}^n)$ and criteria for matrix frames to be $U(r)$ phase retrievable were presented at the AMS Fall Western Virtual Sectional Meeting Special Session on Harmonic Analysis: Geometry, Frames, and Sampling in October 2021. The full presentation can be found at https://cbartondock.github.io/plain-academic/slides/AMSFall2021.pdf.

4. Chapter 2 was a project with Radu V. Balan, Sahil Sidheekh, Tushar Jain, and Maneesh Singh. Chapter 2 was accepted as a conference paper for the 2022 conference Uncertainty in Artificial Intelligence (UAI) that will be in Eindhoven in August 2022. Chapter 2 is also available at https://arxiv.org/abs/2203.11556.

5. Chapter 3 is an ongoing project with Radu Balan and Yonina C. Eldar.

Dedication

For my parents, Alan W. Dock and Mary A. Barton-Dock, for their constant love and support.

Table of Contents

Bibliography 187

List of Tables

List of Figures

Chapter 1: Generalized Phase Retrieval[1]

1.1 Introduction

Problems of "phase loss" type – in which a signal must be reconstructed from only the magnitude of its Fourier transform or Fourier coefficients – are ubiquitous in applications, appearing for example in inverse scattering problems, thin film optics, x-ray crystallography, electron microscopy, astronomy, speech processing, and pure state quantum tomography [5]. The intuitive reasons for this ubiquity are essentially two-fold, the first of which will be familiar to anyone who has studied electromagnetism and optics. Loosely speaking, if a field is described by a linear partial differential equation that admits travelling waves as solutions (for example Maxwell's equations and subsequently the Helmholtz equation), then data about "near field interactions" are encoded in the Fourier transform of the "far field," that is to say the state of the field sufficiently far from the interaction relative to some intrinsic scale of interaction [5]. This principle is perhaps most directly apparent in the Fraunhofer diffraction formula, which describes the diffraction pattern produced by an aperture $\mathcal{A}$ (a compact, hence measurable subset of $\mathbb{R}^2$) when both the source of the incident wave and the measurement apparatus are sufficiently far from the aperture relative to its size. Specifically, the Fraunhofer diffraction formula gives the value of a component of the

[1] In collaboration with Radu V. Balan. This work was submitted for publication in somewhat shortened form to the SIAM Journal of Matrix Analysis.

electromagnetic field U at point x as:

$$U(x) \propto \int_{\mathcal{A}} e^{2\pi i x \cdot y} dy = \mathcal{F}[\mathbb{1}_{\mathcal{A}}](x) \tag{1.1.1}$$

This formula and analogous results for lenses and other near field interactions allow one to predict and in some cases to analytically compute the diffraction pattern produced, but in experimental physics one is often tasked with the opposite problem: to analyze the near field interaction using measurements of the diffraction pattern (or scattering cross section) it produces. The second reason the phase retrieval problem appears in optics and in inverse scattering problems is thus practical: It is usually only possible to measure the magnitude of an oscillating field, not its phase, at different points in space. Indeed, optical instruments typically measure photon flux which is proportional to the squared magnitude of the electromagnetic field [5]. Similarly if one measures the cross sectional density produced by scattering quantum particles off of an interaction potential, then one can infer only the absolute square of the quantum wave function. Thus if one seeks information about the near field interaction (for example the dielectric properties of a lens or the approximate scattering potential) one must attempt to reconstruct the function encoding the near field interaction from the absolute value of its Fourier transform. In practice, one is of course further restricted to making finitely many measurements of the field, which leads directly to the theory of discrete phase retrieval.

Further interest in the phase retrieval problem arose in the context of speech recognition and speech processing. It is known that the human ear is quite reliably "phase deaf," and as such one should not expect the linguistic meaning of an audio waveform containing human language

to depend on its phase. This intuition is made quantitative in the "cepstral analysis" of speech signals in which phase retrieval plays a key role [6].

1.2 Variants of the phase retrieval problem

1.2.1 The continuous phase retrieval problem

The most natural general setting for the continuous phase retrieval problem arrives vis a vis the theory of tempered distributions. Because it is most relevant to the phase retrieval problem we restrict ourselves to a recapitulation of the one dimensional theory of distributions, but one may extend it to tempered distributions on $\mathbb{R}^n$ without encountering serious theoretical difficulties. In order to include all the usual variants of the phase retrieval problem we will also allow complex valued Schwartz functions and tempered distributions, in contrast with the usual presentation of distribution theory. In particular, let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. Then $f \in C^\infty(\mathbb{R} \to \mathbb{K})$ is termed *rapidly decreasing* if for all $N \in \mathbb{N}$

$$\lim_{|x|\to\infty} |x|^N |f(x)| = 0 \tag{1.2.1}$$

In this case the *Schwartz class* $\mathcal{S}_{\mathbb{K}}(\mathbb{R})$ is given by

$$\mathcal{S}_{\mathbb{K}}(\mathbb{R}) = \{f \in C^\infty(\mathbb{R} \to \mathbb{K}) \mid \partial_x^k f \text{ is rapidly decreasing for all } k = 0, 1, \ldots\} \tag{1.2.2}$$

The Schwartz class is equipped with the family of norms $||\cdot||_{\alpha,\beta}$ defined to be $||f||_{\alpha,\beta} = \sup_{x\in\mathbb{R}} |x|^\alpha |\partial_x^\beta f(x)|$ for $f \in \mathcal{S}_{\mathbb{K}}(\mathbb{R})$ and $\alpha, \beta \in \mathbb{N}$ (the natural generalization of $||f||_{\alpha,\beta}$ to $\mathbb{R}^n$

yields only a family of semi-norms, but they nevertheless define a locally convex topology on $\mathcal{S}_{\mathbb{K}}(\mathbb{R})$ and can be upgraded to a family of norms as needed). As such, one defines the tempered distributions $\mathcal{S}'_{\mathbb{K}}(\mathbb{R})$ as the continuous dual of $\mathcal{S}_{\mathbb{K}}(\mathbb{R})$, that is to say a $\mathbb{K}$-linear functional $\phi : \mathcal{S}_{\mathbb{K}}(\mathbb{R}) \to \mathbb{K}$ is an element of $\mathcal{S}'_{\mathbb{K}}(\mathbb{R})$ if and only if for any sequence $(f_m)_{m\geqslant 0} \subset \mathcal{S}_{\mathbb{K}}(\mathbb{R})$ such that $\lim_{m\to\infty} ||f_m||_{\alpha,\beta} \to 0$ for all $\alpha, \beta \in \mathbb{N}$ we have $\lim_{m\to\infty} |\phi(f_m)| \to 0$. Note that one may restrict from $\mathcal{S}'_{\mathbb{C}}(\mathbb{R})$ to $\mathcal{S}'_{\mathbb{R}}(\mathbb{R})$ in the following sense: A distribution $\phi \in \mathcal{S}'_{\mathbb{C}}(\mathbb{R})$ is considered to be real valued if for every real valued test function $f \in \mathcal{S}_{\mathbb{R}}(\mathbb{R})$ one has $\phi(f) \in \mathbb{R}$. Finally we need the definition of the Fourier transform for a tempered distribution: $\mathcal{F} : \mathcal{S}'_{\mathbb{K}}(\mathbb{R}) \to \mathcal{S}'(\mathbb{R})$ is defined via the Fourier transform on $\mathcal{S}_{\mathbb{K}}(\mathbb{R})$ and the Parseval identity as:

$$\mathcal{F}[\psi](\phi) := \psi(\mathcal{F}[\phi]) \tag{1.2.3}$$

Where convenient we will also write $\hat{\psi}$ for $\mathcal{F}[\psi]$. If $\psi \in L^1(\mathbb{R})$ (permitting a slight abuse of notation in identifying $\psi \in L^1(\mathbb{R})$ with the tempered distribution $\phi \mapsto \int_{\mathbb{R}} \overline{\psi}\phi dx$) then of course $\hat{\psi}(\omega) = \int_{\mathbb{R}} e^{-2\pi i\omega x}\psi(x)dx$ is the usual Fourier transform. The phase retrieval problem can then be stated as follows: Recover

$$\psi \in \mathcal{B} \subset \{\phi \in \mathcal{S}'_{\mathbb{K}}(\mathbb{R}) | \hat{\phi} \in L^1_{\text{loc}}(\mathbb{R} \to \mathbb{K})\}/\sim \tag{1.2.4}$$

from $|\mathcal{F}[\psi]|$ where $\psi_1 \sim \psi_2$ if and only if there exists a unimodular scalar $\lambda \in \mathbb{K}$ such that $\psi_1 = \lambda\psi_2$. If $\mathbb{K} = \mathbb{R}$ then $\lambda \in \{1, -1\}$, whereas if $\mathbb{K} = \mathbb{C}$ then $\lambda \in U(1) \simeq \{e^{i\theta} \mid \theta \in [0, 2\pi)\}$. The quotienting out of the overall phase factor λ is necessary because the Fourier transform is $\mathbb{K}$-linear, and thus any constant phase factor would not appear in $|\hat{\phi}|$). The technical requirement

that $\hat{\phi} \in L^1_{\text{loc}}(\mathbb{R} \to \mathbb{K})$ in combination with the Fourier inversion theorem has the effect of fully determining ϕ once $\arg(\hat{\phi})$ is known almost everywhere, thus the problem becomes to determine $\arg(\hat{\phi})$ from $|\hat{\phi}|$ [7]. It should be noted that the continuous phase retrieval problem is usually treated in the complex case, however we include the real case above because its discretization is of interest.

With the problem thus stated, it is clear that the recovery of $\psi \in \mathcal{B}$ is only possible if the collection of functions $\mathcal{B}$ is sufficiently restrictive, since for arbitrary such ψ the phase and magnitude of the Fourier transform are independent. Indeed, if $r \in L^2(\mathbb{R})$ and $\psi = \mathcal{F}^{-1}[r(\omega)e^{-2\pi i p(\omega)}]$ for any $p \in C^\infty(\mathbb{R})$ then $|\hat{\phi}| = |r|$. An example of $\mathcal{B}$ for which recovery is possible is to further restrict to functions having having compact support or supported on the half line $\mathbb{R}_{\geqslant 0}$ [7]. This particular setup is highly relevant to optics problems, in which $\mathcal{B}$ is typically taken to be a subset of functions having compact support (representing the interaction region of the lens or other impediment to the travelling wave solution).

1.2.2 The discrete phase retrieval problem for Fourier measurements

Obtainable from the continuous phase retrieval problem but of distinct theoretical and practical import is the discrete phase retrieval problem for Fourier measurements. In particular if we take

$$\mathcal{B} = \{\sum_{i \in I} z_i \delta(x - x_i) : z \in l^2(I, \mathbb{K})\} \tag{1.2.5}$$

where δ is the Dirac distribution, and I is a countable index set, then for $\phi \in \mathcal{B}$ one has

$$|\hat{\phi}(\omega)| = |\sum_{i \in I} e^{-2\pi i \omega x_i} z_i| \tag{1.2.6}$$

If for example $I = \{0, \ldots, N-1\}$, $x_i = i/N$, and $|\hat{\phi}|$ is only measured for $\omega \in \{0, \ldots, N-1\}$ then the problem becomes to reconstruct $z \in \mathbb{K}^N/\sim$ from measurements of the form

$$|\hat{\phi}(k)| = |\sum_{n=0}^{N-1} z_n e^{-2\pi i n k/N}| = |Z_k| \tag{1.2.7}$$

where $Z \in \mathbb{C}^N$ is the discrete Fourier transform of z. Let $(e_j)_{j=1}^N \in \mathbb{C}^N$ with $(e_j)_k = e^{2\pi i j k/N}$ be the discrete Fourier basis for $\mathbb{C}^N$. Then the discrete phase retrieval problem may be formulated as finding an inverse to the following function:

$$\begin{aligned} \alpha &: \mathbb{K}^N/\sim \;\to\; \mathbb{R}^N \\ \alpha(z) &= \begin{bmatrix} |\langle e_1, z\rangle_{\mathbb{C}}| \\ \vdots \\ |\langle e_N, z\rangle_{\mathbb{C}}| \end{bmatrix} \end{aligned} \tag{1.2.8}$$

Because the map α is not everywhere differentiable, it is often useful to consider instead its entry-wise square:

$$\beta : \mathbb{K}^N / \sim \to \mathbb{R}^N \tag{1.2.9}$$

$$\beta(z) = \begin{bmatrix} |\langle e_1, z\rangle_{\mathbb{C}}|^2 \\ \vdots \\ |\langle e_N, z\rangle_{\mathbb{C}}|^2 \end{bmatrix} \tag{1.2.10}$$

One can also consider the problem for I countably infinite, but in many applied contexts the discrete case suffices since one has access to only finitely many measurements and can reasonably assume that $\hat{\phi}$ is composed of only finitely many frequencies. Moreover, it is known that if I is countably infinite then the phase retrieval analysis map α is never lower Lipschitz with respect to the natural distance (even allowing for non-Fourier frames) [8].

1.2.3 The discrete phase retrieval problem for arbitrary measurements

The formulation of the discrete Fourier phase retrieval problem in (1.2.8) generalizes readily to non-Fourier measurements, and in particular to frames. Recall that if H is a separable Hilbert space then a countable subset $\{f_i\}_{i\in I} \subset H$ is a frame for H if there exist $A, B > 0$ such that for every $w \in H$

$$A||w||_H^2 \leqslant \sum_{i\in I} |\langle w, f_i\rangle_H|^2 \leqslant B||w||_H^2 \tag{1.2.11}$$

For a finite dimensional Hilbert space the notion of a frame is identical to that of a spanning set. A frame for $\mathbb{K}^n$ given by $\{f_i\}_{i=1}^m \subset \mathbb{K}^n$ is called a phase retrievable frame if

$$\begin{aligned} \alpha : \mathbb{K}^n/\sim &\to \mathbb{R}^m \\ \alpha(z) &= \begin{bmatrix} |\langle f_1, z\rangle_{\mathbb{K}}| \\ \vdots \\ |\langle f_m, z\rangle_{\mathbb{K}}| \end{bmatrix} \end{aligned} \tag{1.2.12}$$

is injective (or equivalently if β is injective). Note that if $\mathbb{K} = \mathbb{R}$ one typically restricts to real *measurement vectors* as well, and as such considers measurements of the form $\alpha_k(z) = |\langle f_k, z\rangle_{\mathbb{R}}|$ rather than $|\langle \Re[f_k], z\rangle_{\mathbb{R}} - i\langle \Im[f_k], z\rangle_{\mathbb{R}}|$.

1.2.4 The group theoretic phase retrieval problem

In this chapter we will primarily analyze a further generalization of the discrete phase retrieval problem to non-abelian phases belonging to $U(r)$. It is worth noting, however, that this problem belongs to a large class of interesting group-theoretic phase retrieval problems. Motivated by the fact that for $\mathbb{K} = \mathbb{R}$ we have that $\beta_k(z) = |\langle f_k, z\rangle|^2$ may also be written as $\beta_k(z) = \langle f_k f_k^T, zz^T\rangle_{\mathbb{R}}$ and that analogously for $\mathbb{K} = \mathbb{C}$ we have $\beta_k(z) = \langle f_k f_k^*, zz^*\rangle_{\mathbb{C}}$, we may generalize from $f_k f_k^T$ and $f_k f_k^*$ to arbitrary elements of $\mathrm{Sym}(\mathbb{R}^n)$ (resp. $\mathrm{Sym}(\mathbb{C}^n)$) and re-interpret β as the composition of the resulting linear measurements with a non-linear embedding $\pi(z) = zz^T$ (resp. $\pi(z) = zz^*$) into the space of symmetric operators that encodes the phase loss. By explicitly choosing the embedding to have a particular group of invariances, we can significantly extend the notion of "phase ambiguity" to any unitary group action on a Hilbert

space.

For now we content ourselves to the finite dimensional case: Fix finite dimensional Hilbert spaces H (real or complex) and K (real), a group G, and a unitary representation of G on H – that is to say a linear group action $\psi : G \times H \to H$ that preserves $||\cdot||_H$. Denote by $\sim$ the equivalence relation on H such that $x \sim y$ if and only if there exists $g \in G$ so that $\psi(g, x) = y$. Fix further an embedding $\pi : H \to K$ such that for $x, y \in H$ we have $\pi(x) = \pi(y)$ if and only if $x \sim y$. Then a finite collection $\mathcal{A} = \{A_j\}_{j=1}^m \subset K$ is called (G, π) phase retrievable (we will simply say G phase retrievable when the embedding in question is clear) if the following map is injective:

$$\begin{aligned} &\beta : H/G \to \mathbb{R}^m \\ &\beta_j(z) = \langle A_j, \pi(z) \rangle_K \end{aligned} \tag{1.2.13}$$

Note that if π is surjective (or in fact if $K = \Delta_\pi := \text{Ran}(\pi) - \text{Ran}(\pi) = \{\pi(x) - \pi(y) \mid x, y \in H\}$) then the notion of a G phase retrievable collection corresponds with the notion of a frame for K (and in general any frame for K is automatically G phase retrievable). For this reason, the more interesting case is when Δ_π is a proper subset of K and the collection $\mathcal{A}$ is not a frame for K. As we'll see, in many cases it is not necessary for $\mathcal{A}$ to be a frame for K in order to be phase retrievable – the problem would hardly be very interesting if it were. Indeed, in the $U(r)$ phase retrieval problem (in which $H = \mathbb{C}^{n \times r}$, $K = \text{Sym}(\mathbb{C}^n)$, $G = U(r)$, $\psi(U, z) = zU$ and $\pi(z) = zz^*$) it can be shown that when $r \leqslant n/2$ a generic collection $\mathcal{A}$ of cardinality $|\mathcal{A}| = 4nr - 4r^2 \leqslant n^2$ is $U(r)$ phase retrievable, whereas a frame for $\text{Sym}(\mathbb{C}^n)$ would consist of at least n^2 elements [9].

Some interesting variants of this problem are:

- $H = \mathbb{R}^n$, $K = \text{Sym}(\mathbb{R}^n)$, $G = O(1) = \{1, -1\}$, $\psi(\lambda, x) = \lambda x$, and $\pi(x) = xx^T$. In this case it was shown in [10] that a collection of the form $\{f_j f_j^T\}_{j=1}^m$ will be $O(1)$ phase retrievable if and only if $\{f_j\}_{j=1}^m$ has the so-called complementing property, that is to say that for every $I \subset \{1, \dots, m\}$ either $\{f_j\}_{j \in I}$ or $\{f_j\}_{j \in I^C}$ spans $\mathbb{R}^n$.

- $H = \mathbb{C}^n$, $K = \text{Sym}(\mathbb{C}^n)$, $G = U(1) = \{e^{i\theta} \mid \theta \in [0, 2\pi)\}$, $\psi(\lambda, x) = \lambda x$, and $\pi(x) = xx^*$. It is shown in [11] that a collection of the form $\{f_j f_j^*\}_{j=1}^m$ will be $U(1)$ phase retrievable if and only if for all $u \in \mathbb{C}^n$ with $||u||_2 = 1$ one has $\text{span}_{\mathbb{R}}\{f_j f_j^* u\}_{j=1}^m = \text{span}_{\mathbb{R}}\{iu\}^{\perp}$.

- $H = \mathbb{C}^{n \times r}$ with $r \leqslant n$, $K = \text{Sym}(\mathbb{C}^n)$, $G = U(r)$, $\psi(U, x) = xU$, and $\pi(x) = xx^*$. It is shown in Theorem 1.8.13 that a collection $\{A_j\}_{j=1}^m \subset \text{Sym}(\mathbb{C}^n)$ will be $U(r)$ phase retrievable if and only if for all $U \in \mathbb{C}^{n \times r}$ having orthonormal columns $\text{span}_{\mathbb{R}}\{A_j U\} = \{UK \mid K^* = -K\}^{\perp}$. This result generalizes both to $r > 1$ and to non rank 1 positive semidefinite frame matrices the analogous result in [11].

- $H = \mathbb{R}^{n \times r}$ with $r \leqslant n$, $K = \text{Sym}(\mathbb{R}^n)$, $G = O(r)$ $\psi(R, x) = xR$, $\pi(x) = xx^T$. This problem, as far as I am aware, has not been studied. I would conjecture, however, that it differs little from the case above, namely that $\{A_j\}_{j=1}^m \subset \text{Sym}(\mathbb{R}^n)$ will be $O(r)$ phase retrievable if and only if for all $U \in \mathbb{R}^{n \times r}$ having orthonormal columns $\text{span}_{\mathbb{R}}\{A_j U\} = \{UK \mid K^T = -K\}^{\perp}$.

- $H = \mathbb{C}^n$, $K = \mathbb{R}^n$, $G = U(1) \times \cdots \times U(1)$, $\psi((\theta_1, \dots, \theta_n), x) = \text{diag}(e^{i\theta_1}, \dots, e^{i\theta_n})x$,

$\pi(x) = \text{diag}(xx^T) = \begin{bmatrix} |x_1|^2 \\ \vdots \\ |x_n|^2 \end{bmatrix}$. In this case $\Delta_\pi = \mathbb{R}^n$, that is any element of $\mathbb{R}^n$ can be written as $\pi(x) - \pi(y)$ for some $x, y \in \mathbb{C}^n$, hence the only $U(1) \times \cdots \times U(1)$ phase retrievable subsets of $\mathbb{R}^n$ are the frames for $\mathbb{R}^n$.

1.2.5 Lipschitz analysis of the phase retrieval problem

If one wishes to "make quantitative" the question of phase retrievability, one option is to strengthen the requirement that the measurement map be invertible to a requirement that it be lower Lipschitz, and then compute its lower Lipschitz constant. Doing so of course requires choosing a metric on H/G. Given the generalized phase retrieval problem set out in Section 1.2.4 there are essentially two reasonable choices for metrics on H/G:

(i) The *induced metric* ρ_π (induced by π and the norm distance on K):

$$\begin{aligned} &\rho_\pi : H/G \times H/G \to \mathbb{R} \\ &\rho_\pi(x, y) = ||\pi(x) - \pi(y)||_K \end{aligned} \tag{1.2.14}$$

(ii) The *natural metric* D:

$$\begin{aligned} &D : H/G \times H/G \to \mathbb{R} \\ &D(x, y) = \inf_{g \in G} ||x - \psi(g, y)||_H \end{aligned} \tag{1.2.15}$$

The fact that the natural metric is symmetric and obeys the triangle inequality follows directly from the fact that ψ is assumed to be linear and norm preserving on H and that G is

a group. If G is compact then of course $D(x,y) = \min_{g\in G} ||x - \psi(g,y)||_H$.

We note that if the set $\Delta_\pi := \text{Ran}(\pi) - \text{Ran}(\pi)$ is closed in K then any G phase retrievable collection $\{A_j\}_{j\in I} \subset K$ will give rise to a β analysis map that is lower Lipschitz with respect to the induced distance since if a_0 is the square of the lower Lipschitz constant for $\beta : (H, \rho_\pi) \to \mathbb{R}^m$ then

$$\begin{aligned} a_0 &= \inf_{x,y\in H} \frac{||\beta(x) - \beta(y)||_2}{||\pi(x) - \pi(y)||_K} \\ &= \min_{\substack{W\in\Delta_\pi \\ ||W||_K=1}} \sum_{i\in I} |\langle A_j, W\rangle_K|^2 \end{aligned} \tag{1.2.16}$$

Noting that for K finite dimensional $\Delta_\pi \cap B_1(0)$ is closed and bounded and hence compact. Thus if $a_0 = 0$ there exists $W_0 \in \Delta_\pi \cap B_1(0)$ such that $0 = \sum_{i\in I} |\langle A_j, W_0\rangle_K|^2$. The fact that $W_0 \in \Delta_\pi$ means that there exists $x_0, y_0 \in H$ such that $W_0 = \pi(x_0) - \pi(y_0)$, and the fact that $||W_0||_K = 1$ implies that $x_0 \not\sim y_0$. Thus plugging $W_0 = \pi(x_0) - \pi(y_0)$ into (1.2.16) yields that $\beta(x_0) = \beta(y_0)$, contradicting the fact that $\{A_j\}_{j\in I}$ is G phase retrievable. Thus $a_0 > 0$ when $\mathcal{A}$ is G phase retrievable. The converse is obviously true, so computation of a_0, while potentially very difficult, gives us a way of checking whether a given collection $\mathcal{A} \subset K$ is G phase retrievable. This fact will eventually be employed to prove Theorem 1.8.13, providing equivalent criteria for a collection $\mathcal{A} \subset \text{Sym}(\mathbb{C}^n)$ to be $U(r)$ phase retrievable.

1.3 The $U(r)$ phase retrieval problem

Let $H = \mathbb{C}^{n\times r}$ with $n \geqslant r$ be the Hilbert space of tall matrices with complex entries, equipped with the real inner product $\langle z, w\rangle_{\mathbb{R}} = \Re\text{tr}\{z^* w\}$, where z^* denotes the transpose complex conjugate of z (the hermitian conjugate). We denote by $\langle z, w\rangle_{\mathbb{C}} = \text{tr}\{z^* w\}$ the complex

inner product and by $\text{Ran}(z) = \{zu | u \in \mathbb{C}^r\}$ the range of z as an operator $z : \mathbb{C}^r \to \mathbb{C}^n$. Let $\mathbb{C}_*^{n\times r}$ be the open subset of $\mathbb{C}^{n\times r}$ consisting of full rank tall matrices. For $p \geqslant 1$ we denote by $||z||_p$ the pth Schatten norm of z, that is to say the l_p norm of the singular values of z. The pseudo-inverse of z will be denoted $z^\dagger$. Let $U(r)$ be the Lie group of $r \times r$ matrices with entries in $\mathbb{C}$ satisfying $U^*U = \mathbb{1}$. We denote by $\mathbb{C}^{n\times r}/U(r)$ and $\mathbb{C}_*^{n\times r}/U(r)$ the set of equivalence classes in $\mathbb{C}^{n\times r}$ and $\mathbb{C}_*^{n\times r}$ respectively under the equivalence relation $z \sim w$ if and only if there exists $U \in U(r)$ such that $z = wU$. Let $S^{p,q}(\mathbb{C}^n)$ denote the set of symmetric operators (hermitian matrices) on $\mathbb{C}^n$ having at most p positive and q negative eigenvalues, and $\mathring{S}^{p,q}(\mathbb{C}^n)$ the set of symmetric operators (hermitian matrices) on $\mathbb{C}^n$ having exactly p positive and q negative eigenvalues. The set $\mathbb{C}^{n\times r}/U(r)$ may then be identified with $S^{r,0}(\mathbb{C}^n)$ and $\mathbb{C}_*^{n\times r}/U(r)$ with $\mathring{S}^{r,0}(\mathbb{C}^n)$ via Cholesky decomposition. Being a finite dimensional space, a *frame* for $\mathbb{C}^{n\times r}$ is a collection $\{f_j\}_{j=1}^m \subset \mathbb{C}^{n\times r}$ that spans $\mathbb{C}^{n\times r}$. In particular, $\{f_j\}_{j=1}^m$ is frame if and only if there exist $A, B > 0$ (called *frame bounds*) satisfying $A||z||_2^2 \leqslant \sum_{j=1}^m |\langle f_j, z\rangle_{\mathbb{R}}|^2 \leqslant B||z||_2^2$ for all $z \in \mathbb{C}^{n\times r}$. This condition may also be written $A||z||_2^2 \leqslant \sum_{j=1}^m \langle A_j, zz^*\rangle_{\mathbb{R}} \leqslant B||z||_2^2$ for all $z \in \mathbb{C}^{n\times r}$ where $A_j = f_j f_j^*$. Using this fact, we may extend the concept of a frame for $\mathbb{C}^{n\times r}$ to collections of symmetric matrices $\{A_j\}_{j=1}^m \subset \text{Sym}(\mathbb{C}^n)$. Fix a frame for $\mathbb{C}^{n\times r}$, then that frame is called *generalized phase retrievable* if the following map is injective:

$$\begin{aligned} &\beta : \mathbb{C}^{n\times r}/U(r) \to \mathbb{R}^m \\ &\beta_j(z) = \langle A_j, zz^*\rangle_{\mathbb{R}}, \qquad j = 1, \ldots, m \end{aligned} \tag{1.3.1}$$

This definition is in agreement with the generalized phase retrieval problem laid out in [12] for the case $r = 1$. Note that if $A_j = f_j f_j^*$ then $\beta_j(z) = ||f_j^* z||_2^2$. A breadth of literature

exists on the classical phase retrieval problem where $r = 1$ and $H = \mathbb{C}^n$ or $H = \mathbb{R}^n$, see for example [10] for an explicit construction of Parseval phase retrievable frames and [13] for a proof of the stability of finite dimensional phase retrievability under perturbation of the frame vectors. In contrast to the finite dimensional case, it is shown in [8] that infinite dimensional phase retrieval is never lower-Lipschitz. Probabilistic error bounds for the case of noisy phase retrieval may be found in [14] for frames sampled from a subgaussian distribution satisfying a so called "small ball" assumption. Efficient algorithms exist for doing classical phase retrieval (for example via Wirtinger flow as in [15]), as well for constructing frames with desirable properties (nearly tight with low coherence) as in [16]. See for example [17] for an analysis of the stability statistics for random frames and [18] for the interesting result that a large class of "non-peaky" vectors (so called μ-flat vectors) are recoverable even when frame vectors are chosen as Bernoulli random vectors, a case in which phase retrieval is well known to fail for arbitrary signals. Recently several advances have been made in understanding natural generalizations of the problem to arbitrary symmetric measurement matrices [12], unifying the problem of phase retrieval with that of fusion frame reconstruction. Lipschitz stability questions for the generalized phase retrieval are analyzed in [19]. The generalized phase retrieval problem in the case $r = 1$ has proven amenable to efficient implementations of gradient descent [20] and a probabilistic guarantee of global convergence of first order methods like gradient descent has been obtained in [21] for $O(n \log^3(n))$ frame vectors. In accordance with the classical phase retrieval we also define the α

map as the entry-wise square root of the beta map (here we require that each $A_j \geqslant 0$):

$$\begin{aligned} &\alpha : \mathbb{C}^{n\times r}/U(r) \to \mathbb{R}^m \\ &\alpha_j(z) = \langle A_j, zz^* \rangle_{\mathbb{R}}^{\frac{1}{2}}, \qquad j = 1, \ldots, m \end{aligned} \tag{1.3.2}$$

Note that if we write $A_j = f_j f_j^*$ using Cholesky decomposition then $\alpha_j(z) = ||f_j^* z||_2$. In this paper we will study the global and local Lipschitz properties of these two maps in the case that the frame is generalized phase retrievable. In particular, we analyze the following (squared) global Lipschitz constants:

$$a_0 := \inf_{\substack{x,y\in\mathbb{C}^{n\times r} \\ x\neq y}} \frac{||\beta(x)-\beta(y)||_2^2}{||xx^*-yy^*||_2^2} \quad , \quad b_0 := \sup_{\substack{x,y\in\mathbb{C}^{n\times r} \\ x\neq y}} \frac{||\beta(x)-\beta(y)||_2^2}{||xx^*-yy^*||_2^2} \tag{1.3.3}$$

$$A_0 := \inf_{\substack{x,y\in\mathbb{C}^{n\times r} \\ x\neq y}} \frac{||\alpha(x)-\alpha(y)||_2^2}{||(xx^*)^{\frac{1}{2}}-(yy^*)^{\frac{1}{2}}||_2^2} \quad , \quad B_0 := \sup_{\substack{x,y\in\mathbb{C}^{n\times r} \\ x\neq y}} \frac{||\alpha(x)-\alpha(y)||_2^2}{||(xx^*)^{\frac{1}{2}}-(yy^*)^{\frac{1}{2}}||_2^2} \tag{1.3.4}$$

In doing so we will employ several distance metrics on $\mathbb{C}^{n\times r}/U(r)$ (equivalently on $S^{r,0}(\mathbb{C}^n)$), the relationships between which are contained in Theorem 1.6.4. The Lipschitz properties of α and β are intimately related to the geometry of $S^{r,0}(\mathbb{C}^n)$, which is the subject of Theorem 1.7.4. Theorem 1.7.4 continues the results in [22] on the geometry of the $n \times n$ positive definite matrices $\mathbb{P}(n)$. The main contributions of this work are thus:

- In Section 1.6 we introduce the novel distance

$$d(x,y) := \sqrt{(||x||_2^2 + ||y||_2^2)^2 - 4||x^*y||_1^2} \tag{1.3.5}$$

on $\mathbb{C}^{n\times r}/U(r)$ and in Theorem 1.6.4 provide optimal Lipschitz constants with respect to

natural embeddings of $(\mathbb{C}^{n\times r}/U(r), d)$ into the Euclidean space $(\text{Sym}(\mathbb{C}^n), ||\cdot||_2)$. This new distance metric allows us in 1.8.5 to compute local lower Lipschitz constants for the β map generalizing those in Theorem 2.5 of [23]. 1.6.4 also provides optimal Lipschitz constants with respect to natural embeddings of $(\mathbb{C}^{n\times r}/U(r), D)$ into $(\text{Sym}(\mathbb{C}^n), ||\cdot||_2)$ for the Bures-Wasserstein distance $D(x,y) := \sqrt{||x||_2^2 + ||y||_2^2 - 2||x^*y||_1}$.

- In Section 1.7 Theorem 1.7.4 generalizes Theorem 5 in [22] by providing the geometry not just of manifold of positive definite matrices $\mathbb{P}(n)$ but of the algebraic semi-variety $S^{r,0}(\mathbb{C}^n)$. In particular we manifest a Whitney stratification of $S^{r,0}(\mathbb{C}^n)$, obtain the Riemannian metrics of the stratifying manifolds, and show that this family of metrics is compatible across the strata in the sense that geodesics of lower strata are limiting curves of geodesics in higher strata. In particular this proves that the geodesic in $S^{r,0}(\mathbb{C}^n)$ connecting two matrices of rank $k < r$ is completely contained in $\mathring{S}^{k,0}(\mathbb{C}^n)$. This stratification of the low rank positive-semidefinite matrices is crucial in simplifying the computation of the global lower Lipschitz bounds for β and α in Theorems 1.8.5 and 1.8.8 respectively.

- In Section 1.8 Theorem 1.8.5 provides an explicit formula for the global lower bound a_0 as the minimization over $U(n)$ of the $(2nr - r^2)$th eigenvalue of a family of matrices parametrized by $U(n)$. Theorem 1.8.5 also uses the distance d to provide a generalization of Theorem 2.5 in [23] to the case $r > 1$ and shows that the analog $\hat{Q}_z$ of $\mathcal{R}(\xi)$ can be used to control a_0 to within a factor of 2. We also show in Theorem 1.8.8 that the corresponding generalization of Theorem 2.2 in [23] to the case $r > 1$ is false, namely that $A_0 = 0$ when $r > 1$. Thus in the case $r > 1$ the more recently introduced β map (the entry-wise square of the α map) is a more natural and well behaved analysis map for generalized phase retrieval,

owing primarily to the fact that it lifts to a linear map on the low rank positive semi-definite matrices. It should be noted that Theorem 1.8.8 does not rule out the possibility of a better distance metric with respect to which α is globally lower Lipschitz. Finally, in Theorem 1.8.13 we provide novel conditions for a frame $\{A_j\}_{j=1}^m$ for $\mathbb{C}^{n\times r}$ to be generalized phase retrievable.

We caution the reader that throughout the paper the scalar product $\langle\cdot,\cdot\rangle_{\mathbb{R}}$ is a real inner product, however z^* denotes the conjugate with respect to the complex inner product $\langle\cdot,\cdot\rangle_{\mathbb{C}}$. We also note that the norm $||z||_p$ for $p \geqslant 1$ is the pth Schatten norm of $z \in \mathbb{C}^{n\times r}$ seen as a $\mathbb{C}$-linear operator from $\mathbb{C}^r$ to $\mathbb{C}^n$. Hence the norm $||\cdot||_2$, while it refers to the Schatten 2 norm, is equivalently given as $||z||_2 = \sqrt{\langle z,z\rangle_{\mathbb{R}}} = \sqrt{\langle z,z\rangle_{\mathbb{C}}}$. If z were instead seen as an $\mathbb{R}$-linear operator from $\mathbb{C}^r$ to $\mathbb{C}^n$ then the resulting Schatten p norm would be amplified by a factor $2^{\frac{1}{p}}$ since the multiplicity of each singular value would double.

1.4 A review of quantitative phase retrievability

The question of phase retrievability criteria for frames for $\mathbb{R}^n$ was addressed in [10], in which it was shown that a frame $\mathcal{F}$ is phase retrievable if and only if it satisfies the "complementing property," that is if and only if for every subset $\mathcal{I} \subset \mathcal{F}$ either $\mathcal{I}$ or $\mathcal{F}\backslash\mathcal{I}$ spans $\mathbb{R}^n$. It was moreover shown in [10] that if $m < 2n-1$ then a frame for $\mathbb{R}^n$ of cardinality m will not be phase retrievable and also that a generic frame for $\mathbb{R}^n$ of size $m \geqslant 2n-1$ will be phase retrievable – that is to say the set $\{\mathcal{F} = \{f_1,\ldots,f_m\} \subset \mathbb{R}^n | \mathcal{F} \text{ is phase retrievable}\}$ will be dense in the Zariski topology when $m \geqslant 2n-1$. The question of phase retrievability *criteria* can be made quantitative by asking for which frames the analysis maps α and β are lower Lipschitz with respect

to some natural distance metrics, and computing their lower Lipschitz constants. Intuitively, a frame is phase retrievable if and only if α (resp. β) is injective, thus it is natural to analyze (for a given frame) the lower Lipschitz constant of α (resp. β), which measures"how" injective α (resp. β) is. In answer to this refinement it was shown in [24] that for the α map and the distance $\rho(x,y) = \min\{||x-y||_2, ||x+y||_2\}$ we have:

Theorem 1.4.1. *(See [24] Theorem 4.3.) For any index set $I \subset \{1,\dots,m\}$ let $\mathcal{F}[I] = \{f_k | k \in I\}$ and let $\sigma_1^2[I] = \lambda_{min}\left(\sum_{k\in I} f_k f_k^*\right)$ and $\sigma_n^2[I] = \lambda_{min}\left(\sum_{k\in I} f_k f_k^*\right)$. Then*

$$A_0 := \inf_{\substack{x,y\in\mathbb{R}^n \\ x\not\sim y}} \frac{||\alpha(x)-\alpha(y)||_2^2}{\rho(x,y)^2} = \min_{I\subset\{1,\dots,m\}} \sigma_n^2[I] + \sigma_n^2[I^C] \tag{1.4.1}$$

This result implies in particular that for a phase retrievable frame for $\mathbb{R}^n$ the α map is globally lower Lipschitz. An analogous result was given in [24] for the β map and the distance $||xx^T - yy^T||_1$:

Theorem 1.4.2. *(See [24] Theorem 2.1.) Let $\{f_j\}_{j=1}^m$ be a phase retrievable frame for $\mathbb{R}^n$ and let $R : \mathbb{R}^n \to Sym(\mathbb{R}^n)$ be given by $R(x) = \sum_{j=1}^m |\langle x, f_j\rangle|^2 f_j f_j^T$. Then*

$$a_0 := \inf_{\substack{x,y\in\mathbb{R}^n \\ x\not\sim y}} \frac{||\beta(x)-\beta(y)||_2^2}{||xx^T - yy^T||_1^2} = \min_{\substack{x\in\mathbb{R}^n \\ ||x||_2=1}} \lambda_n(R(x)) > 0 \tag{1.4.2}$$

Regarding the complex case the following phase retrievability criterion was obtained in [11]:

Theorem 1.4.3. *(See [11] Theorem 4.) Let $\{f_j\}_{j=1}^m$ be a frame for $\mathbb{C}^n$. For $u \in \mathbb{C}^n$ denote $S(u) = span_{\mathbb{R}}\{f_j f_j^* u\}_{j=1}^m$. Then the following are equivalent:*

(i) The frame $\{f_j\}_{j=1}^m \subset \mathbb{C}^n$ is phase retrievable.

(ii) $\dim_{\mathbb{R}} S(u) \geqslant 2n-1$ *for every* $u \in \mathbb{C}^n \backslash \{0\}$.

(iii) $S(u) = span_{\mathbb{R}}\{iu\}^{\perp}$ *for every* $u \in \mathbb{C}^n \in \backslash\{0\}$.

In connection to this paper we note that the above result is extended to the case of generalized retrievability of frames for $\mathbb{C}^{n\times r}$ by Theorem 1.8.13. The quantitative lower Lipschitz variant of Theorem 1.4.3 was obtained for the β analysis map in [23], in which it was proved that for the beta map:

Theorem 1.4.4. *(See [23] Theorem 2.3 and Theorem 2.5.) Let* $\{f_j\}_{j=1}^m$ *be a phase retrievable frame for* $\mathbb{C}^n$. *Define* $\mathcal{R} : \mathbb{R}^{2n} \to Sym(\mathbb{R}^{2n})$ *via* $\mathcal{R}(\xi) = \sum_{j=1}^m \Phi_j \xi\xi^T \Phi_j$ *where* $\Phi_j = \phi_j\phi_j^T + J\phi_j\phi_j^T J^T$, $\phi_j = \begin{bmatrix} \Re f_j \\ \\ \Im f_j \end{bmatrix}$ *and* J *is the symplectic form* $\begin{bmatrix} 0 & -\mathbb{I} \\ \\ \mathbb{I} & 0 \end{bmatrix}$. *Then*

$$a_0 := \inf_{\substack{x,y\in\mathbb{C}^n \\ x\not\sim y}} \frac{||\beta(x)-\beta(y)||_2^2}{||xx^*-yy^*||_1^2} = \min_{\substack{\xi\in\mathbb{R}^{2n} \\ ||\xi||_2=1}} \lambda_{2n-1}(\mathcal{R}(\xi)) > 0 \tag{1.4.3}$$

The connection of the above to Theorem 1.4.3 is that the null space of $\mathcal{R}(\xi)$ includes the realification of $\text{span}_{\mathbb{R}}\{i\xi\}$ for every ξ. Theorem 1.4.4 is extended to the case of generalized phase retrievability of frames for $\mathbb{C}^{n\times r}$ by Theorem 1.8.5.

1.5 Applications of Phase Retrieval

1.5.1 Phase Retrieval in Optics: Fraunhofer diffraction

Historically, one of the first applications of phase retrieval was to the inverse problem of inferring an object's structure from the diffraction pattern it generates when it interacts with an

incident electromagnetic field. It is a fundamental property of Maxwell's equations that in the "far field," when the distance from the interfering object is large compared to the object's size, the structure of the object will be encoded in the Fourier transform of the field. Following [2] we re-cap this principle in its most directly apparent form in the Fraunhofer diffraction regime, but it is quite broadly applicable. For the purposes of this example assume we have a monochromatic field

$$V(x,t) = U(x)e^{-i\omega t} \tag{1.5.1}$$

Here V is a component of the electric (or magnetic) field. In this case the wave equation $c^{-2}\partial_t^2 V - \Delta V = 0$ reduces to Helmholtzs' equation

$$(k^2 + \Delta)U = 0 \tag{1.5.2}$$

Where here $k = \frac{\omega}{c} = \frac{2\pi}{\lambda}$ is the wave number. We would like to obtain a representation formula for $U(x)$ satisfying (1.5.2) in terms of its values on a surface containing x. To this end, consider two solutions U and U' of (1.5.2) and let $V \subset \mathbb{R}^3$ be a compact, connected volume with boundary δV and employ Green's divergence theorem to the vector field $U\nabla U' - U'\nabla U$ to obtain:

$$\int_V U\Delta U' - U'\Delta U dV = \oint_{\delta V} U(\hat{n}\cdot\nabla)U' - U'(\hat{n}\cdot\nabla)U dS \tag{1.5.3}$$

This is of course Green's second identity, with $\hat{n}(y)$ being the outwards pointing unit normal vector to the surface δV at point $y \in \delta V$. If we now substitute $\Delta U' = -k^2 U'$ and $\Delta U = -k^2 U$

into (1.5.3) we find that the left-hand side vanishes, thus

$$\oint_{\delta V} U(\hat{n}\cdot\nabla)U' - U'(\hat{n}\cdot\nabla)U dS = 0 \tag{1.5.4}$$

With this in hand, let $V \ni x$ and let $U'(y) = e^{iks}/s$ where $s = ||y-x||_2$. This choice of U', which is of course a Huygens' wavelet emanating from the point x, is singular at $y = x$, so one is forced to pursue a limiting argument by first excluding from V a ball $B_\epsilon(x)$ of radius ϵ centered at x. In this case, and noting that on the surface of said sphere $\hat{n}\cdot\nabla U' = (ik - s^{-1})e^{iks}/s$, (1.5.4) becomes

$$\begin{aligned}\oint_{\delta V} U(\hat{n}\cdot\nabla)\frac{e^{iks}}{s} - \frac{e^{iks}}{s}(\hat{n}\cdot\nabla)U dS &= -\oint_{\delta B_\epsilon(x)} (ik - s^{-1})e^{iks}U(y)/s - e^{iks}U/s dS(y) \\ &= (1-ik\epsilon)e^{ik\epsilon}\oint_{\delta B_\epsilon(x)} U(y)d\Omega(y)\end{aligned} \tag{1.5.5}$$

The left-hand side of (1.5.5) is independent of ϵ, and in the limit $\epsilon \to 0$ the right-hand side converges to $4\pi U(x)$, thus we conclude:

$$U(x) = \frac{1}{4\pi}\oint_{\delta V} U(\hat{n}\cdot\nabla)\frac{e^{iks}}{s} - \frac{e^{iks}}{s}(\hat{n}\cdot\nabla)U dS \tag{1.5.6}$$

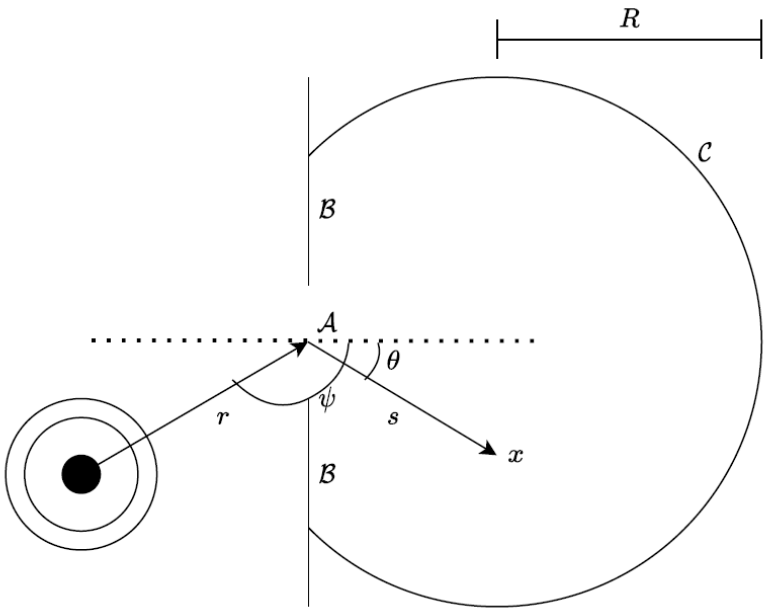


Figure 1.1: From [2].

This result is known as the Fresnel-Kirchoff integral theorem, and follows directly from Green's theorem and (1.5.2). We can apply the Fresnel-Kirchoff theorem to the case of of diffraction through an aperture $\mathcal{A}$ by envisioning an imaginary sphere around that aperture, as in Figure 1.1. In this case (1.5.6) yields

$$U(x) = \frac{1}{4\pi}\left(\int_{\mathcal{A}} + \int_{\mathcal{B}} + \int_{\mathcal{C}}\right) U(\hat{n}\cdot\nabla)\frac{e^{iks}}{s} - \frac{e^{iks}}{s}(\hat{n}\cdot\nabla)U\,dS \tag{1.5.7}$$

At this point some reasonable physical assumptions can be used to arrive at the Fresnel-Kirchoff diffraction formula. In particular if one assumes that R is larger than $c(t-t_0)$ where t_0 is the time of emission then the contribution of the third integral will be zero, since both U and $(\hat{n}\cdot\nabla)U$ will be zero on $\mathcal{C}$. Moreover, Kirchoff assumed that $U|_{\mathcal{B}}$ and its normal derivative are zero, and that $U|_{\mathcal{A}}$ is the same as it would be absent the screen, that is to say $U|_{\mathcal{A}} = Ae^{ikr}/r$ and $(\hat{n}\cdot\nabla)U|_{\mathcal{A}} = \cos(\theta)(ik - r^{-1})e^{ikr}/r$. These assumptions are known as the Kirchoff boundary

conditions. In this case (1.5.7) gives

$$U(x) = \frac{A}{4\pi}\int_{\mathcal{A}} \frac{e^{ik(r+s)}(ik - s^{-1})}{rs}\cos(\theta) + \frac{e^{ik(r+s)}(ik - s^{-1})}{rs}\cos(\psi)dS \tag{1.5.8}$$

In the far field, that is when $\lambda << s$ and $\lambda << r$ the terms $ik - s^{-1}$ and $ik - r^{-1}$ are approximately equal to ik. Thus, in the far field:

$$U(x) = \frac{iA}{2\lambda}\int_{\mathcal{A}} \frac{e^{ik(r+s)}}{rs}[\cos(\theta) + \cos(\psi)]dS \tag{1.5.9}$$

This is the celebrated Fresnel-Kirchoff diffraction formula. On the other hand, if the source x_0 and target x are far from the aperture relative to its size, then $\cos(\theta) + \cos(\psi) \approx 2\cos(\delta)$ where δ is the angle between the screen's normal and $x - x_0$. Moreover, $\frac{1}{rs}$ will be almost constant over $\mathcal{A}$. Thus place the origin in A and let r' and s' be the respective distances of x_0 and x from the origin. In this case

$$U(x) \approx \frac{iA\cos(\delta)}{\lambda r's'}\int_{\mathcal{A}} e^{ik(r+s)}dS \tag{1.5.10}$$

Now parameterize a point in the aperture via $(\xi, \eta, 0)$ so that $dS = d\eta d\xi$, $r^2 = (x_0 - \xi)^2 + (y_0 - \eta)^2 + z_0^2$, $s^2 = (x - \xi)^2 + (y - \eta)^2 + z^2$, $r^2 = x_0^2 + y_0^2 + z_0^2$, and $s^2 = x^2 + y^2 + z^2$. A little algebra yields that $r \approx r' - (x_0\xi + y_0\eta)/r$ and $s \approx s' - (x\xi + y\eta)/s'$ (the neglecting of higher order terms in these expansion is precisely what gives the Fraunhofer diffraction formula) and we obtain the

Fraunhofer diffraction formula

$$U(x) \approx \frac{iA\cos(\delta)}{\lambda r' s'} e^{ik(r'+s')} \int_{\mathcal{A}} e^{ik(x_0/r' - x/s')\xi + (y_0/r' - y/s')\eta} d\eta d\xi \tag{1.5.11}$$

Thus using coordinates $p = (x_0/r' - x/s')/\lambda$ and $q = (y_0/r' - y/s')/\lambda$ yields

$$U(p,q) \propto \int_A e^{-2\pi i(p\xi + q\eta)} d\eta d\xi = \mathcal{F}[\mathbb{1}_{\mathcal{A}}](p,q) \tag{1.5.12}$$

And we obtain that in the far field it is the Fourier transform of the aperture that is encoded in the field. It is typically only possible to measure the *magnitude* of the field for some finite collection $\{x_i\}_{i \in I}$, thus one would like to be able to recover $\mathcal{A}$ (or equivalently $\mathbb{1}_{\mathcal{A}}$) from $\{|\mathcal{F}[\mathbb{1}_{\mathcal{A}}(x_i)]|\}_{i \in I}$.

1.5.2 Phase Retrieval in Inverse Schrodinger Scattering

For simplicity we will consider one dimensional inverse scattering, in which one attempts to recover the scattering potential from the frequency dependent magnitude of the reflected wave. This example follows [7]. Assume a localized potential $V(x)$ so that $V(x) = 0$ for $x < 0$. To the left and right of the support of V the solutions of the time independent Schrodinger equation at spatial frequency k equation are respectively

$$\begin{aligned} \psi_L(x) &= A(k)e^{2\pi ik} + B(k)e^{-2\pi ik} \\ \psi_R(x) &= C(k)e^{2\pi ik} \end{aligned} \tag{1.5.13}$$

Thus $|A(k)|$ is the strength of the incident wave at frequency k, $|B(k)|$ the strength of the reflected wave, and $|C(k)|$ the strength of the transmitted wave. Certain physical potentials without bound

states are completely determined either by $B(k)$ or $C(k)$, but in practice what one is able to measure is $|B(k)|$ (or $|C(k)|$) [7]. In the Born approximation (the high frequency regime) it can be shown that

$$|\int_0^\infty \frac{dV}{dx} e^{2ik} dx| = 4k^2 |B(k)| \tag{1.5.14}$$

Thus one is able to obtain only the *magnitude* of the Fourier transform of $V'(x)$ for each k, and from this one would like to reconstruct $V'(x)$ (and hence $V(x)$).

1.5.3 Phase Retrieval in Speech Processing

The phase retrieval problem arises in a totally different manner in the field of speech processing (and more generally discrete time signal processing). One typically assumes that the speech signal $s \in c(\mathbb{Z})$ is a bounded sequence of the form

$$s(n) = e(n) * \theta(n) = \sum_{k \in \mathbb{Z}} e(k)\theta(n-k) \tag{1.5.15}$$

Where n is the discrete time variable, e is the "excitation signal" containing the actual meaning of the speech and θ models the impulse response of the vocal system to the meaningful excitation (typically θ has finite support) [6]. In recognition and translation tasks one would like therefore to separate out the excitation signal e from the speech signal s. If the signal were instead of the form

$$s(n) = e(n) + w(n) \tag{1.5.16}$$

With w being high frequency noise, linear signal processing has a ready answer in the discrete time Fourier transform and spectral analysis. A low-pass filter of the form $\mathcal{J}[s] = \mathcal{F}^{-1}[\mathbb{1}_{\omega<\omega_0} S]$ where $S(\omega) = \sum_{k\in\mathbb{Z}} s[k]e^{-2\pi ik\omega}$ would be sufficient to isolate $e[n]$ since $\mathcal{J}[s] \approx e$. The convolutional analog to this type of spectral analysis is termed "cepstral analysis" and was introduced by Bogert, Healy, and Tukey in [25] and generalized by Oppenheim and Schafer to "homomorphic signal analysis" in [26]. Analogous to the Fourier transform, the backbone of cepstral analysis is the so-called real cepstrum:

$$\mathcal{H}[s](n) = \mathcal{F}^{-1}[\log|S(\omega)|] = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-2\pi i\omega n}\log|S(\omega)|d\omega \tag{1.5.17}$$

Here n is neither the frequency nor the discrete time, and is termed the "quefrency." The field is replete with a dictionary of such Seussian terms ("rahmonics" replace harmonics, "saphes" replace phases, etc). The complex analog of $\mathcal{H}$ (using the complex logarithm to avoid the phase annihilating absolute value) is used in [26] and has several nice theoretical properties, but turns out to be less useful in practical speech processing tasks than (1.5.17) [6]. The key property of $\mathcal{H}$ that allows it to play the role of the Fourier transform with respect to (1.5.15) is that

$$\mathcal{H}[e * \theta](n) = \mathcal{F}^{-1}[\log|E(\omega)| + \log|\Theta(\omega)|] = \mathcal{H}[e](n) + \mathcal{H}[\theta](n) \tag{1.5.18}$$

Because of this homomorphism, it makes sense to perform filtering (termed "liftering") in the quefrency domain. With appropriate quefrency filters and a forward discrete time Fourier transform one is thus able to approximately isolate $|E(\omega)|$, at which point recovery of the excitation signal $e(n)$ is precisely the phase retrieval problem.

1.5.4 Phase Retrieval in Quantum Tomography

A motivating example for the Lipschitz analysis of α and β is quantum tomography of impure states. A noisy quantum system is modeled as a statistical ensemble over pure quantum states. The standard example is unpolarized light. In such cases, all of the measurable information in the system is contained in a density matrix which, using bra-ket notation, has the form

$$\rho = \sum_{j \in \mathcal{I}} p_j |\psi_j\rangle\langle\psi_j| \tag{1.5.19}$$

where p_j is the ensemble probability that the system is in the pure quantum state $|\psi_j\rangle$ belonging to a Hilbert space H. If we assume the cardinality of $\mathcal{I}$ is finite and equal to r and that the state vectors themselves live in the Hilbert space $\mathbb{C}^n$ then $\rho \in S^{r,0}(\mathbb{C}^n) \cap \{x \in \text{Sym}(\mathbb{C}^n) | \text{tr}\{x\} = 1\}$. The expectation of a given observable A (a symmetric operator on $\mathbb{C}^n$) is therefore

$$\mathbb{E}_\rho[A] = \sum_{j \in \mathcal{I}} p_j \langle\psi_j|A|\psi_j\rangle = \sum_{j \in \mathcal{I}} p_j \text{tr}\{|\psi_j\rangle\langle\psi_j|A\} = \text{tr}\{\rho A\} = \Re\text{tr}\{\rho A\} \tag{1.5.20}$$

By repeatedly measuring the observable A and then allowing the quantum system to relax one may estimate $\text{tr}\{\rho A\}$ (and perhaps higher moments) but the aim is to infer ρ itself. It was shown in [27] that sufficiently many randomly sampled Pauli observables can be used along with methods from compressed sensing (trace minimization, matrix Lasso) to reconstruct a low rank density matrix with high fidelity. In general, if a suite of observables is well-chosen (constitutes a generalized phase-retrievable frame) then the problem of inferring ρ from the expectation values of said observables is subordinate to the problem of phase retrieval on $\mathbb{C}^{n \times r}$. Asking if, for a col-

lection of observables $\{A_j\}_{j=1}^m$, the density matrix ρ is recoverable is equivalent to asking if the map

$$\tilde{\beta} : S^{r,0}(\mathbb{C}^n) \cap \{x \in \text{Sym}(\mathbb{C}^n) | \text{tr}\{x\} = 1\} \to \mathbb{R}^m$$
$$\tilde{\beta}(\rho) = \begin{bmatrix} \langle \rho, A_1 \rangle_{\mathbb{R}} \\ \vdots \\ \langle \rho, A_m \rangle_{\mathbb{R}} \end{bmatrix} \tag{1.5.21}$$

is injective. In fact, given that we can only approximate the expectations using finitely many measurements, we should hope that it is lower Lipschitz with respect to the Frobenius distance. Such stability questions for phase retrievable frames for $\mathbb{C}^n$ (the pure state case) are investigated in [13]. Given that ρ is positive semidefinite and rank at most r there exists a Cholesky factor $z \in \mathbb{C}^{n \times r}$ such that $\rho = zz^*$. Indeed we may take $z \in \mathbb{C}^{n \times r}/U(r)$ since ρ is invariant under $z \to zU$, in which case $\text{tr}\{\rho\} = 1$ if and only if $||z||_2 = 1$. We may therefore concern ourselves with the Lipschitz properties of β restricted to $z \in \mathbb{C}^{n \times r}/U(r)$ with $||z||_2 = 1$, rather than $\tilde{\beta}$. For the time being we consider a Lipschitz analysis of $\beta : \mathbb{C}^{n \times r}/U(r) \to \mathbb{R}^m$, deferring discussion of a possible Lipschitz retract onto the unit sphere. Thus we seek information on the optimal global lower Lipschitz constant of the β map, namely $\sqrt{a_0}$. In the above example if $a_0 > 0$ this means that if we can measure each $\mathbb{E}_\rho[A_j]$ to within error $\epsilon > 0$ then we can obtain an approximation $\hat{\rho}$ to ρ that satisfies

$$||\rho - \hat{\rho}||_2 \leqslant \frac{\epsilon\sqrt{m}}{\sqrt{a_0}} \tag{1.5.22}$$

In addition to quantum state tomography, Lipschitz analysis of spaces of low-rank matrices

is central in a significant number of problems in science and engineering such as: the phase retrieval problem [10, 28], source separation and inverse problems [29], as well as the low-rank matrix completion problem [30].

1.6 Relevant distances and Lipschitz embeddings

Definition 1.6.1. We define the equivalence relation $\sim$ on $\mathbb{C}^{n\times r}$ via

$$x \sim y \iff \exists U \in U(r) | x = yU \tag{1.6.1}$$

and denote by $[x]$ the equivalence class of $x \in \mathbb{C}^{n\times r}$, and by $\mathbb{C}^{n\times r}/U(r)$ the collection of equivalence classes $\{[x]|x \in \mathbb{C}^{n\times r}\}$.

The stability analysis that follows for β and α in Theorems 1.8.5 and 1.8.8 will rely heavily on the following natural metrics on $\mathbb{C}^{n\times r}/U(r)$.

Definition 1.6.2. We define $D, d : \mathbb{C}^{n\times r} \times \mathbb{C}^{n\times r} \to \mathbb{R}$.

$$\begin{aligned} D(x,y) &= \min_{U\in U(r)} ||x - yU||_2 \\ &= \sqrt{||x||_2^2 + ||y||_2^2 - 2||x^*y||_1} \\ d(x,y) &= \min_{U\in U(r)} ||x - yU||_2||x + yU||_2 \\ &= \sqrt{(||x||_2^2 + ||y||_2^2)^2 - 4||x^*y||_1^2} \end{aligned} \tag{1.6.2}$$

We note that another distance on $\mathbb{C}^{n\times r}/U(r)$ given by

$$\begin{aligned} D'(x,y) &= \max_{U\in U(r)} ||x - yU||_2 \\ &= \sqrt{||x||_2^2 + ||y||_2^2 + 2||x^*y||_1} \end{aligned} \tag{1.6.3}$$

and is introduced and analyzed for the $r = 1$ case in [31]. We note merely that $d = D \cdot D'$. This does not imply d is a metric, however in fact we have the following proposition.

Proposition 1. *Both D and d are metrics in the usual sense on $\mathbb{C}^{n\times r}/U(r)$.*

Proof. See 1.9.1. □

The proof of Proposition 1 relies on Lemma 1.9.1, an apparently simple result about the analytic geometry of parallelepipeds in $\mathbb{R}^3$ which may be of independent interest.

The minimizer U can be chosen to be the same for both d and D, and is characterized by the following:

Proposition 2. *The unitary minimizer in both d and D is given by the polar factor in $x^*yU = |x^*y|$. The minimizer will be unique so long as x^*y is full rank. Otherwise, the minimizer will be of the form $U = U_0 + U_1$ where $U_0 = V_0W_0^*$ with $V_0, W_0 \in \mathbb{C}^{r\times rank(x^*y)}$ the matrices whose columns are the right and left singular vectors respectively of the non-zero singular values of x^*y and $U_1 \in \mathbb{C}^{r\times r}$ any matrix such that $U_1U_1^* = \mathbb{P}_{\ker(x^*y)}$ and $U_1^*U_1 = \mathbb{P}_{Ran(x^*y)^\perp}$.*

Proof. See 1.9.2 □

The metrics d and D can be compared to the usual Euclidean distance on $\mathrm{Sym}(\mathbb{C}^n)$ modulo certain embeddings.

Definition 1.6.3. We define $\theta, \pi, \psi : \mathbb{C}^{n\times r} \to S^{r,0}(\mathbb{C}^n)$ as

$$\begin{aligned} \theta(x) &= (xx^*)^{\frac{1}{2}} \\ \pi(x) &= xx^* = \theta(x)^2 \\ \psi(x) &= ||x||_2(xx^*)^{\frac{1}{2}} = ||\theta(x)||_2\theta(x) \end{aligned} \tag{1.6.4}$$

Proposition 3. *The embeddings π, θ, and ψ are rank-preserving, surjective, and injective modulo $\sim$, thus we write $\theta, \pi, \psi : \mathbb{C}^{n\times r}/U(r) \hookrightarrow Sym(\mathbb{C}^n)$.*

Proof. See 1.9.3 □

Theorem 1.6.4. *Let $x, y \in \mathbb{C}^{n\times r}/U(r)$. Then*

(i) $\theta : (\mathbb{C}^{n\times r}/U(r), D) \to (S^{r,0}(\mathbb{C}^n), ||\cdot||_2)$ is a bi-Lipschitz map. In particular,

$$C_n||\theta(x) - \theta(y)||_2 \leqslant D(x, y) \leqslant ||\theta(x) - \theta(y)||_2 \tag{1.6.5}$$

where $C_n = 1$ if $n = 1$ and $C_n = \frac{1}{\sqrt{2}}$ for $n > 1$. The constants C_n and 1 are optimal.

(ii) $\pi : (\mathbb{C}^{n\times r}/U(r), d) \to (S^{r,0}(\mathbb{C}^n), ||\cdot||_1)$ is 1-Lipschitz and $\psi^{-1} : (S^{r,0}(\mathbb{C}^n), ||\cdot||_2) \to (\mathbb{C}^{n\times r}/U(r), d)$ is 2-Lipschitz for $r > 2$ and $\sqrt{2}$-Lipschitz for $r = 1$. In particular,

$$||\pi(x) - \pi(y)||_2 \leqslant ||\pi(x) - \pi(y)||_1 \leqslant d(x, y) \leqslant c_r||\psi(x) - \psi(y)||_2 \tag{1.6.6}$$

where $c_r = \sqrt{2}$ if $r = 1$ and $c_r = 2$ if $r > 1$. The constants 1 and c_r are optimal.

(iii) For $r = 1$

$$\psi(x) = \pi(x) \tag{1.6.7}$$

$$d(x, y) = ||\pi(x) - \pi(y)||_1 \tag{1.6.8}$$

The identity (1.6.8) *was noticed and used in [23], its proof is included here for the benefit of the reader.*

(iv) For $r > 1$*, there is no constant* C *satisfying* $C||\pi(x)-\pi(y)||_2 \geqslant d(x, y)$ *for each* $x, y \in \mathbb{C}^{n\times r}$ *(hence the use of the alternate embedding* ψ*).*

Proof. See 1.9.4 □

Remark 1.6.5. While d and D are evidently not Lipschitz equivalent (they scale differently), they do generate the same topology on $\mathbb{C}^{n\times r}/U(r)$ since $d(x, y) \leqslant D(x, y)^2$ and given sufficiently small $\epsilon > 0$ we have $d(x, y) < ||x||\sqrt{\epsilon} \implies D(x, y) < \epsilon$.

1.7 Geometry of the matrix phase retrieval

It will be essential in the analysis and computation of (1.3.3) to understand the geometry of the spaces $S^{r,0}(\mathbb{C}^n)$. In order to do so, we will demonstrate that $S^{r,0}(\mathbb{C}^n)$ has a Whitney stratification over the smooth Riemannian manifolds $\mathring{S}^{i,0}(\mathbb{C}^n)$ for $i = 0, \ldots, r$ of real dimension $2ni - i^2$. We recall the following definitions, due to John Mather and sourced from [32]:

Definition 1.7.1. Let V_i, V_j be disjoint real manifolds embedded in $\mathbb{R}^d$ such that $\dim V_j > \dim V_i$ and $V_i \cap \overline{V_j}$ non-empty. Let $x \in V_i \cap \overline{V_j}$. Then a triple (V_j, V_i, x) is called $a-$ (resp. $b-$) regular if

(a) If a sequence $(y_n)_{n\geqslant 1} \subset V_j$ converges to x in $\mathbb{R}^d$ and $T_{y_n}(V_j)$ converges in the Grassmannian $\mathrm{Gr}_{\dim V_j}(\mathbb{R}^d)$ to a subspace τ_x of $\mathbb{R}^d$ then $T_x(V_i) \subset \tau_x$.

(b) If sequences $(y_n)_{n\geqslant 1} \subset V_j$ and $(x_n)_{n\geqslant 1} \subset V_i$ converge to x in $\mathbb{R}^d$, the unit vector $(x_n - y_n)/||x_n - y_n||_2$ converges to a vector $v \in \mathbb{R}^d$, and $T_{y_n}(V_j)$ converges in the Grassmannian $\mathrm{Gr}_{\dim V_j}(\mathbb{R}^d)$ to a subspace τ_x of $\mathbb{R}^d$ then $v \in \tau_x$.

Definition 1.7.2. Let V be a real semi-algebraic variety. A disjoint decomposition

$$V = \bigsqcup_{i\in I} V_i, \qquad V_i \cap V_j = \varnothing \text{ for } i \neq j \tag{1.7.1}$$

into smooth manifolds $\{V_i\}_{i\in I}$, termed strata, is a Whitney stratification if

(a) Each point has a neighborhood intersecting only finitely many strata

(b) The boundary sets $\overline{V_j}\backslash V_j$ of each stratum V_j are unions of other strata.

(c) Every triple (V_j, V_i, x) such that $x \in V_i \subset \overline{V_j}$ is a-regular and b-regular as in Definition 1.7.1.

A simple example of a semi-algebraic variety that is not a manifold but admits a Whitney stratification is the cone $\mathcal{C} = \{(x, y)|xy \geqslant 0\} \subset \mathbb{R}^2$ consisting off the first and third quadrant of the coordinate plane. A possible Whitney stratification of this set is given by $V_0 = \{0\}$, $V_1 = \{(x, 0)|x \neq 0\}$, $V_2 = \{(0, y)|y \neq 0\}$, and $V_3 = \{(x, y)|x \neq 0, y \neq 0\}$. In this case note that condition (a) is trivially satisfied since there are only finitely many strata, and moreover that (b) is satisfied since $\overline{V_3}\backslash V_3 = V_0 \cup V_1 \cup V_2$, $\overline{V_2}\backslash V_2 = V_0$, $\overline{V_1}\backslash V_1 = V_0$, and that $\overline{V_0}\backslash V_0 = \phi$ (an empty union of the other strata). That this stratification is both (a) and (b) regular may be readily observed. For example the tangent space at any point of V_3 is simply $\mathbb{R}^2$, and thus

the Grassmanian limit of a convergent sequence of such tangent spaces is also $\mathbb{R}^2$ and certainly contains the one dimensional tangent space at any point of V_2 (identified with the y axis), the one dimensional tangent space at any point of V_1 (identified with the x axis), and the zero dimensional tangent space associated with V_0 (identified with the origin).

We will also need the following:

Definition 1.7.3. Let $\mathcal{M}$ and $\mathcal{N}$ be smooth manifolds and let $\pi : \mathcal{M} \to \mathcal{N}$ be a smooth map. For each $x \in \mathcal{M}$ let

$$T_x(\mathcal{M}) := \{\gamma'(0) | \gamma : [-1, 1] \to \mathcal{M} \text{ is a smooth curve with } \gamma(0) = x\} \tag{1.7.2}$$

be the tangent space of $\mathcal{M}$ at x. Similarly for $T_{\pi(x)}(\mathcal{N})$. Let $D\pi(x) : T_x(\mathcal{M}) \to T_{\pi(x)}(\mathcal{N})$ be the differential of π at x, that is to say $D\pi(x)(v) := \alpha'(0)$ where $\alpha = \pi \circ \gamma$, $\gamma(0) = x$, and $\gamma'(0) = v$ (that $D\pi(x)$ does not depend on the exact choice of curve γ is an elementary result of differential geometry). Then

(a) For each $x \in \mathcal{M}$ define the vertical space at x as:

$$V_{\pi,x}(\mathcal{M}) \subset T_x(\mathcal{M}) := \ker D\pi(x) = \{w \in T_x(\mathcal{M}) | D\pi(x)(w) = 0\} \tag{1.7.3}$$

(b) If $\mathcal{M}$ is equipped with a Riemannian metric $g : \mathcal{M} \times T_x(\mathcal{M}) \times T_x(\mathcal{M}) \to \mathbb{R}$ then we may define the horizontal space at each x via the canonical orthogonal complement of the vertical

space:

$$H_{\pi,x}(\mathcal{M}) \subset T_x(\mathcal{M}) := V_{\pi,x}(\mathcal{M})^{\perp} = \{v \in T_x(\mathcal{M}) | g(x,v,w) = 0 \forall w \in V_{\pi,x}(\mathbb{C}_*^{n\times r})\} \tag{1.7.4}$$

The following proposition will be essential both in proving the geometric results in Theorem 1.7.4 and in the analysis of the Lipschitz constants for β and α set out in Theorems 1.8.5, 1.8.8, and 1.8.12:

Proposition 4. *Let $\pi : \mathbb{C}_*^{n\times r} \to \mathring{S}^{r,0}(\mathbb{C}^n)$ be as in Definition 1.6.3 and let $V_{\pi,x}(\mathbb{C}_*^{n\times r})$ and $H_{\pi,x}(\mathbb{C}_*^{n\times r})$ denote the vertical and horizontal spaces as in Definition 1.7.3 of the manifold $\mathbb{C}_*^{n\times r}$ at x with respect to the embedding π. Here the Riemmanian metric on $\mathbb{C}_*^{n\times r}$ is of course $g : \mathbb{C}_*^{n\times r} \times \mathbb{C}^{n\times r} \times \mathbb{C}^{n\times r} \to \mathbb{R}$ given by $g(x,v,w) = \Re tr\{z^* w\}$. Let $T_{\pi(x)}(\mathring{S}^{r,0}(\mathbb{C}^n))$ denote the tangent space of $\mathring{S}^{r,0}(\mathbb{C}^n)$ at $\pi(x)$. Then*

$$V_{\pi,x}(\mathbb{C}_*^{n\times r}) = \{xK | K \in \mathbb{C}^{r\times r}, K^* = -K\} \tag{1.7.5}$$

$$H_{\pi,x}(\mathbb{C}_*^{n\times r}) = \{Hx + X | H \in \mathbb{C}^{n\times n}, H^* = H = \mathbb{P}_{\mathit{Ran}(x)} H, \; X \in \mathbb{C}^{n\times r}, \mathbb{P}_{\mathit{Ran}(x)} X = 0\} \tag{1.7.6}$$

$$T_{\pi(x)}(\mathring{S}^{r,0}(\mathbb{C}^n)) = \{W \in \mathit{Sym}(\mathbb{C}^n) | \mathbb{P}_{\mathit{Ran}(x)^{\perp}} W \mathbb{P}_{\mathit{Ran}(x)^{\perp}} = 0\} = D\pi(x)(H_{\pi,x}(\mathbb{C}_*^{n\times r})) \tag{1.7.7}$$

Proof. See 1.10.1 □

Employing similar techniques to [22], but generalizing from the manifold of positive definite matrices to the semi-algebraic variety $S^{r,0}(\mathbb{C}^n)$ semidefinite matrices, we prove:

Theorem 1.7.4. *Let π be as in Definition 1.6.3 and the distance D be as in (1.6.2). Then*

(i) $\mathring{S}^{p,q}(\mathbb{C}^n)$ *is a real analytic manifold for each* $p, q > 0$ *of real dimension* $2n(p+q)-(p+q)^2$.

(ii) $\pi : \mathbb{C}_*^{n\times r} \to \mathring{S}^{r,0}(\mathbb{C}^n)$ *can be made into a Riemannian submersion by choosing the following unique Riemannian metric on* $\mathring{S}^{r,0}(\mathbb{C}^n)$*:*

$$h(Z_1, Z_2) = tr\{Z_2^{\|} \int_0^{\infty} e^{-uxx^*} Z_1^{\|} e^{-uxx^*} du\} + \Re tr\{Z_1^{\perp *} Z_2^{\perp} (xx^*)^{\dagger}\} \tag{1.7.8}$$

Where $Z_1, Z_2 \in T_{\pi(x)}(\mathring{S}^{r,0}(\mathbb{C}^n))$, $(xx^*)^{\dagger}$ *denotes the pseudo-inverse of* xx^**, and*

$$Z_i^{\|} = \mathbb{P}_{Ran(x)} Z_i \mathbb{P}_{Ran(x)} \qquad Z_i^{\perp} = \mathbb{P}_{Ran(x)^{\perp}} Z_i \mathbb{P}_{Ran(x)} \tag{1.7.9}$$

(iii) $\mathring{S}^{r,0}(\mathbb{C}^n)$ *equipped with the metric* h *is a Riemannian manifold with* D *as its geodesic distance.*

(iv) The semi-algebraic variety $S^{r,0}(\mathbb{C}^n)$ *admits as an explicit Whitney stratification* $(\mathring{S}^{i,0})_{i=0}^{r}$.

(v) The geometry associated to h *is compatible with the Whitney stratification in the following sense: If* $(A_i)_{i\geqslant 1}, (B_i)_{i\geqslant 1} \subset \mathring{S}^{p,0}$ *have limits* A *and* B *respectively in* $\mathring{S}^{q,0}$ *for* $q < p$ *and if* $\gamma_i : [0,1] \to \mathring{S}^{p,0}$ *are geodesics in* $\mathring{S}^{p,0}$ *connecting* A_i *to* B_i *chosen in such a way that the limiting curve* $\delta : [0,1] \to \overline{\mathring{S}^{p,0}}$ *given by*

$$\delta(t) = \lim_{i\to\infty} \gamma_i(t) \tag{1.7.10}$$

exists, then the image of δ *lies in* $\mathring{S}^{q,0}$ *and is a geodesic curve in* $\mathring{S}^{q,0}$ *connecting* A *to* B.

Proof. See 1.10.2 □

1.8 Computation of Lipschitz bounds

We are primarily interested in computing a_0 and A_0, the squared global lower Lipschitz constants for the β and α analysis maps respectively. Owing to the linearity of the β analysis map when interpreted as in (1.5.21), we will be able to show in Theorem 1.8.5 that the optimal global lower Lipschitz bound a_0 can be obtained via local considerations. For the α analysis map we will be able to show in Theorem 1.8.8 that the optimal global lower Lipschitz bound A_0 is actually zero for $r > 1$. Since the global lower Lipschitz bound for the α analysis map is trivial we emphasize the analysis of the local lower Lipschitz bounds. Recall that

$$a_0 = \inf_{\substack{x,y\in\mathbb{C}^{n\times r}\\ [x]\neq[y]}} \frac{||\beta(x)-\beta(y)||_2^2}{||\pi(x)-\pi(y)||_2^2} = \inf_{\substack{x,y\in\mathbb{C}^{n\times r}\\ [x]\neq[y]}} \frac{\sum_{j=1}^m (\langle xx^*, A_j\rangle_{\mathbb{R}} - \langle yy^*, A_j\rangle_{\mathbb{R}})^2}{||xx^* - yy^*||_2^2} \tag{1.8.1}$$

From purely topological considerations, we may obtain

Proposition 5. *The constant a_0 is strictly positive whenever the map β is injective, equivalently whenever $\{A_j\}_{j=1}^m$ is a generalized phase retrievable frame of symmetric matrices.*

Proof. See 1.11.1 □

Definition 1.8.1. Let $z \in \mathbb{C}^{n\times r}$ have rank k. We will analyze the following four types of local lower Lipschitz bounds for β, the first two with respect to the norm induced metric and the second

two with respect to the metric d:

$$
\begin{aligned}
a_1(z) &= \lim_{R\to 0} \inf_{\substack{x\in\mathbb{C}^{n\times r} \\ ||\pi(x)-\pi(z)||_2<R}} \frac{||\beta(x)-\beta(z)||_2^2}{||\pi(x)-\pi(z)||_2^2} \\
a_2(z) &= \lim_{R\to 0} \inf_{\substack{x,y\in\mathbb{C}^{n\times r} \\ ||\pi(x)-\pi(z)||_2<R \\ ||\pi(y)-\pi(z)||_2<R}} \frac{(||\beta(x)-\beta(y)||_2^2}{||\pi(x)-\pi(y)||_2^2} \\
\hat{a}_1(z) &= \lim_{R\to 0} \inf_{\substack{x\in\mathbb{C}^{n\times r} \\ d(x,z)<R \\ \text{rank}(x)\leqslant k}} \frac{||\beta(x)-\beta(z)||_2^2}{d(x,z)^2} \\
\hat{a}_2(z) &= \lim_{R\to 0} \inf_{\substack{x,y\in\mathbb{C}^{n\times r} \\ d(x,z)<R \\ d(y,z)<R \\ \text{rank}(x)\leqslant k \\ \text{rank}(y)\leqslant k}} \frac{||\beta(x)-\beta(y)||_2^2}{d(x,y)^2}
\end{aligned}
\tag{1.8.2}
$$

Note that in the definition of $\hat{a}_1(z)$ and $\hat{a}_2(z)$ we do not allow the ranks of x and y to exceed that of z. As we shall prove, without the rank constraints these local lower bounds would be zero.

The following two "geometric" local lower bounds will prove helpful in our analysis.

Definition 1.8.2. Let $z \in \mathbb{C}^{n\times r}$ have rank k and let $\hat{z} \in \mathbb{C}_*^{n\times k}$ be such that there exists $U \in U(r)$ with $[\hat{z}|0]U = z$. Let $T_{\pi(\hat{z})}(\mathring{S}^{k,0}(\mathbb{C}^n))$ and $H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k})$ be as 1.7.7 and 1.7.6. We define:

$$
a(z) := \min_{\substack{W\in T_{\pi(\hat{z})}(\mathring{S}^{k,0}(\mathbb{C}^n)) \\ ||W||_2=1}} \sum_{j=1}^{m} |\langle W, A_j\rangle_{\mathbb{R}}|^2 \tag{1.8.3}
$$

$$
\hat{a}(z) := \min_{\substack{w\in H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k}) \\ ||w||_2=1}} \sum_{j=1}^{m} |\langle D\pi(\hat{z})(w), A_j\rangle_{\mathbb{R}}|^2 \tag{1.8.4}
$$

The following two families of matrices, Q_z and $\hat{Q}_z$, indexed by $\mathbb{C}^{n\times r}$, will allow us to write the local lower Lipschitz bounds with respect to $||xx^* - yy^*||_2$ and $d(x,y)$ as eigenvalue problems.

Definition 1.8.3. Given $z \in \mathbb{C}^{n\times r}$ having rank $k > 0$ we define a matrix $Q_z \in \mathbb{R}^{(2nk-k^2)\times(2nk-k^2)}$ in the following way. Let $U_1 \in \mathbb{C}^{n\times k}$ be a matrix whose columns are left singular vectors of z corresponding to non-zero singular values of z, so that $U_1U_1^* = \mathbb{P}_{\text{Ran}(z)}$. Let $U_2 \in \mathbb{C}^{n\times(n-k)}$ be a matrix whose columns are left singular vectors of z corresponding to the zero singular values of z, so that $U_2U_2^* = \mathbb{P}_{\text{Ran}z^\perp}$. Then

$$Q_z := \sum_{j=1}^{m} \begin{bmatrix} \tau(U_1^*A_jU_1) \\ \\ \mu(U_2^*A_jU_1) \end{bmatrix} \begin{bmatrix} \tau(U_1^*A_jU_1) \\ \\ \mu(U_2^*A_jU_1) \end{bmatrix}^T \tag{1.8.5}$$

where the isometric isomorphisms τ and μ are given by

$$\tau : \text{Sym}(\mathbb{C}^k) \to \mathbb{R}^{k^2} \qquad \mu : \mathbb{C}^{p\times q} \to \mathbb{R}^{2pq} \tag{1.8.6}$$

$$\tau(X) = \begin{bmatrix} D(X) \\ \sqrt{2}T(\Re X) \\ \sqrt{2}T(\Im X) \end{bmatrix} \qquad \mu(X) = \text{vec}(\begin{bmatrix} \Re X \\ \\ \Im X \end{bmatrix})$$

where

$$D : \text{Sym}(\mathbb{C}^k) \to \mathbb{R}^k \qquad T : \text{Sym}(\mathbb{R}^k) \to \mathbb{R}^{\frac{1}{2}k(k-1)} \tag{1.8.7}$$

$$D(W) = \begin{bmatrix} X_{11} \\ \vdots \\ X_{kk} \end{bmatrix} \qquad T(X) = \begin{bmatrix} X_{12} \\ X_{13} \\ X_{23} \\ \vdots \\ X_{k-1k} \end{bmatrix}$$

and

$$\text{vec} : \mathbb{R}^{p\times q} \to \mathbb{R}^{pq} \qquad \text{vec}(X) = \text{vec}([X_1|\cdots|X_q]) = \begin{bmatrix} X_1 \\ \vdots \\ X_q \end{bmatrix} \tag{1.8.8}$$

We note that Q_z depends only on $\text{Ran}(z)$, in particular it is invariant under $(U_1, U_2) \to (U_1P, U_2Q)$ for $P \in U(k), Q \in U(n-k)$. We will also refer to Q_z as $Q_{[U_1|U_2]}$ where $[U_1|U_2] \in U(n)$.

Definition 1.8.4. Given $z \in \mathbb{C}^{n\times r}$ having rank $k > 0$ we define a matrix $\hat{Q}_z \in \mathbb{R}^{2nk\times 2nk}$ in the following way. Let $F_j = \mathbb{I}_{k\times k} \otimes j(A_j) \in \mathbb{R}^{2nk\times 2nk}$ where

$$\begin{aligned} &j : \mathbb{C}^{m\times n} \to \mathbb{R}^{2m\times 2n} \\ &j(X) = \begin{bmatrix} \Re X & -\Im X \\ \Im X & \Re X \end{bmatrix} \end{aligned} \tag{1.8.9}$$

is an injective homomorphism. Then

$$\hat{Q}_z := 4\sum_{j=1}^{m} F_j \mu(\hat{z})\mu(\hat{z})^T F_j \tag{1.8.10}$$

With these definitions in mind, we will prove the following:

Theorem 1.8.5. *Let $z \in \mathbb{C}^{n\times r}$ have rank $k > 0$. Then*

(i) The global lower bound a_0 is given as

$$a_0 = \inf_{z \in \mathbb{C}^{n \times r} \setminus \{0\}} a(z) \tag{1.8.11}$$

(ii) The local lower bounds $a_1(z)$ and $a_2(z)$ are squeezed between a_0 and $a(z)$

$$a_0 \leqslant a_2(z) \leqslant a_1(z) \leqslant a(z) \tag{1.8.12}$$

So that in particular

$$a_0 = \inf_{z \in \mathbb{C}^{n \times r} \setminus \{0\}} a_i(z) \tag{1.8.13}$$

(iii) The infimization problem in $a(z)$ may be reformulated as an eigenvalue problem. Let Q_z be the $2nk - k^2 \times 2nk - k^2$ matrix given in Definition 1.8.3. Then

$$a(z) = \lambda_{2nk-k^2}(Q_z) \tag{1.8.14}$$

(iv) For $r = 1$, $\hat{a}(z)$ differs from $a(z)$ by a constant factor, hence for $r = 1$ the infimum $\inf_{z \in \mathbb{C}^{n \times r} \setminus \{0\}} \hat{a}(z)$ is non-zero. For $r > 1$ this infimum is zero and hence there is no non-trivial global lower bound $\hat{a}_0$ analogous to a_0 for the alternate metric d.

(v) The local lower bounds with respect to the alternate metric d satisfy

$$\hat{a}_1(z) = \hat{a}_2(z) = \frac{1}{4||z||_2^2} \hat{a}(z) \tag{1.8.15}$$

(vi) The infimization problem in $\hat{a}(z)$ may be reformulated as an eigenvalue problem. Let $\hat{Q}_z$ be the $2nk \times 2nk$ matrix given in Definition 1.8.4. Then $\hat{a}(z)$ is directly computable as

$$\hat{a}(z) = \lambda_{2nk-k^2}(\hat{Q}_z) \tag{1.8.16}$$

(vii) We have the following local inequality relating $a(z)$ and $\hat{a}(z)$.

$$\frac{1}{4||z||_2^2}\hat{a}(z) \leqslant a(z) \leqslant \frac{1}{2\sigma_k(z)^2}\hat{a}(z) \tag{1.8.17}$$

(viii) Computation of the global lower bound a_0 may be reformulated as the minimization of a continuous quantity over the compact Lie group $U(n)$.

$$a_0 = \min_{\substack{U \in U(n) \\ U=[U_1|U_2] \\ U_1 \in \mathbb{C}^{n\times r} \\ U_2 \in \mathbb{C}^{n\times(n-r)}}} \lambda_{2nr-r^2}(Q_{[U_1|U_2]}) \tag{1.8.18}$$

(ix) While (iv) makes clear that a_0 cannot be upper bounded by $\inf_{z\in\mathbb{C}^{n\times r}\setminus\{0\}} \hat{a}(z)$, we can achieve a similar end by constraining z to have orthonormal columns. Namely

$$\frac{1}{4}\inf_{\substack{z\in\mathbb{C}_*^{n\times r} \\ z^*z=\mathbb{I}_{r\times r}}} \hat{a}(z) \leqslant a_0 \leqslant \frac{1}{2}\inf_{\substack{z\in\mathbb{C}_*^{n\times r} \\ z^*z=\mathbb{I}_{r\times r}}} \hat{a}(z) \tag{1.8.19}$$

Proof. See 1.11.2 □

We now move on to analyzing the local lower Lipschitz bounds for the α map $x \mapsto$

$\langle xx^*, A_j\rangle_{\mathbb{R}}^{\frac{1}{2}}$. This was done for the case $r = 1$ in [23]. Recall that $\theta(x) = (xx^*)^{\frac{1}{2}}$ and that

$$A_0 = \inf_{\substack{x,y\in\mathbb{C}^{n\times r}\\ [x]\neq[y]}} \frac{||\alpha(x)-\alpha(y)||_2^2}{||\theta(x)-\theta(y)||_2^2} = \inf_{\substack{x,y\in\mathbb{C}^{n\times r}\\ [x]\neq[y]}} \frac{\sum_{j=1}^m(\langle xx^*, A_j\rangle_{\mathbb{R}}^{\frac{1}{2}} - \langle yy^*, A_j\rangle_{\mathbb{R}}^{\frac{1}{2}})^2}{||(xx^*)^{\frac{1}{2}} - (yy^*)^{\frac{1}{2}}||_2^2} \tag{1.8.20}$$

In analogy with Definition 1.8.1, we consider the local lower Lipschitz bounds for the α map.

Definition 1.8.6. Let $z \in \mathbb{C}^{n\times r}$ have rank k. We define

$$\begin{aligned}
A_1(z) &= \lim_{R\to 0} \inf_{\substack{x\in\mathbb{C}^{n\times r}\\ ||\theta(x)-\theta(z)||_2\leqslant R\\ \text{rank}(x)\leqslant k}} \frac{||\alpha(x)-\alpha(z)||_2^2}{||\theta(x)-\theta(z)||_2^2}\\
A_2(z) &= \lim_{R\to 0} \inf_{\substack{x,y\in\mathbb{C}^{n\times r}\\ ||\theta(x)-\theta(z)||_2\leqslant R\\ ||\theta(y)-\theta(z)||_2\leqslant R\\ \text{rank}(x)\leqslant k\\ \text{rank}(y)\leqslant k}} \frac{||\alpha(x)-\alpha(y)||_2^2}{||\theta(x)-\theta(y)||_2^2}\\
\hat{A}_1(z) &= \lim_{R\to 0} \inf_{\substack{x\in\mathbb{C}^{n\times r}\\ D(x,z)\leqslant R\\ \text{rank}(x)\leqslant k}} \frac{||\alpha(x)-\alpha(z)||_2^2}{D(x,z)^2}\\
\hat{A}_2(z) &= \lim_{R\to 0} \inf_{\substack{x,y\in\mathbb{C}^{n\times r}\\ D(x,z)\leqslant R\\ D(y,z)\leqslant R\\ \text{rank}(x)\leqslant k\\ \text{rank}(y)\leqslant k}} \frac{||\alpha(x)-\alpha(y)||_2^2}{D(x,y)^2}
\end{aligned} \tag{1.8.21}$$

Definition 1.8.7. Given $z \in \mathbb{C}^{n\times r}$ having rank $k > 0$ we define two matrices $\hat{T}_z, \hat{R}_z \in \mathbb{R}^{2nk\times 2nk}$. Let $I_0(z) \subset \{1,\dots,m\}$ be the indices such that $\alpha_j(z) = 0$ (or equivalently such that α_j is not differentiable) for $j \in I_0(z)$, and let $I(z) = \{1,\dots,m\}\backslash I_0(z)$. Once again let $F_j = \mathbb{I}_{k\times k}\otimes j(A_j) \in$

$\mathbb{R}^{2nk\times 2nk}$, then define $\hat{T}_z$ and $\hat{R}_z$ via

$$\hat{T}_z = \sum_{j\in I(z)} \frac{1}{\mu(\hat{z})^T F_j \mu(\hat{z})} F_j \mu(\hat{z})\mu(\hat{z})^T F_j \tag{1.8.22}$$

$$\hat{R}_z = \sum_{j\in I_0(z)} F_j \tag{1.8.23}$$

With these definitions in mind we prove:

Theorem 1.8.8. *Let $z \in \mathbb{C}^{n\times r}$ have rank $k > 0$. Then*

(i) For $r > 1$ it is the case that $\inf_{z\in\mathbb{C}^{n\times r}\setminus\{0\}} A_i(z) = 0$ for $i = 1, 2$, as such $A_0 = 0$.

(ii) Let $\hat{T}_z$ and $\hat{R}_z$ be as in Definition 1.8.7. Then $\hat{A}_1(z)$ and $\hat{A}_2(z)$ are directly computable as

$$\hat{A}_1(z) = \lambda_{2nk-k^2}(\hat{T}_z + \hat{R}_z) \tag{1.8.24}$$

$$\hat{A}_2(z) = \lambda_{2nk-k^2}(\hat{T}_z) \tag{1.8.25}$$

(iii) We have the following inequality between $A_i(z)$ and $\hat{A}_i(z)$ for $i = 1, 2$, which justifies not treating them separately.

$$\hat{A}_i(z) \leqslant A_i(z) \leqslant \sqrt{2}\hat{A}_i(z) \tag{1.8.26}$$

Proof. See 1.11.3 □

For the sake of completeness we also include the following theorem on the global upper Lipschitz bounds for the α and β analysis maps.

Definition 1.8.9. We define the following (squared) upper Lipschitz constants for β and α respectively:

$$b_0 := \sup_{\substack{x,y\in\mathbb{C}^{n\times r}\\ [x]\neq[y]}} \frac{||\beta(x)-\beta(y)||_2^2}{||xx^*-yy^*||_2^2} \tag{1.8.27}$$

$$B_0 := \sup_{\substack{x,y\in\mathbb{C}^{n\times r}\\ [x]\neq[y]}} \frac{||\alpha(x)-\alpha(y)||_2^2}{||(xx^*)^{\frac{1}{2}}-(yy^*)^{\frac{1}{2}}||_2^2} \tag{1.8.28}$$

A somewhat simplifying alternate upper Lipschitz constant for β is

$$b_{0,1} := \sup_{\substack{x,y\in\mathbb{C}^{n\times r}\\ [x]\neq[y]}} \frac{||\beta(x)-\beta(y)||_2^2}{||xx^*-yy^*||_1^2} \tag{1.8.29}$$

Definition 1.8.10. The β map is the pullback of a linear operator acting on symmetric matrices which we refer to as $\mathcal{A}$. Specifically,

$$\begin{aligned} \mathcal{A} &: \text{Sym}(\mathbb{C}^n) \to \mathbb{R}^m \\ \mathcal{A}_j(X) &= \langle X, A_j\rangle_{\mathbb{R}} \end{aligned} \tag{1.8.30}$$

Definition 1.8.11. When $A_j \geqslant 0$ for each j, we define the operator T_r.

$$\begin{aligned} T_r &: \mathbb{C}^{n\times r} \to (\mathbb{C}^{n\times r})^m \\ T_r(x) &= (A_j^{\frac{1}{2}}x)_{j=1}^m \end{aligned} \tag{1.8.31}$$

In a slight abuse of notation we write for $r = 1$

$$\begin{aligned} T_1 &: \mathbb{C}^n \to \mathbb{C}^{n\times m} \\ T_1(x) &= [A_1^{\frac{1}{2}}x|\cdots|A_m^{\frac{1}{2}}x] \end{aligned} \tag{1.8.32}$$

We compute explicitly b_0, $b_{0,1}$, and B_0 via different norms of the operators $\mathcal{A}$ and T_r, as well as providing formulas for b_0 and B_0 analogous to (1.8.18) and (1.8.25). Specifically, we prove:

Theorem 1.8.12. *Let b_0, $b_{0,1}$, B_0, $\mathcal{A}$, and T_r be as above. Then*

(i) The global upper bound b_0 is given by

$$b_0 = \max_{\substack{U\in U(n)\\ U=[U_1|U_2]\\ U_1\in\mathbb{C}^{n\times r},U_2\in\mathbb{C}^{n\times n-r}}} \lambda_1(Q_{[U_1|U_2]}) \tag{1.8.33}$$

Where Q_U is as in Definition 1.8.3.

(ii) The global upper bound $b_{0,1}$ is given by

$$b_{0,1} = ||\mathcal{A}||^2_{1\to 2} \tag{1.8.34}$$

Additionally if $A_j \geqslant 0$ for all j then

$$b_{0,1} = ||T_r||^4_{2\to(2,4)} = ||T_1||^4_{2\to(2,4)} \tag{1.8.35}$$

Where the $||\cdot||_{2,4}$ norm of a matrix is the l^4 norm of the vector of l^2 norms of its columns.

(iii) The global upper bound B_0 is given by

$$B_0 = \sup_{\substack{z \in \mathbb{C}^{n\times r} \\ z \neq 0}} \lambda_1(\hat{T}_z) = B \tag{1.8.36}$$

Where $\hat{T}_z$ is as in Definition 1.8.7 and B is the optimal upper frame bound for $\{A_j\}_{j=1}^m$.

Proof. See 1.11.4. □

It turns out that Theorem 1.8.5 allows us to find novel algebraic conditions for a frame for $\mathbb{C}^{n\times r}$ to be generalized phase retrievable, generalizing Theorem 4 in [11]. The benefit of condition (vi) over the definition of phase retrievability is that they involve checking a quantity over all $n \times r$ matrices with orthonormal columns, that is to say over the Stiefel manifold of dimension $2nr - r^2$, as opposed to over all pairs of $n \times r$ matrices.

Theorem 1.8.13. *Let $\{A_j\}_{j=1}^m$ be a frame for $\mathbb{C}^{n\times r}$. Then the following are equivalent:*

(i) $\{A_j\}_{j=1}^m$ is generalized phase retrievable.

(ii) For all $U_1 \in \mathbb{C}^{n\times r}$, $U_2 \in \mathbb{C}^{n\times(n-r)}$ such that $[U_1|U_2] \in U(n)$ the $2nr - r^2 \times 2nr - r^2$ matrix

$$Q_{[U_1|U_2]} = \sum_{j=1}^m \begin{bmatrix} \tau(U_1^* A_j U_1) \\ \mu(U_2^* A_j U_1) \end{bmatrix} \begin{bmatrix} \tau(U_1^* A_j U_1) \\ \mu(U_2^* A_j U_1) \end{bmatrix}^T \tag{1.8.37}$$

is invertible.

(iii) For all $z \in \mathbb{C}^{n\times r}$ such that z has orthonormal columns, the $2nr \times 2nr$ matrix

$$\hat{Q}_z = 4\sum_{j=1}^m (\mathbb{I}_{k\times k} \otimes j(A_j))\mu(z)\mu(z)^T(\mathbb{I}_{k\times k} \otimes j(A_j)) \tag{1.8.38}$$

has as its null space precisely the r^2 dimensional $\mathcal{V}_z = \{\mu(u)|u \in V_{\pi,z}(\mathbb{C}_*^{n\times r})\}$.

(iv) For all $U_1 \in \mathbb{C}^{n\times r}$, $U_2 \in \mathbb{C}^{n\times(n-r)}$ such that $[U_1|U_2] \in U(n)$, $H \in Sym(\mathbb{C}^r)$, $B \in \mathbb{C}^{(n-r)\times r}$ there exist $c_1, \dots c_m \in \mathbb{R}$ such that

$$U_1^*(\sum_{j=1}^m c_j A_j)U_1 = H \tag{1.8.39a}$$

$$U_2^*(\sum_{j=1}^m c_j A_j)U_1 = B \tag{1.8.39b}$$

(v) For all $U_1 \in \mathbb{C}^{n\times r}$ with orthonormal columns

$$span_{\mathbb{R}}\{A_j U_1\}_{j=1}^m = \{U_1 K|K \in \mathbb{C}^{r\times r}, K^* = -K\}^{\perp} \tag{1.8.40}$$

(vi) For all $U_1 \in \mathbb{C}^{n\times r}$ with orthonormal columns

$$dim_{\mathbb{R}}\{A_j U_1\}_{j=1}^m \geqslant 2nr - r^2 \tag{1.8.41}$$

Proof. See 1.11.5 □

1.9 Proofs for Section 1.6

1.9.1 Proof of Proposition 1

Proof. Both $d(x, y)$ and $D(x, y)$ are obviously positive and symmetry follows from the fact that that $U(r)$ is a group. Moreover, owing to the compactness of $U(r)$, both $D(x, y)$ and $d(x, y)$ are

zero if and only if there exists U_0 such that $x = yU_0$, that is if and only if $[x] = [y]$. It remains to prove the triangle inequality. For $D(x, y)$ the computation is straightforward and follows from the unitary invariance of the Frobenius norm. If U_1 and U_2 are unitary minimizers for $D(x, z)$ and $D(z, y)$ respectively then

$$\begin{aligned} D(x,z) + D(y,z) &= ||x - zU_1||_2 + ||z - yU_2||_2 \\ &= ||x - zU_1||_2 + ||zU_1 - yU_2U_1||_2 \\ &\geqslant ||x - yU_2U_1||_2 \geqslant D(x,y) \end{aligned} \tag{1.9.1}$$

We note that the above argument also holds for any unitarily invariant norm $|||\cdot|||$ so that each $D_{|||\cdot|||}(x, y) := \min_{U\in U(r)} |||x - yU|||$ is a metric on $\mathbb{C}^{n\times r}/U(r)$. A similar trick can be employed regarding $d(x, y)$, but it requires the following lemma which does not readily generalize to arbitrary unitarily invariant norms or even $p \neq 2$:

Lemma 1.9.1. *The following triangle inequality holds for all* $x, y, z \in \mathbb{C}^{n\times r}$

$$||x - y||_2||x + y||_2 \leqslant ||x - z||_2||x + z||_2 + ||z - y||_2||z + y||_2 \tag{1.9.2}$$

Proof. This is essentially a statement about the geometry of parallelepipeds in $\mathbb{R}^3$, namely that the sum of the product of face diagonals from any two sides sharing a vertex will always exceed the product of the two on the remaining side sharing the vertex. The lemma follows from the

observation that for $x, y \in \mathbb{R}^n$

$$\begin{aligned}
||x-y||_2||x+y||_2 &= \sqrt{(||x||_2^2+||y||_2^2)^2 - 4|\langle x,y\rangle_{\mathbb{R}}|^2} \\
&= \frac{1}{2}\left(||x||_2^2 - ||y||_2^2 + \sqrt{(||x||_2^2+||y||_2^2)^2 - 4|\langle x,y\rangle_{\mathbb{R}}|^2}\right) \\
&\quad - \frac{1}{2}\left(||x||_2^2 - ||y||_2^2 - \sqrt{(||x||_2^2+||y||_2^2)^2 - 4|\langle x,y\rangle_{\mathbb{R}}|^2}\right) \\
&= \lambda_+(xx^T - yy^T) - \lambda_-(xx^T - yy^T) \\
&= ||xx^T - yy^T||_1
\end{aligned} \tag{1.9.3}$$

See the proof of Theorem 1.6.4 for a direct computation of the eigenvalues of $xx^T - yy^T$ (the theorem deals with the complex case but the real case is identical). This identity proves the lemma immediately since the latter obeys the triangle inequality and

$$\begin{aligned}
||x-y||_2||x+y||_2 &= ||\mu(x)-\mu(y)||_2||\mu(x)+\mu(y)||_2 \\
&= ||\mu(x)\mu(x)^T - \mu(y)\mu(y)^T||_1 \\
&\leqslant ||\mu(x)\mu(x)^T - \mu(z)\mu(z)^T||_1 + ||\mu(z)\mu(z)^T - \mu(y)\mu(y)^T||_1 \\
&= ||x-z||_2||x+z||_2 + ||z-y||_2||z+y||_2
\end{aligned} \tag{1.9.4}$$

Where $\mu : \mathbb{C}^{n\times r} \to \mathbb{R}^{2nr}$ is complex matrix vectorization. □

The proposition then follows via a similar argument to (1.9.1), namely if U_1, U_2 are the

minimizers in $d(x,z)$ and $d(z,y)$ respectively then

$$\begin{aligned} d(x,z)+d(z,y) &= ||x-zU_1||_2||x+zU_1||_2 + ||z-yU_2||_2||z+yU_2||_2 \\ &= ||x-zU_1||_2||x+zU_1||_2 + ||zU_1-yU_2U_1||_2||zU_1+yU_2U_1||_2 \\ &\geqslant ||x-yU_2U_1||_2||x+yU_2U_1||_2 \geqslant d(x,y) \end{aligned} \tag{1.9.5}$$

□

1.9.2 Proof of Proposition 2

Proof. Both the trace $\mathrm{tr}\{x^*yU\}$ in that appears in D and its square as it appears in d will be maximized when x^*yU is positive semidefinite, thus we may take the minimizer to be the polar factor for x^*y, the polar factor of course being the unique unitary for which x^*yU is non-negative only when x^*y is full rank. The non-uniqueness of the minimizer arises precisely from the non-uniqueness in choice of polar factor when x^*y does not have full rank. Note that even if y is full rank, x^*y will have rank less than r whenever $\mathrm{Ran}(y) \cap \mathrm{Ran}(x)^\perp \neq 0$. □

1.9.3 Proof of Proposition 3

Proof. Note that the non-zero eigenvalues of $\pi(x)$ are precisely the squares of the singular values of x, the non-zero eigenvalues of $\theta(x)$ agree with the non-zero singular values of x, and the non-zero eigenvalues values of $\psi(x)$ differ from the non-zero singular values of x only by a factor of $||x||_2$. This proves that the embeddings preserve rank. It is readily checked that the embeddings

are surjective and injective modulo $\sim$. In particular for $A \in S^{r,0}(\mathbb{C}^n)$, we have

$$\pi^{-1}(A) = [\text{Cholesky}(A)] \tag{1.9.6}$$

$$\theta^{-1}(A) = [\text{Cholesky}(A^2)] \tag{1.9.7}$$

$$\psi^{-1}(A) = [\text{Cholesky}(A^2/||A||_2)] \tag{1.9.8}$$

where Cholesky(A) is a Cholesky decomposition of A in $\mathbb{C}^{n\times r}$ (note that the Cholesky decomposition is unique up to equivalence class). □

1.9.4 Proof of Theorem 1.6.4

Proof. To prove (1.6.5) we analyze the following quantity:

$$Q(x,y) = \frac{D(x,y)^2}{||\theta(x) - \theta(y)||_2^2} = \frac{||x||_2^2 + ||y||_2^2 - 2||x^*y||_1}{||x||_2^2 + ||y||_2^2 - 2\text{tr}\{(xx^*)^{\frac{1}{2}}(yy^*)^{\frac{1}{2}}\}} \tag{1.9.9}$$

We first note that $||x^*y||_1 = ||(xx^*)^{\frac{1}{2}}(yy^*)^{\frac{1}{2}}||_1$ since $(xx^*)^{\frac{1}{2}}(yy^*)^{\frac{1}{2}}$ and x^*y have the same non-zero singular values. Hence if we define $A = \theta(x) = (xx^*)^{\frac{1}{2}}$ and $B = \theta(y) = (yy^*)^{\frac{1}{2}}$ we can abuse notation slightly and write

$$Q(A,B) = \frac{||A||_2^2 + ||B||_2^2 - 2||AB||_1}{||A||_2^2 + ||B||_2^2 - 2\text{tr}\{AB\}} \tag{1.9.10}$$

Now $\text{tr}\{AB\} \leqslant ||AB||_1$, so we conclude that $Q(x,y) \leqslant 1$. On the other hand this bound is achievable by any x and y for having the same left singular vectors, since in this case A and B commute hence $AB \geqslant 0$ and $||AB||_1 = \text{tr}\{AB\}$. We conclude that the upper Lipschitz constant

is 1, and in particular

$$\sup_{\substack{x,y\in\mathbb{C}^{n\times r}/U(r)\\ x\neq y}} Q(x,y) = \max_{\substack{x,y\in\mathbb{C}^{n\times r}/U(r)\\ x\neq y}} Q(x,y) = 1 \tag{1.9.11}$$

We now turn our attention to the lower bound. It is shown in [33] that for any unitarily invariant norm $|||\cdot|||$ and positive semidefinite matrices A and B the following generalization of the arithmetic-geometric mean inequality holds:

$$4|||AB|||^2 \leqslant |||(A+B)^2||| \tag{1.9.12}$$

We apply this inequality to the nuclear norm and conclude that

$$\begin{aligned} 4||AB||_1 &\leqslant ||(A+B)^2||_1 \\ &= \text{tr}\{(A+B)^2\} \\ &= ||A||_2^2 + ||B||_2^2 + 2\text{tr}\{AB\} \end{aligned} \tag{1.9.13}$$

We employ this fact in the analysis of $Q(x,y)$:

$$\begin{aligned} Q(A,B) &= \frac{1}{2}\cdot\frac{2||A||_2^2 + 2||B||_2^2 - 4||AB||_1}{||A||_2^2 + ||B||_2^2 - 2\text{tr}\{AB\}} \\ &\geqslant \frac{1}{2}\cdot\frac{2||A||_2^2 + 2||B||_2^2 - (||A||_2^2 + ||B||_2^2 + 2\text{tr}\{AB\})}{||A||_2^2 + ||B||_2^2 - 2\text{tr}\{AB\}} = \frac{1}{2} \end{aligned} \tag{1.9.14}$$

This implies a lower Lipschitz constant of at least $\frac{1}{\sqrt{2}}$. For the trivial case $n = r = 1$ the ratio is 1. To prove the constant of $\frac{1}{\sqrt{2}}$ is optimal for $n > 1$, let e_1 and e_2 be any two orthogonal unit vectors in $\mathbb{C}^n$ and let $x = e_1$ and $(y_j)_{j\geqslant 1}$ be given by $y_j = \sqrt{1-\frac{1}{j^2}}e_1 + \frac{1}{j}e_2$. Define $A = \theta(x)$

and $B_j = \theta(y_j)$, then both A and each B_j have unit norm and are rank 1 hence are idempotent, so that

$$\begin{aligned} AB_j &= (xx)^{\frac{1}{2}}(y_j y_j^*)^{\frac{1}{2}} = xx^* y_j y_j^* \\ &= \langle x, y_j \rangle_{\mathbb{R}} x y_j^* \\ &= (1 - \frac{1}{j^2}) e_1 e_1^* + \frac{\sqrt{1 - \frac{1}{j^2}}}{j} e_1 e_2^* \end{aligned} \tag{1.9.15}$$

Thus $\text{tr}\{AB_j\} = 1 - \frac{1}{j^2}$. On the other hand, $||AB_j||_1 = ||x^* y_j||_1 = |\langle x, y_j \rangle_{\mathbb{R}}| = \sqrt{1 - \frac{1}{j^2}}$. We find

$$\begin{aligned} \lim_{j \to \infty} Q(A, B_j) &= \lim_{j \to \infty} \frac{1 - ||AB_j||_1}{1 - \text{tr}\{AB_j\}} \\ &= \lim_{j \to \infty} j^2 (1 - \sqrt{1 - \frac{1}{j^2}}) = \frac{1}{2} \end{aligned} \tag{1.9.16}$$

Thus we conclude

$$\inf_{\substack{x, y \in \mathbb{C}^{n \times r} \\ x \neq y}} Q(x, y) = \frac{1}{2} \tag{1.9.17}$$

We now concern ourselves with proving (1.6.6). To prove the lower bound, let U_0 be the minimizer in $d(x, y)$. Then

$$\begin{aligned} ||\pi(x) - \pi(y)||_1 &= ||xx^* - yy^*||_1 \\ &= ||\frac{1}{2}(x - yU_0)(x + yU_0)^* + \frac{1}{2}(x + yU_0)(x - yU_0)^*||_2 \\ &\leqslant \frac{1}{2}||(x - yU_0)(x + yU_0)^*||_1 + \frac{1}{2}||(x - yU_0)(x + yU_0)^*||_1 \\ &\leqslant ||x - yU_0||_2 ||x + yU_0||_2 = d(x, y) \end{aligned} \tag{1.9.18}$$

This implies a lower Lipschitz constant of at least 1, but in fact this constant is optimal since the two are equal for $r = 1$. Turning our attention to the upper bound, we will in fact prove the following stronger inequality:

$$||\psi(x) - \psi(y)||_2 \geqslant \frac{1}{4}d(x,y)^2 + \frac{1}{4}D(x,y)^4 + (||x||_2 - ||y||_2)^2\left(||x^*y||_1 + \frac{1}{2}(||x||_2 + ||y||_2)^2\right) \tag{1.9.19}$$

We prove (1.9.19) by direct computation:

$$\begin{aligned}
&||\psi(x) - \psi(y)||_2^2 - \frac{1}{4}d(x,y)^2 \\
&\qquad = ||x||_2^4 + ||y||_2^4 - 2||x||_2||y||_2\text{tr}\{(xx^*)^{\frac{1}{2}}(yy^*)^{\frac{1}{2}}\} - \frac{1}{4}\left((||x||_2^2 + ||y||_2^2)^2 - 4||x^*y||_1^2\right) \\
&\qquad = \frac{3}{4}||x||_2^4 + \frac{3}{4}||y||_2^4 + ||x^*y||_1^2 - \frac{1}{2}||x||_2^2||y||_2^2 - 2||x||_2||y||_2\text{tr}\{(xx^*)^{\frac{1}{2}}(yy^*)^{\frac{1}{2}}\} \\
&\qquad \geqslant \frac{3}{4}||x||_2^4 + \frac{3}{4}||y||_2^4 + ||x^*y||_1^2 - \frac{1}{2}||x||_2^2||y||_2^2 - 2||x||_2||y||_2||(xx^*)^{\frac{1}{2}}(yy^*)^{\frac{1}{2}}||_1 \\
&\qquad = \frac{1}{4}(||x||_2^2 - ||y||_2^2)^2 + \frac{1}{2}||x||_2^4 + \frac{1}{2}||y||_2^4 + ||x^*y||_1^2 - 2||x||_2||y||_2||x^*y||_1
\end{aligned} \tag{1.9.20}$$

We then note that

$$\begin{aligned}
\frac{1}{4}D(x,y)^4 &= \frac{1}{4}(||x||^2 + ||y||^2 - 2||x^*y||_1)^2 \\
&= \frac{1}{4}||x||_2^4 + \frac{1}{4}||y||_2^4 + \frac{1}{2}||x||_2^2||y||_2^2 + ||x^*y||_1^2 - (||x||_2^2 + ||y||_2^2)||x^*y||_1
\end{aligned} \tag{1.9.21}$$

So that if we add and subtract $\frac{1}{4}D(x,y)^4$ from (1.9.20) we obtain the result

$$\begin{aligned} &||\psi(x)-\psi(y)||_2^2 - \frac{1}{4}d(x,y)^2 \\ &\qquad \geqslant \frac{1}{2}(||x||_2^2 - ||y||_2^2)^2 + \frac{1}{4}D(x,y)^4 + (||x||_2 - ||y||_2)^2||x^*y||_1 \\ &\qquad = \frac{1}{4}D(x,y)^4 + (||x||_2 - ||y||_2)^2\left((||x^*y||_1 + \frac{1}{2}(||x||_2 + ||y||_2)^2\right) \end{aligned} \tag{1.9.22}$$

This immediately proves that $2||\psi(x)-\psi(y)||_2 \geqslant d(x,y)$ and hence that the upper Lipschitz constant in (1.6.6) is at most 2. For $r = 1$, we will prove shortly claim (iii), implying that $d(x,y) = ||\pi(x)-\pi(y)||_1 = ||\psi(x)-\psi(y)||_1$, hence in this case the optimal constant is $\sqrt{2}$, owing to the fact that $\psi(x)-\psi(y)$ will have rank at most 2 and in that case $d(x,y) = ||\psi(x)-\psi(y)||_1 \leqslant \sqrt{2}||\psi(x)-\psi(y)||_2$. For $r > 1$, however, we show that the upper Lipschitz constant of 2 is optimal by considering a sequence of matrices in $\mathbb{C}^{n\times 2}$. As before let e_1 and e_2 be any unit orthonormal vectors in $\mathbb{C}^n$. Let $x = [e_1|0]$, $(y_j)_{j\geqslant 1}$ be given by $y_j = [\sqrt{1-\frac{1}{j^2}}e_1|\frac{1}{j}e_2]$. As before let $A = \theta(x)$, $B_n = \theta(y_j)$. We first note that A and each B_j commute and are positive semidefinite, so that AB_j is also positive semidefinite and we have $\text{tr}\{AB_j\} = ||AB_j||_1$ and the inequality in (1.9.20) is actually an equality. This makes clear the impediment to a rank 1 sequence achieving the upper Lipschitz constant of 2: A and B_j could not be made to commute without x and y_j lying in the same equivalence class. Finally, we observe that $||x||_2 = ||y_j||_2 = 1$ so the remainder term in (1.9.19) disappears and we obtain

$$||\psi(x)-\psi(y_j)||_2^2 = \frac{1}{4}d(x,y)^2 + \frac{1}{4}D(x,y)^4 \tag{1.9.23}$$

We note moreover that $d(x,y)^2 = D(x,y)^2(||x||_2^2 + ||y||_2^2 + 2||x^*y||_1)$ so that

$$\begin{aligned} \frac{||\psi(x) - \psi(y_j)||_2^2}{d(x,y_j)^2} &= \frac{1}{4}\left(1 + \frac{D(x,y_j)^4}{d(x,y_j)^2}\right) \\ &= \frac{1}{4}\left(1 + \frac{1 - ||x^*y_j||_1}{1 + ||x^*y_j||_1}\right) \end{aligned} \tag{1.9.24}$$

Now $||x^*y_j||_1 = ||\begin{bmatrix} e_1^* \\ \hline 0 \end{bmatrix}\begin{bmatrix} \sqrt{1-\frac{1}{j^2}} & 0 \\ 0 & \frac{1}{j} \end{bmatrix}\Big[e_1|e_2\Big]||_1 = \sqrt{1-\frac{1}{j^2}}$ so that

$$\lim_{j\to\infty} \frac{||\psi(x) - \psi(y_j)||_2^2}{d(x,y_j)^2} = \lim_{j\to\infty} \frac{1}{4}\left(1 + \frac{1 - \sqrt{1-\frac{1}{j^2}}}{1 + \sqrt{1+\frac{1}{j^2}}}\right) = \frac{1}{4} \tag{1.9.25}$$

Thus we have proven claims (i) and (ii). To prove the first claim of (iii) note that for $r = 1$, $(xx^*)^{\frac{1}{2}} = \frac{xx^*}{||x||_2}$. The second part of (iii) follows from direct computation of $||xx^* - yy^*||_1$ via the method of moments. Clearly $xx^* - yy^*$ will have one positive and one negative eigenvalue, which we denote λ_+ and λ_-. In this case

$$\begin{aligned} \lambda_+ + \lambda_- &= \text{tr}\{xx^* - yy^*\} \\ &= ||x||_2^2 - ||y||_2^2 \\ \lambda_+\lambda_- &= \frac{1}{2}\left(\text{tr}\{xx^* - yy^*\}^2 - \text{tr}\{(xx^* - yy^*)^2\}\right) \\ &= ||x||^2||y||^2 - |\langle x, y\rangle_{\mathbb{R}}|^2 \end{aligned} \tag{1.9.26}$$

A little bit of algebra then yields

$$\lambda_\pm = \frac{1}{2}\left(||x||_2^2 - ||y||_2^2 \pm \sqrt{(||x||^2 + ||y||^2)^2 - 4|\langle x, y\rangle_{\mathbb{R}}|^2}\right) \tag{1.9.27}$$

Thus we find $||xx^* - yy^*||_1 = \lambda_+ - \lambda_- = \sqrt{(||x||^2 + ||y||^2)^2 - 4|\langle x, y\rangle_{\mathbb{R}}|^2} = d(x, y)$. It strikes the authors that this is a minor miracle. Finally, to prove claim (iv) consider x and y having a common basis of singular vectors with singular values $(\sigma_i)_{i=1}^r$ and $(\mu_i)_{i=1}^r$ respectively. Then

$$||\pi(x) - \pi(y)||_2^2 = \sum_{i=1}^{r} (\sigma_i^2 - \mu_i^2)^2 \tag{1.9.28}$$

$$d(x, y)^2 = \sum_{i,j=1}^{r} (\sigma_i + \mu_i)^2 (\sigma_j - \mu_j)^2 \tag{1.9.29}$$

The latter is obviously larger, consistent with (1.6.6). If it were additionally the case that $d(x, y) \leqslant C||\pi(x) - \pi(y)||_2$ we would have

$$\sum_{i \neq j} (\sigma_i + \mu_i)^2 (\sigma_j - \mu_j)^2 \leqslant (C - 1) \sum_{i=1}^{r} (\sigma_i^2 - \mu_i^2)^2 \tag{1.9.30}$$

In the case $r = 1$ the left hand side is zero and so we may take $C = 1$. For $r > 1$, in contradiction of the above take $\sigma_1 = \mu_1 = \delta$, $\sigma_2 \neq \mu_2$ and all other singular values zero. We then would obtain

$$4\delta^2 (\sigma_2 - \mu_2)^2 \leqslant (C - 1)(\sigma_2^2 - \mu_2^2)^2 \tag{1.9.31}$$

There is evidently no such C since δ may be chosen arbitrarily large. Thus claim (v) is proved, justifying the use of the alternate embedding ψ in (1.6.6). This concludes the proof of Theorem 1.6.4. □

1.10 Proofs for Section 1.7

1.10.1 Proof of Proposition 4

Proof. The proof of (1.7.5) is by direct computation. Namely

$$V_{\pi,x}(\mathbb{C}_*^{n\times r}) = \ker D\pi(x) = \{w \in \mathbb{C}^{n\times r} | xw^* + wx^* = 0\} \tag{1.10.1}$$

We would like to obtain a direct parametrization, however, and note that

$$\begin{aligned} w \in V_{\pi,x}(\mathbb{C}_*^{n\times r}) &\iff wx^* = \tilde{K} && \tilde{K} \in \mathbb{C}^{n\times n}, \tilde{K}^* = -\tilde{K}, \mathbb{P}_{\mathrm{Ran}(x)}\tilde{K} = \tilde{K} \\ &\iff wx^* = xKx^* && K \in \mathbb{C}^{r\times r}, K^* = -K \\ &\iff w = xK && K \in \mathbb{C}^{r\times r}, K^* = -K \end{aligned} \tag{1.10.2}$$

In the first line note that w is recoverable from such a $\tilde{K}$ via $w = \tilde{K}x(x^*x)^{-1}$. In the second note that $K = (xx^*)^\dagger x^* \tilde{K} x (xx^*)^\dagger$. The third "if and only if" is obtained by right multiplying

$x(x^*x)^{-1}$. The horizontal space is then computable as $V_{\pi,x}(\mathbb{C}_*^{n\times r})^\perp$:

$$\begin{aligned}
w \in H_{\pi,x}(\mathbb{C}_*^{n\times r}) &\iff \Re\mathrm{tr}\{w^*xK\} = 0 \quad \forall K \in \mathbb{C}^{n\times n}, K^* = -K \\
&\iff x^*w = \tilde{H} \qquad \tilde{H} \in \mathbb{C}^{r\times r}, \tilde{H}^* = \tilde{H} \\
&\iff x^*w = x^*Hx \qquad H \in \mathbb{C}^{n\times n}, H^* = H, \mathbb{P}_{\mathrm{Ran}(x)}H = H \\
&\iff \mathbb{P}_{\mathrm{Ran}(x)}w = Hx \quad H \in \mathbb{C}^{n\times n}, H^* = H, \mathbb{P}_{\mathrm{Ran}(x)}H = H \\
&\iff w = Hx + X \qquad H \in \mathbb{C}^{n\times n}, H^* = H = \mathbb{P}_{\mathrm{Ran}(x)}H, X \in \mathbb{C}^{n\times r}, \mathbb{P}_{\mathrm{Ran}(x)}X = 0
\end{aligned} \tag{1.10.3}$$

The second line follows from the fact that $\mathbb{C}^{n\times n}$ decomposes orthogonally into Hermitian and skew-Hermitian matrices. In the second note that $H = (x^*x)^{-1}x\tilde{H}x^*(x^*x)^{-1}$. The third follows from left multiplying by $(xx^*)^\dagger x$. Finally, the tangent space can be parametrized via the horizontal space as its image through $D\pi(x)$ as

$$\begin{aligned}
T_{\pi(x)}(\mathring{S}^{r,0}(\mathbb{C}^n)) &= D\pi(x)(H_{\pi,x}(\mathbb{C}_*^{n\times r})) \\
&= \{Hxx^* + xx^*H + xX^* + Xx^* | H \in \mathbb{C}^{n\times n}, H^* = H, \mathbb{P}_{\mathrm{Ran}(x)}H = H, \mathbb{P}_{\mathrm{Ran}(x)}X = 0\}
\end{aligned} \tag{1.10.4}$$

This provides a direct parametrization, but for our purposes the simpler indirect description given by (1.7.7) will be more useful. It is clear from (1.10.4) that $T_{\pi(x)}(\mathring{S}^{r,0}(\mathbb{C}^n)) \subset \{W \in \mathrm{Sym}(\mathbb{C}^n) | \mathbb{P}_{\mathrm{Ran}(x)^\perp}W\mathbb{P}_{\mathrm{Ran}(x)^\perp} = 0\}$. To prove the reverse, note that if $W \in \mathrm{Sym}(\mathbb{C}^n)$ and $\mathbb{P}_{\mathrm{Ran}(x)^\perp}W\mathbb{P}_{\mathrm{Ran}(x)^\perp}$ then $W = W_1 + W_2 + W_2^*$ where $\mathbb{P}_{\mathrm{Ran}(x)}W_1\mathbb{P}_{\mathrm{Ran}(x)} = W_1$ and $\mathbb{P}_{\mathrm{Ran}(x)}W_2\mathbb{P}_{\mathrm{Ran}(x)^\perp} = W_2$. Any such W_2 is representable as xX^* where X is as in the description of the horizontal

space. Indeed, take $X = W_2^* x (x^* x)^{-1}$. Finally, the Sylvester equation $xx^* H + Hxx^* = W_1$ has the unique solution

$$H = \int_0^\infty e^{-txx^*} W_1 e^{-txx^*} dt \tag{1.10.5}$$

□

1.10.2 Proof of Theorem 1.7.4

Proof. To prove (i) in relatively short order we employ the following theorem:

Theorem 1.10.1 (see [34] and [35] Appendix B)**.** *Let $\phi : G \times M \to M$ be a smooth action of a Lie group G on a smooth manifold M. If the action is semi-algebraic, then orbits of ϕ are smooth submanifolds of M.*

We apply this theorem in the case of $\mathring{S}^{p,q}(\mathbb{C}^n)$. Sylvester's Inertia Theorem says that $A \in \mathring{S}^{p,q}(\mathbb{C}^n)$ if and only if $A = K I_{p,q} K^*$ for some $K \in \mathrm{GL}(\mathbb{C}^n)$ where $I_{p,q} = \mathrm{diag}(1, \ldots, 1, -1, \ldots, -1, 0, \ldots, 0)$ is the matrix of inertia indices. Thus $\mathring{S}^{p,q}(\mathbb{C}^n)$ is precisely the orbit of $I_{p,q}$ under the smooth Lie group action:

$$\begin{aligned} \psi : \mathrm{GL}(\mathbb{C}^n) \times \mathbb{C}^{n \times n} &\to \mathbb{C}^{n \times n} \\ \psi(K, L) &= KLK^* \end{aligned} \tag{1.10.6}$$

Noting that $\psi(KJ, L) = \psi(K, \psi(J, L))$ for $K, J \in \mathrm{GL}(\mathbb{C}^n)$. We need to check that the action is

semi-algebraic. For a fixed $L \in \mathbb{C}^{n\times n}$ the action has as its graph

$$\begin{aligned}&\left\{(K,Y)\middle| K \in \mathrm{GL}(\mathbb{C}^n), Y = KLK^*\right\}\\ =&\left\{(k_{ij},y_{ij})\middle| i,j \in 1,\dots,n, \mathrm{Det}(k_{ij}) \neq 0, y_{ij} - Q_{ij}(k_{ij}) = 0\right\}\end{aligned} \tag{1.10.7}$$

where each Q_{ij} is a quadratic polynomial in $(k_{ij})_{i,j=1}^n$ determined by L. This set is manifestly semi-algebraic, so by Theorem 1.10.1 each $\mathring{S}^{p,q}(\mathbb{C}^n)$ is a smooth submanifold of $\mathbb{C}^{n\times n}$. To prove that the dimension of $\mathring{S}^{p,q}(\mathbb{C}^n)$ is given by $2n(p+q) - (p+q)^2$ note that the $\dim \mathring{S}^{p,q}(\mathbb{C}^n) = \dim \mathring{S}^{p+q,0}$ since matrix absolute value

$$\begin{aligned}&|\cdot| : \mathring{S}^{p,q}(\mathbb{C}^n) \to \mathring{S}^{p+q,0}\\ &|A| = (AA^*)^{\frac{1}{2}}\end{aligned} \tag{1.10.8}$$

is surjective and injective of up to permutation of eigenvalues. The dimension of $\mathring{S}^{p+q,0}$ can be computed from $T_{\pi(x)}(\mathring{S}^{r,0}(\mathbb{C}^n))$ as found in Lemma 4. Taking $r = p+q$ then

$$\dim T_{\pi(x)}(\mathring{S}^{r,0}(\mathbb{C}^n)) = n^2 - (n-r)^2 = 2nr - r^2 = 2n(p+q) - (p+q)^2 \tag{1.10.9}$$

It remains to prove analyticity of $\mathring{S}^{r,0}(\mathbb{C}^n)$. It is proved in Lemma 3.11 of [36] that $\mathring{S}^{1,0}(\mathbb{C}^n)$ is real analytic. The proof in the general case is analagous. First note that owing to Sylvester's inertia theorem $\mathrm{GL}(\mathbb{C}^n)$ acts transitively on $\mathring{S}^{p,q}(\mathbb{C}^n)$ via conjugation, since if $X, Y \in \mathring{S}^{p,q}(\mathbb{C}^n)$ then we may obtain $G_1, G_2 \in \mathrm{GL}(\mathbb{C}^n)$ so that $G_1 X G_1^* = I_{p,q} = G_2 Y G_2^*$, hence $(G_2^{-1}G_1)X(G_2^{-1}G_1)^* = Y$. It remains to obtain that the stabilizer group is closed in $\mathrm{GL}(\mathbb{C}^n)$ so that we can invoke the homogeneous space construction theorem. If $Z \in \mathring{S}^{p,q}(\mathbb{C}^n)$ then $Z = zI_{p,q}z^*$ for some

$z = U_z \left[\begin{array}{c} \Lambda_z \\ \hline 0 \end{array}\right] V_z^* \in \mathbb{C}_*^{n\times r}$. The stabilizer group at Z is given by $T \in \mathrm{GL}(\mathbb{C}^n)$ such that $Tz \in \{zU | U \in U(p,q)\}$. In a basis $e_1, \dots e_n$ for $\mathbb{C}^n$ where $e_1, \dots e_r$ span $\mathrm{Ran}(z)$ and $e_{r+1}, \dots, e_n$ span $\mathrm{Ran}(z)^\perp$ the stabilizer is therefore given by

$$\mathbb{H}_Z^{r,0} = \left\{ \left[\begin{array}{c|c} \Lambda_z U \Lambda_z^{-1} & M_1 \\ \hline 0 & M_2 \end{array}\right] \middle| \quad U \in U(p,q), M_1 \in \mathbb{C}^{r\times n-r}, M_2 \in \mathbb{C}^{r\times r}, \det(M_2) \neq 0 \right\} \quad (1.10.10)$$

It is easy to see that $\mathbb{H}_Z^{r,0}$ is a (relatively) closed subset of $\mathrm{GL}(\mathbb{C}^n)$, hence by the homogeneous space construction theorem $\mathring{S}^{r,0}(\mathbb{C}^n)$ is diffeomorphic to the analytic manifold $\mathrm{GL}(\mathbb{C}^n)/\mathbb{H}_Z^{r,0}$. This concludes the proof of (i). Claims (ii) and (iii) represent slight generalizations over the analogous results in [22] for positive definite matrices, but the same key theorems apply. Namely, we employ the following:

Theorem 1.10.2 (see [37] Proposition 2.28)**.** *Let (M, g) be a Riemannian manifold and let G be a compact Lie group of isometries acting freely on M. Then let $N = M/G$ and $\pi : M \to N$ be the quotient map. Then there exists a unique Riemannian metric h on N so that $\pi : (M, g) \to (N, h)$ is a Riemannian submersion; and in particular that $D\pi(z) : H_{\pi,z} \to T_{\pi(z)}(N)$ is isometric for each $z \in M$.*

Theorem 1.10.3 (see [37] Proposition 2.109)**.** *If $\pi : (M, g) \to (N, h)$ is a Riemannian submersion and γ is a geodesic in (M, g) such that $\dot\gamma(0)$ is horizontal (i.e. $\dot\gamma(0) \in H_{\pi,\gamma(0)}$) then*

(i) $\dot\gamma(t)$ is horizontal for all t

(ii) $\pi \circ \gamma$ is a geodesic in (N, h) of the same length as γ

In our case we are interested in the geometry of $\mathbb{C}_*^{n\times r}/U(r)$, where $\mathbb{C}_*^{n\times r}$ is an open subset of $\mathbb{C}^{n\times r}$ and is therefore a smooth Riemannian manifold of constant metric when equipped with the standard real inner product on $\mathbb{C}^{n\times r}$

$$\langle A, B\rangle_{\mathbb{R}} = \Re\mathrm{tr}\{A^*B\} \tag{1.10.11}$$

The relevant compact Lie group of isometries will be $U(r)$, acting by matrix multiplication on the right. We note that while $U(r)$ does not act freely on $\mathbb{C}^{n\times r}$, it does act freely on $\mathbb{C}_*^{n\times r}$ since for $x \in \mathbb{C}_*^{n\times r}$ and $W \in U(r)$

$$x = xW \iff x^*x = x^*xW \iff (x^*x)^{-1}(x^*x) = W \iff \mathbb{I}_{r\times r} = W \tag{1.10.12}$$

Therefore by Theorem 1.10.2 there exists a metric h on $\mathbb{C}_*^{n\times r}/U(r)$ such that the differential of π at x

$$\begin{aligned} &D\pi(x) : (H_{\pi,x}(\mathbb{C}_*^{n\times r}), \langle\cdot,\cdot\rangle_{\mathbb{R}}) \to (T_{\pi(x)}(S^{r,0}(\mathbb{C}^n)), h) \\ &D\pi(x)(w) = xw^* + wx^* \end{aligned} \tag{1.10.13}$$

is an isometric isomorphism. Indeed

$$h(Z_1, Z_2) = \langle D\pi(x)^\dagger Z_1, D\pi(x)^\dagger Z_2\rangle_{\mathbb{R}} \tag{1.10.14}$$

Where $D\pi(x)^\dagger$ is the pseudo-inverse of the linear operator $D\pi(x)$. In this case, for $w_1, w_2 \in$

$H_{\pi,x}(\mathbb{C}_*^{n\times r})$

$$h(D\pi(w_1), D\pi(w_2)) = \langle D\pi(x)^\dagger D\pi(w_1), D\pi(x)^\dagger D\pi(w_2)\rangle_{\mathbb{R}} = \langle w_1, w_2\rangle_{\mathbb{R}} \tag{1.10.15}$$

We now determine h explicitly. Namely, if $Z_1, Z_2 \in T_{\pi(x)}(\mathring{S}^{r,0}(\mathbb{C}^n)) = D\pi(H_{\pi,x}(\mathbb{C}_*^{n\times r}))$ then $Z_i = D\pi(x)(H_i x + X_i)$ where H_i, X_i are as in (1.7.6). We must have

$$\begin{aligned} h(Z_1, Z_2) &= \Re\mathrm{tr}[(H_1x + X_1)^*(H_2x + X_2)] \\ &= \Re\mathrm{tr}[x^*H_1H_2x] + \Re\mathrm{tr}[X_1^*X_2] \end{aligned} \tag{1.10.16}$$

We define $Z_i^{\|} := \mathbb{P}_{\mathrm{Ran}(x)} Z_i \mathbb{P}_{\mathrm{Ran}(x)} = xx^*H_i + H_ixx^*$ and $Z_i^{\perp} := \mathbb{P}_{\mathrm{Ran}(x)^\perp} Z_i \mathbb{P}_{\mathrm{Ran}(x)} = X_ix^*$. Then

$$\begin{aligned} H_i &= \int_0^\infty e^{-txx^*} Z_i^{\|} e^{-txx^*} dt \\ X_i &= Z_i^{\perp} x(x^*x)^{-1} \end{aligned} \tag{1.10.17}$$

Plugging these expressions into (1.10.16) yields the expression

$$\begin{aligned} h(Z_1, Z_2) &= \Re\mathrm{tr}\{xx^* \int_0^\infty e^{-txx^*} Z_1^{\|} e^{-txx^*} dt \int_0^\infty e^{-sxx^*} Z_2^{\|} e^{-sxx^*} ds\} + \Re\mathrm{tr}\{Z_1^{\perp *} Z_2^{\perp} (xx^*)^\dagger\} \\ &:= h_0(Z_1, Z_2) + h_1(Z_1, Z_2) \end{aligned} \tag{1.10.18}$$

The first term in (1.10.18) $h_0(Z_1, Z_2)$ can be simplified via the change of coordinates $u = t + s$

and $v = t - s$ as

$$\begin{aligned}
h_0(Z_1, Z_2) &= \int_0^\infty \int_0^\infty \Re\text{tr}\{e^{-xx^*(t+s)} Z_1^{\|} e^{-xx^*(t+s)} xx^* Z_2^{\|}\} ds dt \\
&= \frac{1}{2} \int_0^\infty \int_{-u}^{u} \Re\text{tr}\{e^{-uxx^*} Z_1^{\|} e^{-uxx^*} xx^* Z_2^{\|}\} dv du \\
&= \int_0^\infty u \Re\text{tr}\{e^{-uxx^*} Z_1^{\|} e^{-uxx^*} xx^* Z_2^{\|}\} du \\
&= \int_0^\infty u\text{tr}\{e^{-uxx^*} Z_1^{\|} e^{-uxx^*} xx^* Z_2^{\|} + Z_2^{\|} xx^* e^{-uxx^*} Z_1^{\|} e^{-uxx^*}\} du \\
&= -\text{tr}\{Z_2^{\|} \int_0^\infty u \frac{\partial}{\partial u} e^{-uxx^*} Z_1^{\|} e^{-uxx^*} du\} \\
&= \text{tr}\{Z_2^{\|} \int_0^\infty e^{-uxx^*} Z_1^{\|} e^{-uxx^*} du\} \\
&= \langle H_1, Z_2 \rangle_{\mathbb{R}} = \langle Z_1, H_2 \rangle_{\mathbb{R}}
\end{aligned} \tag{1.10.19}$$

Where the last equality follows from cycling under the trace immediately and then repeating the same calculation. With this metric in hand we have shown (ii), namely that the map

$$\pi : (\mathbb{C}_*^{n\times r}, \langle \cdot, \cdot \rangle_{\mathbb{R}}) \to (\mathring{S}^{r,0}(\mathbb{C}^n), h) \tag{1.10.20}$$

is a Riemannian submersion. To prove (iii), let $A, B \in \mathring{S}^{r,0}(\mathbb{C}^n)$ and let xx^* and yy^* be their respective Cholesky decompositions, so that $x, y \in \mathbb{C}_*^{n\times r}$. Consider the following straight line curve in $\mathbb{C}^{n\times r}$:

$$\begin{aligned}
\sigma_{x,y} &: [0,1] \to \mathbb{C}^{n\times r} \\
\sigma_{x,y}(t) &= (1-t)x + tyU
\end{aligned} \tag{1.10.21}$$

Where U is a polar factor such that $x^*yU = |x^*y|$ (equivalently U is a minimizer of the distance

D, as in Proposition 2). The claim is that we will be able to apply Theorem 1.10.3 to the push-forward of $\sigma_{x,y}$, proving that it is a geodesic connecting $A = \pi(x)$ to $B = \pi(yU)$. Specifically, we would like to prove

$$\sigma_{x,y}(t) \in \mathbb{C}_*^{n\times r} \qquad \forall t \in [0,1] \tag{1.10.22}$$

$$\dot{\sigma}_{x,y}(0) \in H_{\pi,x}(\mathbb{C}_*^{n\times r}) \tag{1.10.23}$$

We first prove (1.10.22), namely that $\sigma_{x,y}(t)$ does not drop rank as t varies from 0 to 1 even though $\mathbb{C}_*^{n\times r}$ is not convex. The endpoints $\sigma_{x,y}(0) = x$ and $\sigma_{x,y}(1) = yU$ are of course full rank, so it is enough to prove it for $t \in (0,1)$. Consider $x^*\sigma_{x,y}(t)$:

$$x^*\sigma_{x,y}(t) = (1-t)\underbrace{x^*x}_{\in \mathbb{P}(r)} \quad + \quad t\underbrace{x^*yU}_{|x^*y| \in PSD(r)} \in \mathbb{P}(r) \text{ for } t \in (0,1) \tag{1.10.24}$$

This implies that $\sigma_{x,y}(t) \in \mathbb{C}_*^{n\times r}$ for $t \in (0,1)$, so (1.10.22) is proved. Let $v = \dot{\sigma}_{x,y}(0) = yU - x$. Then

$$\begin{aligned}
x^*v &= -x^*x + x^*yU = -x^*x + (x^*yy^*x)^{\frac{1}{2}} \\
\mathbb{P}_{\text{Ran}(x)}v &= -(xx^*)^\dagger xx^*x + (xx^*)^\dagger x(x^*yy^*x)^{\frac{1}{2}} \\
\mathbb{P}_{\text{Ran}(x)}v &= \underbrace{(-\mathbb{P}_{\text{Ran}(x)} + (xx^*)^\dagger x(x^*yy^*x)^{\frac{1}{2}}x^*(xx^*)^\dagger)}_{H}\, x \\
v &= Hx + X, \quad \mathbb{P}_{\text{Ran}(x)}X = 0, \quad H^* = \mathbb{P}_{\text{Ran}(x)}H = H
\end{aligned} \tag{1.10.25}$$

Hence (1.10.23) is proved and so by Theorem 1.10.3 we have that $\gamma_{A,B} := \pi \circ \sigma_{x,y}$ is a geodesic

on $(\mathring{S}^{r,0}(\mathbb{C}^n), h)$ connecting A and B. We find specifically that this geodesic is given by

$$\begin{aligned}\gamma_{A,B}(t) &= \pi((1-t)x + tyU) \\ &= ((1-t)x + tyU)((1-t)x + tyU)^* \\ &= (1-t)^2 xx^* + t^2 yy^* + t(1-t)(xU^*y^* + yUx^*)\end{aligned} \tag{1.10.26}$$

Clearly $A = xx^*$ and $B = yy^*$, but what about xU^*y^* and yUx^*? Fortunately, a minor miracle occurs. Namely,

$$\begin{aligned}(yUx^*)^2 &= yUx^*yUx^* = yU|x^*y|x^* = y(|x^*y|U^*)^*x^* = y(x^*y)^*x^* = yy^*xx^* \\ (xU^*y^*)^2 &= xU^*y^*xU^*y^* = x(x^*yU)^*U^*y^* = x|x^*y|U^*y^* = xx^*yy^*\end{aligned} \tag{1.10.27}$$

Thus in fact xU^*y^* and yUx^* are matrix square roots (not necessarily symmetric, but having positive non-zero eigenvalues) for BA and AB respectively. We obtain the following expression for the family of geodesics on $\mathring{S}^{r,0}(\mathbb{C}^n)$ connecting A and B

$$\gamma_{A,B}(t) = (1-t)^2 xx^* + t^2 yy^* + t(1-t)(xU_0^*y^* + yU_0x^*) + t(1-t)(xU_1^*y^* + yU_1x^*) \tag{1.10.28}$$

Where U_0 and U_1 are as in Proposition 2. The fact that the form of this expression is independent of r is somewhat surprising, and motivates claims (iv) and (v). In order to prove (iv) we must first check that the collection of smooth manifolds $(\mathring{S}^{i,0}(\mathbb{C}^n))_{i=0}^r$ provide a stratification of the cone $S^{r,0}(\mathbb{C}^n)$ (conditions (a) and (b) of Definition 1.7.2). Condition (a) is satisfied trivially and

for (b) we note that

$$\overline{\mathring{S}^{i,0}(\mathbb{C}^n)}\backslash\mathring{S}^{i,0}(\mathbb{C}^n) = \{0\} \cup \mathring{S}^{1,0} \cup \cdots \cup S^{i-1,0} \tag{1.10.29}$$

It remains to check that whenever $p > q$ the triple $(\mathring{S}^{p,0}(\mathbb{C}^n), \mathring{S}^{q,0}(\mathbb{C}^n), A)$ is a-regular and b-regular for $A \in \mathring{S}^{q,0} \subset \overline{\mathring{S}^{p,0}}$. It was noted by John Mather in Proposition 2.4 of [38] that b-regularity implies a-regularity, but we will use a-regularity in our proof of b-regularity so we need to prove a-regularity first. Specifically, a-regularity in this case states that if $(A_i)_{i\geqslant 1} \subset \mathring{S}^{p,0}(\mathbb{C}^n)$ converges to $A \in \mathring{S}^{q,0}(\mathbb{C}^n)$ and if $T_{A_i}(\mathring{S}^{p,0}(\mathbb{C}^n))$ converges in Grassmannian sense to the vector space τ_A then $T_A(\mathring{S}^{q,0}(\mathbb{C}^n)) \subset \tau_A$. Upon examining the form of the tangent space as given by (1.7.7) it becomes clear that convergence of the tangent spaces $T_{A_i}(\mathring{S}^{p,0}(\mathbb{C}^n))$ is equivalent to convergence of $\text{Ran}A_i$ to a space we denote L, so that the Grassmannian limit of the tangent spaces is given by

$$\tau_A = \{W \in \text{Sym}(\mathbb{C}^n) | \mathbb{P}_{L^\perp} W \mathbb{P}_{L^\perp} = 0\} \tag{1.10.30}$$

It is evident that L should contain as a subspace $\text{Ran}A$, and that this would prove that the stratification given is a-regular. Indeed, if $A_i = U_i \Lambda_i U_i^*$ is the low rank diagonalization of A_i so that $\Lambda_i = \text{diag}(\lambda_1, \ldots, \lambda_p)$ is the diagonal matrix of non-zero eigenvalues of A_i and $U_i U_i^* = \mathbb{P}_{\text{Ran}_{A_i}}$, $U_i^* U_i = \mathbb{I}_{p\times p}$ then by compactness we can obtain a subsequence of $(U_i)_{i\geqslant 1}$ that converges to a matrix U such that the columns of U are precisely an orthonormal basis for L. In this case, we may write $A = U\Lambda U^*$ since $A = \lim_{i\to\infty} U_i \Lambda_i U_i^*$ and the sequences of eigenvalues converge

(some to zero), so that if $U = [u_1|\cdots|u_p]$ then

$$\text{Ran}A = \text{span}\{u_i|\Lambda_{ii} \neq 0\} \subset \text{span}\{u_i\}_{i=1}^{p} = L \tag{1.10.31}$$

Thus, owing to (1.10.30) and the description of the tangent space in (1.7.7) we conclude that $\mathbb{T}_A(\mathring{S}^{q,0}(\mathbb{C}^n)) \subset \tau_A$ and our stratification is a-regular. As for b-regularity, let $(A_i)_{i\geqslant 1} \subset \mathring{S}^{p,0}(\mathbb{C}^n)$, $A \in \mathring{S}^{q,0}(\mathbb{C}^n)$, and τ_A be as before (specifically we assume the Grassmannian limit defining τ_A converges) and let $(B_i)_{i\geqslant 1} \subset \mathring{S}^{q,0}(\mathbb{C}^n)$ be convergent also to A such that the following limit exists

$$Q = \lim_{i\to\infty} Q_i := \lim_{i\to\infty} \frac{A_i - B_i}{||A_i - B_i||_2} \tag{1.10.32}$$

We claim that $Q \in \tau_A$. Specifically, let $\Theta_i = A_i - \mathbb{P}_{\text{Ran}(A_i)} B_i \mathbb{P}_{\text{Ran}(A_i)}$ and $\Psi_i = \mathbb{P}_{\text{Ran}(A_i)} B_i \mathbb{P}_{\text{Ran}(A_i)} - B_i$. Then either $\Psi_i = 0$, in which case $Q_i = \Theta_i/||\Theta_i||_2$, or $\Psi_i \neq 0$, so that

$$Q_i = \frac{||\Theta_i||_2}{||A_i - B_i||_2} \frac{\Theta_i}{||\Theta_i||_2} + \frac{||\Psi_i||_2}{||A_i - B_i||_2} \frac{\Psi_i}{||\Psi_i||_2} \tag{1.10.33}$$

We will obtain convergent subsequences for the sequences of unit norm matrices $\Theta_i/||\Theta_i||_2$ and $\Psi_i/||\Psi_i||_2$, but first note that

$$\frac{||\Theta_i||_2}{||A_i - B_i||_2} = \frac{||\mathbb{P}_{\text{Ran}(A_i)}(A_i - B_i)\mathbb{P}_{\text{Ran}(A_i)}||_2}{||A_i - B_i||_2} \leqslant 1 \tag{1.10.34}$$

Hence $||\Psi_i||_2/||A_i - B_i||_2$ is also a bounded sequence (if it were not Q_i would fail to converge). Next note that for i sufficiently large $\Psi_i = \mathbb{P}_{\text{Ran}(A_i)} B_i \mathbb{P}_{\text{Ran}(A_i)} - B_i$ is the difference of two matrices in $\mathring{S}^{q,0}(\mathbb{C}^n)$, both converging to A. Therefore, owing to the fact that $\mathring{S}^{q,0}(\mathbb{C}^n)$

is an analytic manifold, any convergent subsequence of $\Psi_i/||\Psi_i||_2$ will have its limit lying in $T_A(\mathring{S}^{q,0}(\mathbb{C}^n))$ (see for example Lemma 4.12 in [39]). Owing to the already proved a-regularity we conclude that the limit of any convergent subsequence of $\Psi_i/||\Psi_i||_2$ lies in τ_A. Similarly, $\Theta_i = \mathbb{P}_{\text{Ran}(A_i)}(A_i - B_i)\mathbb{P}_{\text{Ran}(A_i)}$ hence any convergent subsequence of $\Theta_i/||\Theta_i||_2$ must lie in τ_A. Thus we may obtain a subsequence such that the sequences of real numbers $||\Theta_{i_j}||_2/||A_{i_j} - B_{i_j}||_2$ and $||\Psi_{i_j}||_2/||A_{i_j} - B_{i_j}||_2$ converge to some $\alpha, \beta \in \mathbb{R}$ and the sequences of unit norm matrices $\Theta_{i_j}/||\Theta_{i_j}||_2$ and $\Psi_{i_j}/||\Psi_{i_j}||_2$ converge to some $\hat{\Theta}, \hat{\Psi} \in \tau_A$. Since $(Q_i)_{i\geqslant 1}$ converges, we find that

$$Q = \alpha\hat{\Theta} + \beta\hat{\Psi} \in \tau_A \tag{1.10.35}$$

Thus the stratification $(\mathring{S}^{i,0}(\mathbb{C}^n))_{i=0}^r$ is b-regular and in particular is a Whitney stratification of $S^{r,0}(\mathbb{C}^n)$.

In order to prove (v), let $A_i = x_i x_i^*$ and $B_i = y_i y_i^*$ be Cholesky decompositions of A_i and B_i such that $x_i, y_i \in \mathbb{C}^{n\times p}$ and note that we are told the following limit exists at each t

$$\delta(t) = \lim_{i\to\infty}(1-t)^2 x_i x_i^* + t^2 y_i y_i^* + t(1-t)(x_i U_i^* y_i^* + y_i U_i x_i^*) \tag{1.10.36}$$

Where $U_i \in U(p)$ is such that $x_i^* y_i U_i \geqslant 0$. We note that since $(A_i)_{i\geqslant 1}$ and $(B_i)_{i\geqslant 1}$ converge we may obtain convergent subsequences for their Cholesky factors x_i and y_i ($||x_i||_2$ and $||y_i||_2$ must both be bounded or else A_i and B_i would not converge). We may also obtain a convergent subsequence for $(U_i)_{i\geqslant 1}$ owing to the compactness of $U(p)$. Denote these subsequential limits by x, y, and U respectively and consider a combined subsequential indexing such that each occurs. Let V_x and V_y be the matrices of right singular vectors for x and y so that $x = [\hat{x}|0]V_x$ and

$y = [\hat{y}|0]V_y$ for some $\hat{x}, \hat{y} \in \mathbb{C}_*^{n\times q}$. Then clearly

$$\delta(t) = (1-t)^2\hat{x}\hat{x}^* + t^2\hat{y}\hat{y}^* + t(1-t)(\hat{x}\hat{U}^*\hat{y}^* + \hat{y}\hat{U}\hat{x}^*) \tag{1.10.37}$$

Where $\hat{U}$ is the upper left $q \times q$ block of $V_yUV_x^*$. We will prove that in fact

$$V_yUV_x^* = \left[\begin{array}{c|c} \hat{U} & 0 \\ \hline 0 & \tilde{U} \end{array}\right] \tag{1.10.38}$$

In particular, this will imply that $\hat{U} \in U(q)$ since $V_yUV_x^* \in U(p)$ hence the upper left $q \times q$ blocks of $(V_yUV_x^*)(V_yUV_x^*)^*$ and $(V_yUV_x^*)^*(V_yUV_x^*)$ must both be equal to the $q \times q$ identity matrix. In order to prove (1.10.38), note that $U = VW^*$ where

$$x^*y = W\left[\begin{array}{c|c} \Sigma & 0 \\ \hline 0 & 0 \end{array}\right]V^* \tag{1.10.39}$$

is a singular value decomposition of x^*y. On the other hand if

$$\hat{x}^*\hat{y} = P\left[\begin{array}{c|c} \Lambda & 0 \\ \hline 0 & 0 \end{array}\right]Q^* \tag{1.10.40}$$

is a singular value decomposition for $\hat{x}^*\hat{y}$ then

$$x^*y = V_x^* \underbrace{\left[\begin{array}{c|c} P & 0 \\ \hline 0 & \tilde{P} \end{array}\right]}_{W} \left[\begin{array}{cc|c} \Lambda & 0 & \\ 0 & 0 & 0 \\ \hline & 0 & 0 \end{array}\right] \underbrace{\left[\begin{array}{c|c} Q & 0 \\ \hline 0 & \tilde{Q} \end{array}\right]}_{V^*} V_y \tag{1.10.41}$$

Where $\tilde{P}, \tilde{Q} \in U(p-q)$ are in general arbitrary, but may of course be chosen in accordance with W and V. Thus

$$V_y U V_x^* = V_y V W^* V_x = \begin{bmatrix} PQ & 0 \\ 0 & \tilde{P}\tilde{Q} \end{bmatrix} \tag{1.10.42}$$

is as in (1.10.38). The question remains whether $\hat{x}^*\hat{y}\hat{U} \geqslant 0$, but we note that

$$\begin{aligned} x^*yU &= V_x^* \begin{bmatrix} \hat{x}^*\hat{y} & 0 \\ 0 & 0 \end{bmatrix} V_y U \\ &= V_x^* \begin{bmatrix} \hat{x}^*\hat{y} & 0 \\ 0 & 0 \end{bmatrix} V_y U V_x^* V_x \\ &= V_x^* \begin{bmatrix} \hat{x}^*\hat{y} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{U} & 0 \\ 0 & \tilde{U} \end{bmatrix} V_x \\ &= V_x^* \begin{bmatrix} \hat{x}^*\hat{y}\hat{U} & 0 \\ 0 & 0 \end{bmatrix} V_x \end{aligned} \tag{1.10.43}$$

Thus x^*yU will be positive semidefinite only if $\hat{x}^*\hat{y}\hat{U}$ is positive semidefinite, and since $x^*yU = \lim_{i\to\infty} x_i^* y_i U_i = \lim_{i\to\infty} |x_i^* y_i| \geqslant 0$ we conclude that $\hat{x}^*\hat{y}\hat{U} \geqslant 0$. A nearly identical proof shows that $Ux^*y \geqslant 0$. We conclude that δ is a geodesic in $\mathring{S}^{q,0}(\mathbb{C}^n)$ connecting A and B. □

1.11 Proofs for Section 1.8

1.11.1 Proof of Proposition 5

Proof. We may first note that $\langle xx^*, A_j\rangle_{\mathbb{R}} - \langle yy^*, A_j\rangle_{\mathbb{R}} = \langle xx^* - yy^*, A_j\rangle_{\mathbb{R}}$. The expression (1.3.3) then becomes

$$a_0 = \inf_{\substack{L\in S^{r,r}(\mathbb{C}^n)\\ ||L||_2=1}} \sum_{j=1}^{m} \langle L, A_j\rangle^2 \tag{1.11.1}$$

The claim follows by contradiction if $S^{r,r}$ is closed. Explicitly, if $S^{r,r}$ is closed then $S^{r,r} \cap \{x \in \mathbb{C}^{n\times n} : ||x||_2 = 1\}$ is compact. Assume $a_0 = 0$, then there exists $L_0 \in S^{r,r} \cap \{x \in \mathbb{C}^{n\times n} : ||x||_2 = 1\}$ so that

$$0 = \sum_{j=1}^{m} \langle L_0, A_j\rangle^2 \tag{1.11.2}$$

This implies that the map β is not injective since, in particular, if $xx^* = (L_0)_+$ and $yy^* = (L_0)_-$ then $xx^* \neq yy^*$ since $||L_0||_2 = 1$ but $\beta(x) = \beta(y)$. It remains to show that the spaces $S^{p,q}$ and in particular $S^{r,r}$ are closed. Consider the map $\eta : \mathbb{C}^{n\times n} \to \{0,\ldots,n\}^2$ with $\eta(A) = (\text{rank}(A_+), \text{rank}(A_-))$ taking A to its Sylvester indices (p,q). Then η is continuous with respect to the usual topology on $\mathbb{C}^{n\times n}$ and with respect to the "upper box" topology τ_{ub} on $\{0,\ldots,n\}^2$

generated by the base

$$\mathcal{B}_{\text{ub}} = \{\{x, \dots, n\} \times \{y, \dots, n\} | (x, y) \in \{0, \dots, n+1\}\} \tag{1.11.3}$$

The maps $A \to A_{\pm}$ are continuous and it is well known that $\text{rank}(A+B) \geqslant \text{rank}(A)$ whenever $||B||_{2\to 2} < \sigma_{p+q}(A)$, hence η is continuous. Moreover $\{0, \dots, p\} \times \{0, \dots, q\}$ is closed in τ_{ub} hence $S^{p,q}$, its pullback through the continuous map η, is closed in $\mathbb{C}^{n\times n}$. □

1.11.2 Proof of Theorem 1.8.5

Proof. We first prove that $a_0 = \inf_{z\in\mathbb{C}^{n\times r}} a(z)$. We note that

$$a_0 = \inf_{\substack{x,y\in\mathbb{C}^{n\times r} \\ xx^* \neq yy^*}} \frac{1}{||xx^* - yy^*||_2^2} \sum_{j=1}^{m} |\langle xx^* - yy^*, A_j\rangle_{\mathbb{R}}|^2 \tag{1.11.4}$$

We may change coordinates to $z = \frac{1}{2}(x+y)$ and $w = x - y$ so that

$$a_0 = \inf_{\substack{z,w\in\mathbb{C}^{n\times r} \\ zw^* + wz^* \neq 0}} \frac{1}{||zw^* + wz^*||_2^2} \sum_{j=1}^{m} |\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}}|^2 \tag{1.11.5}$$

Recall that z has rank k, and therefore we may take $z = [\hat{z}|0]U$ for $\hat{z} \in \mathbb{C}_*^{n\times k}$ and $U \in U(r)$. We then define $\hat{w} \in \mathbb{C}^{n\times k}$ via the first k columns of wU^* then $zw^* + wz^* = \hat{z}\hat{w}^* + \hat{w}\hat{z}^* = D\pi(\hat{z})(\hat{w})$,

so that in fact we may take $\hat{w} \in H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k})\backslash\{0\}$. We obtain

$$\begin{aligned} a_0 &= \inf_{z\in\mathbb{C}^{n\times r}\backslash\{0\}} \inf_{\hat{w}\in H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k})\backslash\{0\}} \frac{1}{||D\pi(\hat{z})(\hat{w})||_2^2} \sum_{j=1}^m |\langle D\pi(\hat{z})(\hat{w}), A_j\rangle_{\mathbb{R}}|^2 \\ &= \inf_{z\in\mathbb{C}^{n\times r}\backslash\{0\}} \min_{\substack{W\in T_{\pi(\hat{z})}(\mathring{S}^{k,0}(\mathbb{C}^n)) \\ ||W||_2=1}} \sum_{j=1}^m |\langle W, A_j\rangle_{\mathbb{R}}|^2 \\ &= \inf_{\substack{z\in\mathbb{C}^{n\times r} \\ ||z||_2=1}} \min_{\substack{W\in T_{\pi(\hat{z})}(\mathring{S}^{k,0}(\mathbb{C}^n)) \\ ||W||_2=1}} \sum_{j=1}^m |\langle W, A_j\rangle_{\mathbb{R}}|^2 \\ &= \inf_{\substack{z\in\mathbb{C}^{n\times r} \\ ||z||_2=1}} a(z) \end{aligned} \tag{1.11.6}$$

This proves (1.8.11). The first two inequalities of (1.8.12) are clear from the definitions of the quantities involved, namely $a_0 \leqslant a_2(z) \leqslant a_1(z)$. It remains to prove that $a_1(z) \leqslant a(z)$. We will need the following families of real-linear subspaces of $\mathbb{C}^{n\times r}$ indexed by $z \in \mathbb{C}^{n\times r}$.

$$H_z = \{Hz + X | H \in \mathbb{C}^{n\times n}, H^* = H = \mathbb{P}_{\mathbf{Ran}(z)} H, X \in \mathbb{C}^{n\times r}, \mathbb{P}_{\mathbf{Ran}(z)} X = 0, X\mathbb{P}_{\ker(z)} = 0\} \tag{1.11.7}$$

$$\Delta_z = \{w \in \mathbb{C}^{n\times r} | \quad \exists \rho > 0 \quad \forall |\epsilon| < \rho \quad z^*(z + \epsilon w) \geqslant 0\} \tag{1.11.8}$$

$$\Gamma_z = \{y \in \mathbb{C}^{n\times r} | \mathbb{P}_{\mathbf{Ran}(z)} y = 0, \quad y\mathbb{P}_{\ker(z)} = y\} \tag{1.11.9}$$

Lemma 1.11.1. *The space Δ_z is alternately characterized as*

$$\Delta_z = \{w \in \mathbb{C}^{n\times r} | z^* w = w^* z\} \tag{1.11.10}$$

And is thus manifestly a real-linear subspace. Moreover, Δ_z decomposes orthogonally into

$$\Delta_z = H_z \oplus \Gamma_z \tag{1.11.11}$$

Finally, if $z = [\hat{z}|0]U$ for $\hat{z} \in \mathbb{C}_*^{n\times k}$ then

$$H_z = \left[H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k}) \middle| 0 \right] U \tag{1.11.12}$$

Proof. Clearly a necessary and sufficient condition for $w \in \Delta_z$ is that $z^*w = w^*z$, for in this case take $|\epsilon| < \sigma_k(z)/||w||_2$. We can use this condition to obtain a parametrization for Δ_z:

$$\begin{aligned}
w \in \Delta_z &\iff z^*w = w^*z \\
&\iff z^*w = \tilde{H} \qquad \tilde{H} \in \mathbb{C}^{r\times r}, \tilde{H}^* = \tilde{H} = \mathbb{P}_{\ker(z)^\perp}\tilde{H} \\
&\iff z^*w = z^*Hz \quad H \in \mathbb{C}^{n\times n}, H^* = H = \mathbb{P}_{\mathrm{Ran}(z)}H \\
&\iff w = Hz + X \quad H \in \mathbb{C}^{n\times n}, H^* = H = \mathbb{P}_{\mathrm{Ran}(z)}H, X \in \mathbb{C}^{n\times r}, \mathbb{P}_{\mathrm{Ran}(z)}X = 0
\end{aligned} \tag{1.11.13}$$

This proves (1.11.11), with orthogonality easily verified. To prove (1.11.12) note that if $z = [\hat{z}|0]U$ for $\hat{z} \in \mathbb{C}_*^{n\times k}$, $U \in U(r)$, and $w = Hz + X \in H_z$ then the condition $X\mathbb{P}_{\ker(z)} = 0$ implies

$X = [\tilde{X}|0]U$ for $\tilde{X} \in \mathbb{C}^{n\times k}$ and $\mathbb{P}_{\text{Ran}(z)}X = 0$ if and only if $\mathbb{P}_{\text{Ran}(z)}\tilde{X} = 0$. Thus

$$\begin{aligned} H_z &= \{H[\hat{z}|0]U + [\tilde{X}|0]U | H \in \mathbb{C}^{n\times n}, H^* = H = \mathbb{P}_{\text{Ran}(z)}H, \tilde{X} \in \mathbb{C}^{n\times k}, \mathbb{P}_{\text{Ran}(z)}\tilde{X} = 0\} \\ &= \{[H\hat{z} + \tilde{X}|0]U | H \in \mathbb{C}^{n\times n}, H^* = H = \mathbb{P}_{\text{Ran}(\hat{z})}, \tilde{X} \in \mathbb{C}^{n\times k}, \mathbb{P}_{\text{Ran}(\hat{z})}\tilde{X} = 0\} \\ &= [H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k})|0]U \end{aligned} \tag{1.11.14}$$

□

With this lemma in mind, we may transform $a_1(z)$ into a linear minimization problem over Δ_z. Namely

$$\begin{aligned} a_1(z) &= \lim_{R\to 0} \inf_{\substack{x\in\mathbb{C}^{n\times r} \\ ||xx^*-zz^*||_2<R}} \frac{\sum_{j=1}^m |\langle xx^* - zz^*, A_j\rangle_{\mathbb{R}}|^2}{||xx^* - zz^*||_2^2} \\ &= \lim_{R\to 0} \inf_{\substack{x\in\mathbb{C}^{n\times r} \\ ||xx^*-zz^*||_2<R \\ z^*x\geqslant 0}} \frac{\sum_{j=1}^m |\langle xx^* - zz^*, A_j\rangle_{\mathbb{R}}|^2}{||xx^* - zz^*||_2^2} \end{aligned} \tag{1.11.15}$$

We can add the $z^*x \geqslant 0$ constraint without altering the infimimum since doing so amounts to a choice of representative for x, but x only appears as $\pi(x) = xx^*$. We now show the following lemma, implying that we may instead minimize over $||x - z||_2 < R$.

Lemma 1.11.2. *For all $z \in \mathbb{C}^{n\times r}$ and $\epsilon > 0$ there exists $\delta > 0$ such that if $z^*x \geqslant 0$ and $||zz^* - xx^*||_2 < \delta$ then $||z - x||_2 < \epsilon$.*

Proof. We begin with the fact that the operation

$$\begin{aligned} \zeta &: PSD(n) \to PSD(n) \\ \zeta(A) &= \sqrt{\mathrm{tr}A}\sqrt{A} \end{aligned} \tag{1.11.16}$$

is continuous with respect to the topology induced by the Frobenius norm. Note that $\zeta(xx^*) = ||x||_2(xx^*)^{\frac{1}{2}} = \psi(x)$ (the embedding ψ as given in Definition 1.6.3). Therefore, given any $z \in \mathbb{C}^{n\times r}$ and ϵ_1 there exists δ such that

$$||xx^* - zz^*||_2 < \delta \implies ||||x||_2(xx^*)^{\frac{1}{2}} - ||z||_2(zz^*)^{\frac{1}{2}}||_2 < \epsilon_1 \tag{1.11.17}$$

The latter expression here is of course $||\psi(x) - \psi(z)||_2$, which satisfies $||\psi(x) - \psi(z)||_2 \geqslant \frac{1}{2}D(x,z)^2$ by (1.9.19). If $z^*x \geqslant 0$ then $D(x,z) = ||x - z||_2$, so if we take $\epsilon_1 = \frac{\epsilon^2}{2}$ then the above δ satisfies the lemma. □

With this lemma in hand we may freely replace $||xx^* - zz^*||_2$ by $||x - z||_2$ in the infimization constraint for $a_1(z)$ (note that the converse of the lemma is immediate since π is continuous with respect to the topology induced by the Frobenius norm). After doing so, we change variables

from x to $w = x - z$ so that

$$
\begin{aligned}
a_1(z) &= \lim_{R\to 0} \inf_{\substack{x\in\mathbb{C}^{n\times r} \\ ||x-z||_2<R \\ z^*x\geqslant 0}} \frac{\sum_{j=1}^m |\langle xx^* - zz^*, A_j\rangle_{\mathbb{R}}|^2}{||xx^* - zz^*||_2^2} \\
&= \lim_{R\to 0} \inf_{\substack{w\in\mathbb{C}^{n\times r} \\ ||w||_2<R \\ z^*(z+w)\geqslant 0}} \frac{\sum_{j=1}^m |\langle zw^* + wz^* + ww^*, A_j\rangle_{\mathbb{R}}|^2}{||zw^* + wz^* + ww^*||_2^2} \\
&= \lim_{R\to 0} \inf_{\substack{w\in\Delta_z \\ ||w||_2<R}} \frac{\sum_{j=1}^m |\langle zw^* + wz^* + ww^*, A_j\rangle_{\mathbb{R}}|^2}{||zw^* + wz^* + ww^*||_2^2} \\
&\leqslant \lim_{R\to 0} \inf_{\substack{w\in H_z \\ ||w||_2<R}} \frac{\sum_{j=1}^m |\langle zw^* + wz^* + ww^*, A_j\rangle_{\mathbb{R}}|^2}{||zw^* + wz^* + ww^*||_2^2} \\
&= \lim_{R\to 0} \inf_{\substack{w\in H_z \\ ||w||_2<R}} \frac{\sum_{j=1}^m |\langle zw^* + wz^* + ww^*, A_j\rangle_{\mathbb{R}}|^2}{||zw^* + wz^*||_2^2 + ||ww^*||_2^2 + 4\Re\text{tr}\{zw^*ww^*\}} \\
&\leqslant \lim_{R\to 0} \inf_{\substack{w\in H_z \\ ||w||_2<R}} \frac{\sum_{j=1}^m |\langle zw^* + wz^* + ww^*, A_j\rangle_{\mathbb{R}}|^2}{||zw^* + wz^*||_2^2(1 + 4\frac{\Re\text{tr}\{zw^*ww^*\}}{||zw^*+wz^*||_2^2})}
\end{aligned}
\tag{1.11.18}
$$

We need to show that the ratio

$$
R(w) = 4\frac{|\Re\text{tr}\{zw^*ww^*\}|}{||zw^* + wz^*||_2^2} \tag{1.11.19}
$$

is $O(||w||)$ when $w \in H_z$. We employ the parametrization of H_z given in (1.11.7) and note that for $w = Hz + X$

$$
||zw^* + wz^*||_2^2 = 2(||z^*Hz||_2^2 + ||zz^*H||_2^2 + ||zX^*||_2^2) \tag{1.11.20}
$$

$$
\Re\text{tr}\{zw^*ww^*\} = \Re\text{tr}\{z^*H^2zz^*Hz\} + \Re\text{tr}\{X^*Xz^*Hz\} \tag{1.11.21}
$$

Thus we find

$$\begin{aligned} R(w) &\leqslant \frac{2|\Re\text{tr}\{z^*H^2zz^*Hz\}| + 2|\Re\text{tr}\{X^*Xz^*Hz\}|}{||z^*Hz||_2^2 + ||zz^*H||_2^2 + ||zX^*||_2^2} \\ &\leqslant 2\frac{|\Re\text{tr}\{z^*H^2zz^*Hz\}|}{||z^*Hz||_2^2} + 2\frac{|\Re\text{tr}\{X^*Xz^*Hz\}|}{||zX^*||_2^2 + ||z^*Hz||_2^2} \\ &\leqslant 2\frac{||z^*H^2z||_2}{||z^*Hz||_2} + \frac{||X^*X||_2}{||zX^*||_2} \end{aligned} \tag{1.11.22}$$

Up until this point we have not used the fact that $H\mathbb{P}_{\text{Ran}(z)} = H = \mathbb{P}_{\text{Ran}(z)}H$ and $X\mathbb{P}_{\ker(z)} = 0$. We do so now by noting that if $z = U_1\Lambda V^*$ for $U_1 \in \mathbb{C}^{n\times k}$ such that $U_1U_1^* = \mathbb{P}_{\text{Ran}(z)}$, $\Lambda = \text{diag}(\sigma_1(z), \ldots, \sigma_k(z))$ is the diagonal matrix of ordered singular values $\sigma_1(z) \geqslant \cdots \geqslant \sigma_k(z) > 0$, and $V_1 \in \mathbb{C}^{r\times k}$ such that $V_1V_1^* = \mathbb{P}_{\ker(z)^\perp}$ then

$$\begin{aligned} ||z^*H^2z|| &= ||\Lambda U_1^*H^2U_1\Lambda||_2 \leqslant \sigma_1(z)^2||U_1^*H^2U_1||_2 = \sigma_1(z)^2\sqrt{\text{tr}\{\mathbb{P}_{\text{Ran}(z)}H^2\mathbb{P}_{\text{Ran}(z)}H^2\}} = \sigma_1(z)^2||H^2||_2 \\ ||z^*Hz|| &= ||\Lambda U_1^*HU_1\Lambda||_2 \geqslant \sigma_k(z)^2||U_1^*HU_1||_2 = \sigma_k(z)^2\sqrt{\text{tr}\{\mathbb{P}_{\text{Ran}(z)}H\mathbb{P}_{\text{Ran}(z)}H\}} = \sigma_k(z)||H||_2 \\ ||zX^*||_2 &= ||\Lambda V_1^*X^*||_2 = ||\Lambda(XV_1)^*||_2 \geqslant \sigma_k(z)||XV_1||_2 = \sigma_k(z)\sqrt{\text{tr}\{X\mathbb{P}_{\ker(z)^\perp}X^*\}} = \sigma_k(z)||X||_2 \end{aligned} \tag{1.11.23}$$

Thus if $\kappa(z) = \sigma_1(z)/\sigma_k(z)$ is the condition number of z we find

$$\begin{aligned} R(w) &\leqslant 2\kappa(z)^2 \frac{||H^2||_2}{||H||_2} + \sigma_k(z)^{-1} \frac{||X^*X||_2}{||X||_2} \\ &\leqslant 2\kappa(z)^2 ||H||_2 + \sigma_k^{-1}(z)||X||_2 \\ &\leqslant 2\kappa(z)^2 \sigma_k(z)^{-1} ||Hz||_2 + \sigma_k^{-1}(z)||X||_2 \\ &\leqslant \frac{\sqrt{2}\max(2\kappa(z)^2, 1)}{\sigma_k(z)} \sqrt{||Hz||_2^2 + ||X||_2^2} \\ &= \underbrace{\frac{2\sqrt{2}\kappa(z)^2}{\sigma_k(z)}}_{C(z)} ||w||_2 \end{aligned} \tag{1.11.24}$$

Thus returning to $a_1(z)$ we obtain

$$\begin{aligned} a_1(z) &\leqslant \lim_{R\to 0} \inf_{\substack{w\in H_z \\ ||w||_2<R}} \frac{\sum_{j=1}^m |\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}}|^2}{||zw^* + wz^*||_2^2} (1 + 2C(z)||w||_2) \\ &= \inf_{\substack{w\in H_z \\ w\neq 0}} \frac{\sum_{j=1}^m |\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}}|^2}{||zw^* + wz^*||_2^2} \\ &= \inf_{\substack{w\in H_{\pi,\hat{z}} \\ \hat{w}\neq 0}} \frac{\sum_{j=1}^m |\langle \hat{z}\hat{w}^* + \hat{w}\hat{z}^*, A_j\rangle_{\mathbb{R}}|^2}{||\hat{z}\hat{w}^* + \hat{w}\hat{z}^*||_2^2} \\ &= \min_{\substack{W\in T_{\pi(\hat{z})}(\mathring{S}^{k,0}(\mathbb{C}^n)) \\ ||W||_2=1}} \sum_{j=1}^m |\langle W, A_j\rangle_{\mathbb{R}}|^2 \\ &= a(z) \end{aligned} \tag{1.11.25}$$

This proves (1.8.12). In order to prove (1.8.14) we will employ an explicit parametrization of $T_{\pi(\hat{z})}(\mathring{S}^{k,0}(\mathbb{C}^n))$ implied by (1.7.7). The condition on $W \in \mathrm{Sym}(\mathbb{C}^n)$ in (1.7.7) that $\mathbb{P}_{\mathrm{Ran}(z)^\perp} W \mathbb{P}_{\mathrm{Ran}(z)^\perp} = 0$ implies that

$$W \in T_{\pi(\hat{z})}(\mathring{S}^{k,0}(\mathbb{C}^n)) \iff W = W_1 + \frac{1}{2}(W_2 + W_2^*) \tag{1.11.26}$$

For $W_1, W_2 \in \mathbb{C}^{n\times n}$ where $\mathbb{P}_{\text{Ran}(z)}W_1 = W_1 = W_1^*$, $\mathbb{P}_{\text{Ran}(z)}W_2 = 0$, and $W_2\mathbb{P}_{\text{Ran}(z)} = W_2$. In other words, if $U_1 \in \mathbb{C}^{n\times k}$ and $U_2 \in \mathbb{C}^{n\times n-k}$ are as in Definition 1.8.3 then

$$T_{\pi(\hat{z})}(\mathring{S}^{k,0}) = \{U_1AU_1^* + \frac{1}{2}(U_2BU_1^* + U_1B^*U_2^*)|A \in \text{Sym}(\mathbb{C}^k), B \in \mathbb{C}^{n-k\times k}\} \tag{1.11.27}$$

We will now employ the fact that the maps τ and μ in (1.8.6) are isometries. Specifically, if $A, B \in \text{Sym}(\mathbb{C}^n)$ then $\langle A, B\rangle_{\mathbb{R}} = \tau(A)^T\tau(B)$ and if $X, Y \in \mathbb{C}^{n\times r}$ then $\langle X, Y\rangle_{\mathbb{R}} = \mu(X)^T\mu(Y)$. With this in mind, we obtain that for $W \in T_{\pi(\hat{z})}(\mathring{S}^{k,0})$

$$\begin{aligned}
\sum_{j=1}^m |\langle W, A_j\rangle_{\mathbb{R}}|^2 &= \sum_{j=1}^m |\langle U_1AU_1^* + \frac{1}{2}(U_2BU_1^* + U_1B^*U_2^*), A_j\rangle_{\mathbb{R}}|^2 \\
&= \sum_{j=1}^m |\langle U_1AU_1^*, A_j\rangle_{\mathbb{R}} + \langle U_2BU_1^*, A_j\rangle_{\mathbb{R}}|^2 \\
&= \sum_{j=1}^m |\langle A, U_1^*A_jU_1\rangle_{\mathbb{R}} + \langle B, U_2^*A_jU_1\rangle_{\mathbb{R}}|^2 \\
&= \sum_{j=1}^m \left(\begin{bmatrix} \tau(A) \\ \mu(B) \end{bmatrix}^T \begin{bmatrix} \tau(U_1^*A_jU_1) \\ \mu(U_2^*A_jU_1) \end{bmatrix} \right)^2 \\
&= \begin{bmatrix} \tau(A) \\ \mu(B) \end{bmatrix}^T \left(\sum_{j=1}^m \begin{bmatrix} \tau(U_1^*A_jU_1) \\ \mu(U_2^*A_jU_1) \end{bmatrix} \begin{bmatrix} \tau(U_1^*A_jU_1) \\ \mu(U_2^*A_jU_1) \end{bmatrix}^T \right) \begin{bmatrix} \tau(A) \\ \mu(B) \end{bmatrix} \\
&= \mathcal{W}^T Q_z \mathcal{W}
\end{aligned} \tag{1.11.28}$$

Where $\mathcal{W} = \begin{bmatrix} \tau(A) \\ \mu(B) \end{bmatrix} \in \mathbb{R}^{k^2+2k(n-k)} = \mathbb{R}^{2nk-k^2}$. Meanwhile, again owing to the fact that τ and μ are isometries, we find that for $W \in T_{\pi(\hat{z})}(\mathring{S}^{k,0})$ we have $||W||_2 = ||\mathcal{W}||_2$. Thus returning to

our computation of $a(z)$

$$\begin{aligned} a(z) &= \min_{\substack{W \in T_{\pi(\hat{z})}(\hat{S}^{k,0}(\mathbb{C}^n)) \\ ||W||_2 = 1}} \sum_{j=1}^{m} |\langle W, A_j \rangle_{\mathbb{R}}|^2 \\ &= \min_{\substack{\mathcal{W} \in \mathbb{R}^{2nk-k^2} \\ ||\mathcal{W}||_2 = 1}} \mathcal{W}^T Q_z \mathcal{W} \\ &= \lambda_{2nk-k^2}(Q_z) \end{aligned} \tag{1.11.29}$$

This concludes the proof of $(i) - (iii)$. As for (iv) and (v) note that when rank$(x) \leqslant k$ then we may find $P \in U(r)$ such that $x = [\hat{x}|0]P$ for $\hat{x} \in \mathbb{C}^{n \times k}$ and moreover $d(x,z) = d(\hat{x},\hat{z})$ and $xx^* - zz^* = \hat{x}\hat{x}^* - \hat{z}\hat{z}^*$. Thus

$$\begin{aligned} \hat{a}_1(z) &= \lim_{R \to 0} \inf_{\substack{x \in \mathbb{C}^{n \times r} \\ d(z,x) < R \\ \text{rank}(x) \leqslant k}} \frac{\sum_{j=1}^{m} |\langle xx^* - zz^*, A_j \rangle_{\mathbb{R}}|^2}{d(x,z)^2} \\ &= \lim_{R \to 0} \inf_{\substack{\hat{x} \in \mathbb{C}^{n \times k} \\ d(\hat{x},\hat{z}) < R}} \frac{\sum_{j=1}^{m} |\langle \hat{x}\hat{x}^* - \hat{z}\hat{z}^*, A_j \rangle_{\mathbb{R}}|^2}{d(\hat{x},\hat{z})^2} \end{aligned} \tag{1.11.30}$$

The constraint rank$(x) \leqslant k$ is therefore equivalent to the assumption that $z \in \mathbb{C}_*^{n \times k}$. Hence, in order to avoid a plethora of hats we will assume $z \in \mathbb{C}_*^{n \times k}$. This assumption simplifies the situation considerably since in this case $\Delta_z = H_{\pi,z}$. As we shall see, if the Γ_z component of Δ_z were to be non-trivial, the local lower bounds $\hat{a}_1(z)$ and $\hat{a}_2(z)$ would be zero. We next note that $d(x,z) = ||x-z||_2||x+z||_2$ precisely when $x^*z = z^*x \geqslant 0$, which may be achieved without loss of generality in $\hat{a}_1(z)$ via choice of representative for x. Thus, keeping in mind that $z \in \mathbb{C}_*^{n \times k}$, we

find

$$\begin{aligned}\hat{a}_1(z) &= \lim_{R\to 0} \inf_{\substack{x\in\mathbb{C}^{n\times k}\\ d(z,x)<R}} \frac{\sum_{j=1}^m |\langle xx^* - zz^*, A_j\rangle_{\mathbb{R}}|^2}{d(x,z)^2} \\ &= \lim_{R\to 0} \inf_{\substack{x\in\mathbb{C}^{n\times k}\\ ||x-z||_2\cdot||x+z||_2<R\\ x^*z=z^*x\geqslant 0}} \frac{\sum_{j=1}^m |\langle z(x-z)^* + (x-z)z^* + (x-z)(x-z)^*, A_j\rangle_{\mathbb{R}}|^2}{||x-z||_2^2\cdot||x+z||_2^2}\end{aligned} \tag{1.11.31}$$

In analogy with our analysis of $a_1(z)$ we change variables from x to $w = x - z$ and are thus able to linearize the infimization constraint, since for $||w||_2 < \sigma_k(z)$ we have that $z^*(z+w) \geqslant 0$ if and only if $z^*w = w^*z$, or in other words if and only if $z \in \Delta_z \iff z \in H_{\pi,z}$ (the vertical component of Δ_z, namely Γ_z, is trivial for $z \in \mathbb{C}_*^{n\times k}$). We also exploit the fact that D and d generate the same topology and therefore instead of $||w||_2||2z+w||_2 < R$ we may simply take $||w||_2 < R$.

$$\begin{aligned}\hat{a}_1(z) &= \lim_{R\to 0} \inf_{\substack{w\in H_{\pi,z}\\ ||w||_2<R}} \frac{\sum_{j=1}^m |\langle zw^* + wz^* + ww^*, A_j\rangle_{\mathbb{R}}|^2}{||w||_2^2||2z+w||_2^2} \\ &= \frac{1}{4||z||_2^2} \lim_{R\to 0} \inf_{\substack{w\in H_{\pi,z}\\ ||w||_2<R}} \frac{1}{||w||_2^2} \sum_{j=1}^m |\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}}|^2(1+O(||w||_2^2)) \\ &= \frac{1}{4||z||_2^2} \inf_{\substack{w\in H_{\pi,z}\\ ||w||_2=1}} \sum_{j=1}^m |\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}}|^2 \\ &= \frac{1}{4||z||_2^2}\hat{a}(z)\end{aligned} \tag{1.11.32}$$

We now consider $\hat{a}_2(z)$. In a manner precisely analogous to (1.11.30) the constraint in $\hat{a}_2(z)$ that $\text{rank}(x) \leqslant k$ and $\text{rank}(y) \leqslant k$ is equivalent to the assumption that $z \in \mathbb{C}_*^{n\times k}$. We first employ the

unitary freedom of x and y to note that

$$\begin{aligned}
\hat{a}_2(z) &= \lim_{R\to 0} \inf_{\substack{x,y\in\mathbb{C}^{n\times k}\\ d(x,z)<R\\ d(y,z)<R}} \frac{\sum_{j=1}^m |\langle xx^* - yy^*, A_j\rangle_{\mathbb{R}}|^2}{d(x,y)^2} \\
&= \lim_{R\to 0} \inf_{\substack{x,y\in\mathbb{C}^{n\times k}\\ ||x-z||_2||x+z||_2<R\\ ||y-z||_2||y+z||_2<R\\ x^*z=z^*x\succcurlyeq 0\\ y^*z=z^*y\succcurlyeq 0}} \frac{\sum_{j=1}^m |\langle xx^* - yy^*, A_j\rangle_{\mathbb{R}}|^2}{d(x,y)^2} \\
&= \lim_{R\to 0} \inf_{\substack{x,y\in\mathbb{C}^{n\times k}\\ ||x-z||_2<R\\ ||y-z||_2<R\\ x^*z=z^*x\\ y^*z=z^*y}} \frac{\sum_{j=1}^m |\langle xx^* - yy^*, A_j\rangle_{\mathbb{R}}|^2}{d(x,y)^2}
\end{aligned} \tag{1.11.33}$$

We now weaken the infimization constraints and obtain a lower bound. We note that $x^*z = z^*x$ and $y^*z = z^*y$ taken together imply that $(x-y)^*z = z^*(x-y)$, and also that the denominator $d(x,y)^2 \leqslant ||x-y||_2^2||x+y||_2^2$. Thus, changing variables to $\xi = x - z$ and $\eta = y - z$ we obtain

$$\begin{aligned}
\hat{a}_2(z) &\geqslant \lim_{R\to 0} \inf_{\substack{\xi,\eta\in\mathbb{C}^{n\times k}\\ ||\xi||_2<R\\ ||\eta||_2<R\\ z^*(\xi-\eta)=(\xi-\eta)^*z}} \frac{\sum_{j=1}^m |\langle z(\xi-\eta)^* + (\xi-\eta)z^* + \xi\xi^* - \eta\eta^*, A_j\rangle_{\mathbb{R}}|^2}{||\xi-\eta||_2^2||2z+\xi+\eta||_2^2} \\
&= \frac{1}{4||z||_2^2} \lim_{R\to 0} \inf_{\substack{\xi,\eta\in\mathbb{C}^{n\times k}\\ ||\xi||_2<R\\ ||\eta||_2<R\\ z^*(\xi-\eta)=(\xi-\eta)^*z}} \frac{\sum_{j=1}^m |\langle z(\xi-\eta)^* + (\xi-\eta)z^*, A_j\rangle_{\mathbb{R}}|^2}{||\xi-\eta||_2^2}(1+O(||\xi||_2^2+||\eta||_2^2)) \\
&= \frac{1}{4||z||_2^2} \lim_{R\to 0} \inf_{\substack{\xi,\eta\in\mathbb{C}^{n\times k}\\ ||\xi||_2<R\\ ||\eta||_2<R\\ z^*(\xi-\eta)=(\xi-\eta)^*z}} \frac{\sum_{j=1}^m |\langle z(\xi-\eta)^* + (\xi-\eta)z^*, A_j\rangle_{\mathbb{R}}|^2}{||\xi-\eta||_2^2} \\
&= \frac{1}{4||z||_2^2} \lim_{R\to 0} \inf_{\substack{\xi,\eta\in\mathbb{C}^{n\times k}\\ ||\xi-\eta||_2<2R\\ z^*(\xi-\eta)=(\xi-\eta)^*z}} \frac{\sum_{j=1}^m |\langle z(\xi-\eta)^* + (\xi-\eta)z^*, A_j\rangle_{\mathbb{R}}|^2}{||\xi-\eta||_2^2}
\end{aligned} \tag{1.11.34}$$

The last line is an equality rather than an inequality owing to homogeneity in $\xi - \eta$. Changing variables once more to $w = \xi - \eta$ and using the fact that for $z \in \mathbb{C}_*^{n\times k}$ $z^*w = w^*z \iff w \in \Delta_z \iff w \in H_{\pi,z}(\mathbb{C}_*^{n\times k})$ gives

$$\begin{aligned}\hat{a}_2(z) &\geqslant \frac{1}{4||z||_2^2} \lim_{R\to 0} \inf_{\substack{w\in H_{\pi,z}(\mathbb{C}_*^{n\times k}) \\ ||w||_2<2R}} \frac{\sum_{j=1}^m |\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}}|^2}{||w||_2^2} \\ &= \frac{1}{4||z||_2^2} \inf_{\substack{w\in H_{\pi,z}(\mathbb{C}_*^{n\times k}) \\ ||w||_2=1}} \sum_{j=1}^m |\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}}|^2 \\ &= \hat{a}(z) = \hat{a}_1(z)\end{aligned} \tag{1.11.35}$$

The reverse inequality $\hat{a}_2(z) \leqslant \hat{a}_1(z)$ is immediate from the definitions of $\hat{a}_1(z)$ and $\hat{a}_2(z)$, thus (1.8.15) is proved. We now turn to explicit computation of $\hat{a}(z)$ as the smallest non-zero eigenvalue of $\hat{Q}_z$. As with the computation of $a(z)$ we rely on several embeddings. Specifically we define

$$\begin{aligned} l &: \mathbb{C}^{n\times k} \to \mathbb{R}^{2n\times k} & j &: \mathbb{C}^{n\times k} \to \mathbb{R}^{2n\times 2k} \\ l(X) &= \begin{bmatrix} \Re X \\ \Im X \end{bmatrix} & j(X) &= \begin{bmatrix} \Re X & -\Im X \\ \Im X & \Re X \end{bmatrix} \end{aligned} \tag{1.11.36}$$

Note that j is an injective homomorphism and moreover that

$$j(X) = \begin{bmatrix} l(X) & Jl(X) \end{bmatrix} \tag{1.11.37}$$

where $J \in \mathbb{R}^{2n\times 2n}$ is the symplectic form

$$J = \begin{bmatrix} 0 & -\mathbb{I}_{n\times n} \\ \mathbb{I}_{n\times n} & 0 \end{bmatrix} \tag{1.11.38}$$

Note that $Jj(X) = j(X)J$ for all $X \in \mathbb{C}^{n\times n}$.The embedding l is isometric, and the embedding j is isometric up to a constant since for $X, Y \in \mathbb{C}^{n\times k}$ we have $\langle X, Y\rangle_{\mathbb{R}} = \langle l(X), l(Y)\rangle_{\mathbb{R}} = \frac{1}{2}\langle j(X), j(Y)\rangle_{\mathbb{R}}$. The embedding j is furthermore a structure preserving homomorphism since for $p \in \mathbb{C}^{n\times k}, q \in \mathbb{C}^{k\times l}$ we have that $j(p)l(q) = l(pq)$, $j(pq) = j(p)j(q)$, and $j(p^*) = j(p)^T$. We will also employ the isometric embedding vec defined in the obvious way in (1.8.8). We will need the fact that if $A \in \mathbb{R}^{n\times k}$ and $B \in \mathbb{R}^{k\times l}$ then

$$\text{vec}(AB) = (\mathbb{I}_{l\times l} \otimes A)\text{vec}(B) \tag{1.11.39}$$

Note that this further implies that for $x, y \in \mathbb{R}^{n\times k}$ and $F \in \mathbb{R}^{n\times n}$ we have that

$$\text{vec}(x)^T(\mathbb{I}_{k\times k} \otimes F)\text{vec}(y) = \text{vec}(x)^T\text{vec}(Fy) = \langle x, Fy\rangle_{\mathbb{R}} = \text{tr}\{x^T Fy\} \tag{1.11.40}$$

With this in mind we find that for $z \in \mathbb{C}_*^{n\times k}$ and $w \in H_{\pi,z}(\mathbb{C}_*^{n\times k})$

$$\begin{aligned}
|\langle D\pi(z)(w), A_j\rangle_{\mathbb{R}}|^2 &= 4|\langle wz^*, A_j\rangle_{\mathbb{R}}|^2 \\
&= \langle j(wz^*), A_j\rangle^2 \\
&= \langle j(w), A_j j(z)\rangle^2 \\
&= \Big(\text{vec}(j(w))^T \text{vec}(j(A_j)j(z))\Big)^2 \\
&= \Big(\text{vec}(j(w))^T (\mathbb{I}_{2k\times 2k} \otimes j(A_j))\text{vec}(j(z))\Big)^2 \\
&= 4\Big(\text{vec}(l(w))^T (\mathbb{I}_{k\times k} \otimes j(A_j))\text{vec}(l(z))\Big)^2 \\
&= 4W^T F_j Z Z^T F_j W
\end{aligned} \tag{1.11.41}$$

where $W = \mu(w)$, $Z = \mu(z)$ and $F_j = \mathbb{I}_{k\times k} \otimes j(A_j)$. This should not be too surprising since in fact

$$\begin{aligned}
\beta_j(z) &= \langle zz^*, A_j\rangle_{\mathbb{R}} \\
&= \langle z, A_j z\rangle_{\mathbb{R}} \\
&= \frac{1}{2}\langle j(z), j(A_j)j(z)\rangle \\
&= \frac{1}{2}\text{vec}(j(z))^T \text{vec}(j(A_j)j(z)) \\
&= \frac{1}{2}\text{vec}(j(z))^T (\mathbb{I}_{2k\times 2k} \otimes j(A_j))\text{vec}(j(z)) \\
&= \text{vec}(l(z))^T (\mathbb{I}_{k\times k} \otimes j(A_j))\text{vec}(l(z)) = Z^T F_j Z
\end{aligned} \tag{1.11.42}$$

Thus when β_j is viewed as map from $\mathbb{R}^{2nk}$ to $\mathbb{R}$ we find that $|D\beta_j(Z)(W)|^2 = 4W^T F_j Z Z^T F_j W$. Returning to $a(z)$ we first note that the constraint $w \in H_{\pi,z}(\mathbb{C}_*^{n\times k})$ precisely avoids the "trivial" kernel of dimension k^2 common to each $F_j Z Z^T F_j$. Specifically, we note that $Z^T F_j V = 0$ for

$V \in \mathcal{V}_z \subset \mathbb{R}^{2nk}$ where

$$\mathcal{V}_z = \{\text{vec}(Jl(z)S + l(z)A) | S \in \text{Sym}(\mathbb{R}^k), A \in \text{Asym}(\mathbb{R}^k)\} \tag{1.11.43}$$

Namely if $V \in \mathcal{V}_z$ and $\eta = Jl(z)S + l(z)A \in \mathbb{R}^{2n\times r}$ for $A \in \text{Asym}(\mathbb{R}^k)$ and $S \in \text{Sym}(\mathbb{R}^k)$ so that $V = \text{vec}(\eta)$ then

$$\begin{aligned} Z^T F_j V &= \text{vec}(l(z))^T (\mathbb{I}_{k\times k} \otimes j(A_j))\text{vec}(\eta) \\ &= \text{tr}\{l(z)^T j(A_j)\eta\} \\ &= \text{tr}\{l(z)^T j(A_j)(Jl(z)S + l(z)A)\} \\ &= \text{tr}\{l(z)^T j(A_j)Jl(z)S\} + \text{tr}\{l(z)^T j(A_j)l(z)A\} \\ &= 0 \end{aligned} \tag{1.11.44}$$

The last line follows from the fact that $j(A_j)$ is symmetric and $j(A_j)J$ is anti-symmetric since $(j(A_j)J)^* = -Jj(A_j) = -j(A_j)J$. The reason that $w \in H_{\pi,z}(\mathbb{C}_*^{n\times k})$ avoids this common kernel is that in fact $\mathcal{V}_z = \mu(V_{\pi,z}(\mathbb{C}_*^{n\times k}))$. Recall that

$$V_{\pi,z}(\mathbb{C}_*^{n\times k}) = \{zK | K \in \text{Asym}(\mathbb{C}^k)\} \tag{1.11.45}$$

We may decompose $K \in \text{Asym}(\mathbb{C}^n)$ as $K = A + iS$ where $A \in \text{Asym}(\mathbb{R}^n)$ and $S \in \text{Sym}(\mathbb{R}^n)$.

Hence if $u \in V_{\pi,z}(\mathbb{C}_*^{n\times k})$ then on the one hand $j(u) = [l(u)|Jl(u)]$ and on the other

$$j(u) = j(zK) = j(z)j(K) = [l(z)|Jl(z)] \begin{bmatrix} A & -S \\ S & A \end{bmatrix} = [l(z)A + Jl(z)S| - l(z)S + Jl(z)A] \tag{1.11.46}$$

From which we may clearly identify $l(u) = l(z)A + Jl(z)S$, thus

$$\mathcal{V}_z = \{\mu(u)|u \in V_{\pi,z}(\mathbb{C}_*^{n\times k})\} \tag{1.11.47}$$

The map μ is an isometry, so if $w \in H_{\pi,z}(\mathbb{C}_*^{n\times k})$ then the image $W = \mu(w)$ lies precisely in the orthogonal complement of $\mathcal{V}_z$. Thus

$$\begin{aligned} \hat{a}(z) &= \min_{\substack{w\in H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k}) \\ ||w||_2=1}} \sum_{j=1}^{m} |\langle D\pi(\hat{z})(w), A_j\rangle_{\mathbb{R}}|^2 \\ &= \min_{\substack{W\in\mathbb{R}^{2nk} \\ W\perp\mathcal{V}_z \\ ||W||_2=1}} W^T(4\sum_{j=1}^{m} F_j Z Z^T F_j)W \\ &= \lambda_{2nk-k^2}(\hat{Q}_z) \end{aligned} \tag{1.11.48}$$

Note that at this point the hats return and $Z = \mu(\hat{z})$. Eigenvalues are continuous with respect to matrix entries, and $\hat{Q}_z$ is manifestly continuous with respect to z. As a result of this and the fact that $k \mapsto 2nk - k^2$ is monotone increasing for $k \leqslant n$ we conclude that $\hat{a}(z)$ approaches zero whenever z approaches a drop in rank. Indeed, $\hat{a}(z)$ jumps discontinuously to a non-zero value once the surface of lower rank is actually reached, but this cannot prevent $\inf_{z\in\mathbb{C}^{n\times r}} \hat{a}(z)$ from

being zero, thus there is no hope of defining a non-zero global lower bound $\hat{a}_0$. This concludes the proof of claims (iv)-(vi).

Claim (vii) gives local control of $a(z)$ in terms of $\hat{a}(z)$. We first prove that the the inequality (1.8.17) holds. To do so we consider the following operators:

$$\begin{aligned} &\Pi_1(\hat{z}) : (T_{\pi(\hat{z})}(\mathring{S}^{k,0}(\mathbb{C}^n)), ||\cdot||_2) \to (\mathbb{R}^m, ||\cdot||_2) \\ &\Pi_1(\hat{z})(W) = (\text{tr}\{WA_j\})_{j=1}^m \end{aligned} \tag{1.11.49}$$

$$\begin{aligned} &\Pi_2(\hat{z}) : (H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k}), ||\cdot||_2) \to (\mathbb{R}^m, ||\cdot||_2) \\ &\Pi_2(\hat{z})(w) = (\text{tr}\{(\hat{z}w^* + w\hat{z}^*)A_j\})_{j=1}^m = \Pi_1(\hat{z})D\pi(\hat{z})w \end{aligned} \tag{1.11.50}$$

Note that $a(z)$ and $\hat{a}(z)$, defined respectively in (1.8.3) and (1.8.4), are expressible in terms of the operator norms of the pseudo-inverses of $\Pi_1(\hat{z})$ and $\Pi_2(\hat{z})$.

$$\begin{aligned} a(z) &= ||\Pi_1(\hat{z})^\dagger||_*^{-2} \\ \hat{a}(z) &= ||\Pi_2(\hat{z})^\dagger||_*^{-2} \end{aligned} \tag{1.11.51}$$

We may therefore obtain operator-theoretic inequalities relating $a(z)$ and $\hat{a}(z)$, namely

$$\begin{aligned} ||\Pi_2(\hat{z})^\dagger||_* &= ||D\pi(\hat{z})^{-1}\Pi_1(\hat{z})^\dagger||_* \leqslant ||D\pi(\hat{z})^{-1}||_*||\Pi_1(\hat{z})^\dagger||_* \\ ||\Pi_1(\hat{z})^\dagger||_* &= ||D\pi(\hat{z})\Pi_2(\hat{z})^\dagger||_* \leqslant ||D\pi(\hat{z})||_*||\Pi_2(\hat{z})^\dagger||_* \end{aligned} \tag{1.11.52}$$

Hence

$$||D\pi(\hat{z})||_*^{-2}\hat{a}(z) \leqslant a(z) \leqslant ||D\pi(\hat{z})^{-1}||_*^2\hat{a}(z) \tag{1.11.53}$$

It remains only to compute appropriate bounds for $||D\pi(\hat{z})||_*^{-2}$ and $||D\pi(z)^{-1}||_*^2$ in order to prove (1.8.17). First note that

$$||D\pi(\hat{z})^{-1}||_*^2 = \sup_{W\in\mathbb{T}_{\pi(\hat{z})}(\hat{S}^{k,0}(\mathbb{C}^n))\setminus\{0\}} \frac{||D\pi(\hat{z})^{-1}(W)||_2^2}{||W||_2^2} = \left(\inf_{w\in H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k})\setminus\{0\}} \frac{||\hat{z}w^* + w\hat{z}^*||_2^2}{||w||_2^2}\right)^{-1} \tag{1.11.54}$$

Next note that for $w = H\hat{z} + X \in H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k})$ we have $||w||_2^2 = ||H\hat{z}||_2^2 + ||X||_2^2$ and $||\hat{z}w^* + w\hat{z}||_2^2 = 2(||\hat{z}^*H\hat{z}||_2^2 + ||\hat{z}\hat{z}^*H||_2^2 + ||\hat{z}X^*||_2^2)$ thus

$$\begin{aligned}
||D\pi(\hat{z})^{-1}||_*^{-2} &= \inf_{w\in H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k})\setminus\{0\}} \frac{||\hat{z}w^* + w\hat{z}^*||_2^2}{||w||_2^2} \\
&= 2 \inf_{\substack{H\in \mathbf{Sym}(\mathbb{C}^n),\mathbb{P}_{\mathbf{Ran}_{(\hat{z})}}H=H \\ X\in\mathbb{C}^{n\times k},\mathbb{P}_{\mathbf{Ran}_{(\hat{z})}}X=0}} \frac{||\hat{z}^*H\hat{z}||_2^2 + ||\hat{z}\hat{z}^*H||_2^2 + ||\hat{z}X^*||_2^2}{||H\hat{z}||_2^2 + ||X||_2^2} \\
&\geqslant 2 \inf_{\substack{H\in \mathbf{Sym}(\mathbb{C}^n),\mathbb{P}_{\mathbf{Ran}_{(\hat{z})}}H=H \\ X\in\mathbb{C}^{n\times k},\mathbb{P}_{\mathbf{Ran}_{(\hat{z})}}X=0}} \frac{||\hat{z}^*H\hat{z}||_2^2 + ||\hat{z}X^*||_2^2}{||H\hat{z}||_2^2 + ||X||_2^2} \\
&\geqslant 2\sigma_k(\hat{z})^2 \inf_{\substack{H\in \mathbf{Sym}(\mathbb{C}^n),\mathbb{P}_{\mathbf{Ran}_{(\hat{z})}}H=H \\ X\in\mathbb{C}^{n\times k},\mathbb{P}_{\mathbf{Ran}_{(\hat{z})}}X=0}} \frac{||H\hat{z}||_2^2 + ||X||_2^2}{||H\hat{z}||_2^2 + ||X||_2^2} \\
&= 2\sigma_k(z)^2
\end{aligned} \tag{1.11.55}$$

Hence $||D\pi(\hat{z})^{-1}||_*^2 \leqslant \frac{1}{2\sigma_k(z)^2}$. For the opposing bound note that

$$\begin{aligned} ||D\pi(\hat{z})||_*^2 &= \sup_{w \in H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k})\setminus\{0\}} \frac{||\hat{z}w^* + w\hat{z}^*||_2^2}{||w||_2^2} \\ &\leqslant \sup_{w \in H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k})\setminus\{0\}} \frac{||\hat{z}w^* + w\hat{z}^*||_1^2}{||w||_2^2} \\ &\leqslant \sup_{w \in H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k})\setminus\{0\}} \frac{4||\hat{z}w^*||_1^2}{||w||_2^2} \\ &\leqslant 4||z||_2^2 \end{aligned} \tag{1.11.56}$$

Hence $||D\pi(\hat{z})||_*^{-2} \geqslant \frac{1}{4||z||_2^2}$, proving (1.8.17). We note that choosing $w = \hat{z} \in H_{\pi,\hat{z}}(\mathbb{C}_*^{n\times k})$ proves that in fact $||D\pi(\hat{z})||_{2\to 1} = \frac{1}{2||z||_2}$. Finally, the claimed bounds in (1.8.17) are tight in the case $\text{rank}(z) = 1$, since in this case the inequality is equivalent to the norm inequality for $W \in \mathbb{C}^{n\times n}$

$$\frac{1}{\sqrt{\text{rank}(W)}}||W||_1 \leqslant ||W||_2 \leqslant ||W||_1 \tag{1.11.57}$$

Specifically if $W \in T_{\pi(z)}(\mathring{S}^{1,0}(\mathbb{C}^n))$ for $z \in \mathbb{C}_*^n$ then $W = zw^* + wz^*$ for some $w \in H_{\pi,z}(\mathbb{C}_*^n) \subset \mathbb{C}^n$ and has rank at most 2. Moreover we have that

$$||W||_1 = ||zw^* + wz^*||_1 = \frac{1}{2}||(z+w)(z+w)^* - (z-w)(z-w)^*||_1 \tag{1.11.58}$$

Recall (1.6.8) that for $x, y \in \mathbb{C}^n$ we have that $||xx^* - yy^*||_1 = d(x,y)$ and that $d(x,y) = ||x-y||_2||x+y||_2$ when $x^*y \geqslant 0$. Let $x = z + w$ and $y = z - w$, and note that in this case $w \in H_{\pi,z}(\mathbb{C}_*^n)$ implies $x^*y = z^*z + w^*z - z^*w - w^*w = z^*z - w^*w \geqslant 0$ for $||w||_2$ sufficiently

small. Thus for $||w||_2$ or equivalently $||W||_2$ sufficiently small,

$$||W||_1 = \frac{1}{2}||(z+w)-(z-w)||_2||(z+w)+(z-w)||_2 = 2||z||_2||w||_2 \tag{1.11.59}$$

The condition that $||W||_2$ be sufficiently small is of no issue since the ratio in $a(z)$ is homogeneous in $||W||_2$, hence recalling that $\text{rank}(W) \leqslant 2$ (1.11.57) implies

$$\sqrt{2}||z||_2||w||_2 \leqslant ||W||_2 \leqslant 2||z||_2||w||_2 \tag{1.11.60}$$

Thus for $\text{rank}(z) = 1$ the inequality (1.11.57) is equivalent to

$$\frac{1}{4||z||_2^2}\hat{a}(z) \leqslant a(z) \leqslant \frac{1}{2||z||_2^2}\hat{a}(z) \tag{1.11.61}$$

which is recognizable as (1.8.17) since if $\text{rank}(z) = 1$ then $||z||_2^2 = \sigma_1(z)^2$ and hence since (1.11.57) is tight so too is (1.8.17). This concludes the proof of (vii).

To prove $(viii)$ we combine (1.8.11) and (1.8.14) to obtain the following formula for computing a_0:

$$a_0 = \min_{k=1,\ldots,r} \min_{\substack{U\in U(n) \\ U=[U_1|U_2] \\ U_1\in\mathbb{C}^{n\times k} \\ U_2\in\mathbb{C}^{n\times(n-k)}}} \lambda_{2nk-k^2}(Q_U) \tag{1.11.62}$$

Recalling that

$$Q_{[U_1|U_2]} = \sum_{j=1}^{m} \begin{bmatrix} \tau(U_1^* A_j U_1) \\ \\ \mu(U_2^* A_j U_1) \end{bmatrix} \begin{bmatrix} \tau(U_1^* A_j U_1) \\ \\ \mu(U_2^* A_j U_1) \end{bmatrix}^T \tag{1.11.63}$$

Finally, we need to prove that the minimum over k in fact occurs at $k = r$. We may write

$$a_0 = \min_{k=1,\dots,r} \inf_{z \in \mathbb{C}_*^{n\times k}} \min_{W \in T_{\pi(z)}(\mathring{S}^{k,0}(\mathbb{C}^n))} \frac{1}{||W||_2^2} \sum_{j=1}^{m} |\langle W, A_j \rangle_{\mathbb{R}}|^2 \tag{1.11.64}$$

Then note that if $\hat{z} \in \mathbb{C}_*^{n\times k}$ and $\tilde{z} \in \mathbb{C}_*^{n\times(r-k)}$ is such that $\hat{z}^*\tilde{z} = 0$ then $z = [\hat{z}|\tilde{z}] \in \mathbb{C}_*^{n\times r}$ and moreover, recalling the parametrization of the tangent space (1.7.7) (or alternately that the stratification is a-regular), we find that $T_{\pi(z)}(\mathring{S}^{r,0}(\mathbb{C}^n)) \supset T_{\pi(\hat{z})}(\mathring{S}^{k,0}(\mathbb{C}^n))$ since $\text{Ran}(z)^\perp = \text{Ran}(\hat{z})^\perp \cap \text{Ran}(\tilde{z})^\perp$. Thus, in fact

$$a_0 = \min_{\substack{U \in U(n) \\ U=[U_1|U_2] \\ U_1 \in \mathbb{C}^{n\times r} \\ U_2 \in \mathbb{C}^{n\times(n-r)}}} \lambda_{2nr-r^2}(Q_U) \tag{1.11.65}$$

We now set out to prove (ix), specifically to control a_0 using an infimization of $\hat{a}(z)$ rather than of $a(z)$ by including the additional constraint that $z^*z = \mathbb{I}_{r\times r}$. With this constraint we may write any $w \in H_{\pi,z}(\mathbb{C}_*^{n\times r})$ as $w = z\tilde{H} + X$ where $\tilde{H} \in \text{Sym}(\mathbb{C}^r)$ and $X \in \mathbb{C}^{n\times r}$ satisfies $\mathbb{P}_{\text{Ran}(z)}X = 0$ (equivalently X satisfies $z^*X = 0$). We note that for z satisfying the constraint

$$||w||_2^2 = ||\tilde{H}||_2^2 + ||X||_2^2 \tag{1.11.66}$$

$$||zw^* + wz^*||_2^2 = 4||\tilde{H}||_2^2 + 2||X||_2^2 \tag{1.11.67}$$

Hence referring to (1.8.3) and (1.8.4) we find that for $z^*z = \mathbb{I}_{r\times r}$

$$\frac{1}{4}\hat{a}(z) \leqslant a(z) \leqslant \frac{1}{2}\hat{a}(z) \tag{1.11.68}$$

Note that a direct application of (1.8.17) to the case where z has orthonormal columns would lead to the lower constant being $\frac{1}{4r}$ rather than $\frac{1}{4}$.The form (1.8.18) for a_0 tells us that $a(z)$ depends only on the range of z, and that we may obtain a_0 via

$$a_0 = \inf_{\substack{z\in\mathbb{C}_*^{n\times r}\\ z^*z=\mathbb{I}_{r\times r}}} a(z) \tag{1.11.69}$$

Thus

$$\frac{1}{4}\inf_{\substack{z\in\mathbb{C}_*^{n\times r}\\ z^*z=\mathbb{I}_{r\times r}}} \hat{a}(z) \leqslant a_0 \leqslant \frac{1}{2}\inf_{\substack{z\in\mathbb{C}_*^{n\times r}\\ z^*z=\mathbb{I}_{r\times r}}} \hat{a}(z) \tag{1.11.70}$$

This concludes the proof of (ix) and Theorem 1.8.5. □

Remark 1.11.3. For $r = 1$ the inequality (1.8.17) tells us that

$$\frac{1}{4||z||_2^2}\hat{a}(z) \leqslant a(z) \leqslant \frac{1}{2||z||_2^2}\hat{a}(z) \tag{1.11.71}$$

But in fact, as was proved in [23], more is true. Namely if the nuclear norm is used in the definition of a_0 instead of the Frobenius norm so that

$$a_0^1 = \inf_{\substack{x,y\in\mathbb{C}^{n\times r}\\ x\neq y}} \frac{\sum_{j=1}^m(\langle xx^*, A_j\rangle_{\mathbb{R}} - \langle yy^*, A_j\rangle_{\mathbb{R}})^2}{||xx^* - yy^*||_1^2} \tag{1.11.72}$$

And similarly in the definition of $a(z)$ so that

$$a^1(z) = \min_{\substack{W \in T_{\pi(\hat{z})}(\hat{S}^{k,0}(\mathbb{C}^n)) \\ ||W||_1 = 1}} \sum_{j=1}^{m} |\langle W, A_j \rangle_{\mathbb{R}}|^2 \tag{1.11.73}$$

then

$$a_0^1 = \inf_{z \in \mathbb{C}^{n \times r} \setminus \{0\}} a^1(z) \tag{1.11.74}$$

$$a^1(z) = \frac{1}{4||z||_2^2} \hat{a}(z) \tag{1.11.75}$$

Remark 1.11.4. For $r = 1$, Q_z is orthogonally equivalent to the restriction of $\hat{Q}_z$ to the orthogonal complement of its null space, giving a correspondence between (1.8.14) and (3.5) in [40] when the frame is positive semidefinite ($A_j = f_j f_j^*$). Specifically, if $r = 1$ then we may take $U_1 = \frac{z}{||z||_2} =: e_1$ and $U_2 = [e_2, \ldots, e_n]$ where $e_1, \ldots, e_n$ forms an orthonormal basis for $\mathbb{C}^n$ with respect to the complex inner product $\langle \cdot, \cdot \rangle_{\mathbb{C}}$. Thus

$$\begin{aligned} \tau(U_1^* A_j U_1) &= \frac{|\langle z, f_j \rangle_{\mathbb{C}}|^2}{||z||_2^2} = \frac{1}{||z||_2} \langle e_1, f_j \rangle_{\mathbb{C}} \langle f_j, z \rangle_{\mathbb{C}} \\ \mu(U_2^* A_j U_1) &= \frac{1}{||z||_2} l\left(\begin{bmatrix} \langle e_2, f_j \rangle_{\mathbb{C}} \langle f_j, z \rangle_{\mathbb{C}} \\ \vdots \\ \langle e_n, f_j \rangle_{\mathbb{C}} \langle f_j, z \rangle_{\mathbb{C}} \end{bmatrix} \right) \end{aligned} \tag{1.11.76}$$

Note that $\tau(U_1^* A_j U_1)$ is real, hence if we insert a single 0 in the middle of $\mu(U_2^* A_j U_1)$ between

vec($\Re(U_2^* A_j U_1)$) and vec($\Im(U_2^* A_j U_1)$) we obtain

$$\begin{bmatrix} \tau(U_1^* A_j U_1) \\ \text{vec}(\Re(U_2^* A_j U_1)) \\ 0 \\ \text{vec}(\Im(U_2^* A_j U_1)) \end{bmatrix} = \frac{1}{||z||_2} l\left(\begin{bmatrix} \langle e_1, f_j \rangle_{\mathbb{C}} \langle f_j, z \rangle_{\mathbb{C}} \\ \vdots \\ \langle e_n, f_j \rangle_{\mathbb{C}} \langle f_j, z \rangle_{\mathbb{C}} \end{bmatrix} \right) = \frac{1}{||z||_2} l(U^* A_j z) = \frac{1}{||z||_2} j(U)^T j(A_j) l(z) \tag{1.11.77}$$

Where in the last inequality the algebraic properties of l and j are employed. Thus (up to a row and column of zeros)

$$Q_z = j(U)^T \left\{ \frac{1}{||z||_2^2} \sum_{j=1}^{m} j(A_j) l(z) l(z)^T j(A_j) \right\} j(U) \tag{1.11.78}$$

In accordance with the notation of [40] we denote $\xi = l(z)$, $\phi_j = l(f_j)$, and $\Phi_j = j(A_j) = \phi_j \phi_j^T + J \phi_j \phi_j^T J^T$ so that the above becomes

$$Q_z = j(U)^T \left\{ \frac{1}{||\xi||_2^2} \sum_{j=1}^{m} \Phi_j \xi \xi^T \Phi_j \right\} j(U) \tag{1.11.79}$$

Finally note that the column of $j(U)$ corresponding to the the row and column of zeros on the left hand side is $Jl(z)/||z||_2 = J\xi/||\xi||_2$, thus if we multiply on the left by $j(U)$ and on the right by

$j(U)^T$ we obtain

$$j(U)Q_z j(U)^T = (\mathbb{I} - \mathbb{P}_{J\xi})\left\{ \frac{1}{||\xi||_2^2} \sum_{j=1}^{m} \Phi_j \xi \xi^T \Phi_j \right\} (\mathbb{I} - \mathbb{P}_{J\xi}) \tag{1.11.80}$$

1.11.3 Proof of Theorem 1.8.8

Proof. As was the case for $\hat{a}_1(z)$ and $\hat{a}_2(z)$ the rank constraints in $A_1(z)$, $A_2(z)$, $\hat{A}_1(z)$, and $\hat{A}_2(z)$ allow us to assume that $z \in \mathbb{C}_*^{n\times k}$ rather than $\mathbb{C}^{n\times r}$. As before, this is done because without this assumption the resulting lower bounds would be zero for every z not full rank. We begin with the analysis of $\hat{A}_1(z)$, the simpler of the local lower bounds (we will show (x) that $A_i(z)$ differ from $\hat{A}_i(z)$ only by a constant factor, and hence will not analyze them separately). As we have done several times before we will employ the right hand unitary freedom of the variable x to require that $z^*x \geqslant 0$, and then make the change of variables from x to $w = x - z$.

$$\begin{aligned}
\hat{A}_1(z) &= \lim_{R\to 0} \inf_{\substack{x\in\mathbb{C}^{n\times k} \\ xx^* \neq zz^* \\ D(x,z)<R}} \frac{1}{D(x,z)^2} \sum_{j=1}^{m} |\langle xx^*, A_j\rangle^{\frac{1}{2}} - \langle zz^*, A_j\rangle^{\frac{1}{2}}|^2 \\
&= \lim_{R\to 0} \inf_{\substack{w\in\mathbb{C}^{n\times k} \\ zw^*+wz^*+ww^*\neq 0 \\ ||w||_2<R \\ z^*(z+w)\geqslant 0}} \frac{1}{||w||_2^2} \sum_{j=1}^{m} |\langle (z+w)(z+w)^*, A_j\rangle^{\frac{1}{2}} - \langle zz^*, A_j\rangle^{\frac{1}{2}}|^2 \\
&= \lim_{R\to 0} \inf_{\substack{w\in\mathbb{C}^{n\times k} \\ zw^*+wz^*+ww^*\neq 0 \\ ||w||_2<R \\ w\in\Delta_z}} \frac{1}{||w||_2^2} \left\{ \sum_{j\in I_0(z)} \langle ww^*, A_j\rangle_{\mathbb{R}} + \sum_{j\in I(z)} \frac{|\langle zw^*+wz^*+ww^*, A_j\rangle_{\mathbb{R}}|^2}{|\langle (z+w)(z+w)^*, A_j\rangle^{\frac{1}{2}} + \langle zz^*, A_j\rangle^{\frac{1}{2}}|^2} \right\}
\end{aligned} \tag{1.11.81}$$

Where $I_0(z) = \{j \in \{1,\ldots,m\} | \alpha_j(z) = 0\}$ are the indices for which α_j is zero (and hence not differentiable) and $I(z) = \{j \in \{1,\ldots,m\} | \alpha_j(z) \neq 0\}$ are the indices for which α_j is not zero

(and hence is differentiable). Thus, since z is full rank we know that $\Delta_z = H_{\pi,z}(\mathbb{C}_*^{n\times k})$ and since $zw^* + wz^* + ww^* \neq 0 \iff w \neq 0$ for $w \in H_{\pi,z}(\mathbb{C}_*^{n\times k})$ and sufficiently small in norm, we obtain

$$\begin{aligned}\hat{A}_1(z) &= \lim_{R\to 0} \inf_{\substack{w\in H_{\pi,z}(\mathbb{C}_*^{n\times k}) \\ 0<||w||_2<R}} \frac{1}{||w||_2^2}\Big\{ \sum_{j\in I_0(z)} \langle ww^*, A_j\rangle_{\mathbb{R}} + \sum_{j\in I(z)} \frac{|\langle zw^* + wz^* + ww^*, A_j\rangle_{\mathbb{R}}|^2}{|\langle (z+w)(z+w)^*, A_j\rangle^{\frac{1}{2}} + \langle zz^*, A_j\rangle^{\frac{1}{2}}|^2}\Big\} \\ &= \lim_{R\to 0} \inf_{\substack{w\in H_{\pi,z}(\mathbb{C}_*^{n\times k}) \\ 0<||w||_2<R}} \frac{1}{||w||_2^2}\Big\{ \sum_{j\in I_0(z)} \langle ww^*, A_j\rangle_{\mathbb{R}} + \sum_{j\in I(z)} \frac{|\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}}|^2}{4\langle zz^*, A_j\rangle} + O(||w||^3)\Big\} \\ &= \min_{\substack{w\in H_{\pi,z}(\mathbb{C}_*^{n\times k}) \\ ||w||_2=1}} \frac{1}{||w||_2^2}\Big\{ \sum_{j\in I_0(z)} \langle ww^*, A_j\rangle_{\mathbb{R}} + \sum_{j\in I(z)} \frac{|\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}}|^2}{4\langle zz^*, A_j\rangle}\Big\}\end{aligned} \tag{1.11.82}$$

Now recall from (1.11.41) and (1.11.42) respectively that $|\langle zw^*+wz^*, A_j\rangle_{\mathbb{R}}|^2 = |\langle D\pi(z)(w), A_j\rangle_{\mathbb{R}}|^2 = 4W^TF_jZZ^TF_jW$ and $\langle ww^*, A_j\rangle = \beta_j(w) = W^TF_jW$. Thus the above is

$$\hat{A}_1(z) = \min_{\substack{W\in\mathbb{R}^{2nk} \\ W\perp\mathcal{V}_z \\ ||W||_2=1}} W^T\Big\{ \sum_{j\in I_0(z)} F_j + \sum_{j\in I(z)} \frac{F_jZZ^TF_j}{Z^TF_jZ}\Big\}W \tag{1.11.83}$$

As has already been noted in (1.11.44) the null space of each $F_jZZ^TF_j$ contains $\mathcal{V}_z$, but in fact so does the null space of each F_j for $j \in I_0(z)$ since in this case $F_j\mu(zK) = (\mathbb{I}_{k\times k} \otimes j(A_j))\text{vec}(l(zK)) = \text{vec}(j(A_j)l(zk)) = \text{vec}(l(A_jzK)) = 0$. Thus we obtain finally that

$$\hat{A}_1(z) = \lambda_{2nk-k^2}\Big(\sum_{j\in I_0(z)} F_j + \sum_{j\in I(z)} \frac{F_j\mu(\hat{z})\mu(\hat{z})^TF_j}{\mu(\hat{z})^TF_j\mu(\hat{z})}\Big) \tag{1.11.84}$$

Note that in addition to proving (1.8.24) this also proves $(viii)$ as this form makes clear that, owing to continuity of eigenvalues, infimizing $\hat{A}_1(z)$ over z will give zero (and hence so too will

infimizing $\hat{A}_2(z)$ over z since $\hat{A}_2(z) \leqslant \hat{A}_1(z)$). Specifically the number of possibly non-zero eigenvalues of $\hat{R}_z + \hat{T}_z$ is $2nk - k^2$ and is thus monotone increasing in rank, and thus a sequence $(z_i)_{i \geqslant 1} \subset \mathbb{C}_*^{n \times r}$ approaching a surface of lower rank k will have $\lambda_{2nr-r^2}(\hat{R}_z + \hat{T}_z)$ approach zero. Somewhat more remarkably, (1.11.84) actually gives us $\hat{A}_2(z)$ as an eigenvalue problem also. Specifically, we prove that the "differentiable" terms in $\hat{A}_2(z)$ are equal to those in $\hat{A}_1(z)$ and that in fact these are the only terms which contribute to $\hat{A}_2(z)$. We define

$$\begin{aligned}
\hat{A}_2^I(z) &= \lim_{R \to 0} \inf_{\substack{x,y \in \mathbb{C}^{n \times r} \\ D(x,z) < R \\ D(y,z) < R \\ \text{rank}(x) \leqslant k \\ \text{rank}(y) \leqslant k}} \frac{\sum_{k \in I(z)} |\alpha_k(x) - \alpha_k(y)|^2}{D(x,y)^2} \\
\hat{A}_2^{I_0}(z) &= \lim_{R \to 0} \inf_{\substack{x,y \in \mathbb{C}^{n \times r} \\ D(x,z) < R \\ D(y,z) < R \\ \text{rank}(x) \leqslant k \\ \text{rank}(y) \leqslant k}} \frac{\sum_{k \in I_0(z)} |\alpha_k(x) - \alpha_k(y)|^2}{D(x,y)^2} \\
\hat{A}_1^I(z) &= \lim_{R \to 0} \inf_{\substack{x \in \mathbb{C}^{n \times r} \\ D(z,x) < R \\ \text{rank}(x) \leqslant k}} \frac{\sum_{k \in I(z)} |\alpha_k(x) - \alpha_k(z)|^2}{D(x,z)^2} \\
\hat{A}_1^{I_0}(z) &= \lim_{R \to 0} \inf_{\substack{x \in \mathbb{C}^{n \times r} \\ D(z,x) < R \\ \text{rank}(x) \leqslant k}} \frac{\sum_{k \in I_0(z)} |\alpha_k(x) - \alpha_k(z)|^2}{D(x,z)^2}
\end{aligned} \tag{1.11.85}$$

So that $\hat{A}_2(z) \geqslant \hat{A}_2^{I_0}(z) + \hat{A}_2^I(z) \geqslant \hat{A}_2^I(z)$, $\hat{A}_2^I(z) \leqslant \hat{A}_1^I(z)$, and $\hat{A}_2^{I_0}(z) \leqslant \hat{A}_1^{I_0}(z)$. Applying the mean value theorem to the functions $g_k : [0,1] \to \mathbb{R}, g_k(c) = \alpha_k((1-c)x + cy)$ for $k \in I(z)$ we see that there exist $c_k \in [0,1]$ so that $\alpha_k(y) - \alpha_k(x) = g(1) - g(0) = g'(c_k) = D\alpha_k((1 - c_k)x + c_k y)(y - x)$ (recall that these are precisely the k for which said differential exists, and the differential is taken with respect to the real vector space structure). Hence, replacing the rank constraints with the assumption that $z \in \mathbb{C}_*^{n \times k}$ and aligning both x and y with z so that $z^* x \geqslant 0$

and $z^*y \geqslant 0$ we have:

$$\hat{A}_2^I(z) = \lim_{R\to 0} \inf_{\substack{x,y\in\mathbb{C}^{n\times k} \\ ||x-z||<R \\ ||y-z||<R \\ z^*x\geqslant 0 \\ z^*y\geqslant 0}} \frac{\sum_{k\in I(z)} |D\alpha_k((1-c_k)x + c_k y)(y-x)|^2}{D(x,y)^2} \tag{1.11.86}$$

Using the fact that $D(x,y) \leqslant ||y-x||_2$ and writing $x = z + \xi$ and $y = z + \eta$ we obtain that

$$\hat{A}_2^I(z) \geqslant \lim_{R\to 0} \inf_{\substack{\eta,\xi\in\Delta_z \\ ||\xi||<R \\ ||\eta||<R}} \frac{\sum_{k\in I(z)} |D\alpha_k(z + (1-c_k)\xi + c_k\eta)(\eta-\xi)|^2}{||\eta-\xi||_2^2} \tag{1.11.87}$$

The trick of linearizing the conic constraints here to $\xi, \eta \in \Delta_z$ is crucial since it allows us to strictly weaken the constraints in the infimum by taking $w = \eta - \xi$ so that, after using the continuity of $D\alpha_k$ (α_k is continuously differentiable when differentiable)

$$\begin{aligned}
\hat{A}_2^I(z) &\geqslant \lim_{R\to 0} \inf_{\substack{\eta,\xi\in\Delta_z \\ ||\xi||_2<R \\ ||\eta||_2<R}} \frac{\sum_{k\in I(z)} |D\alpha_k(z + (1-c_k)\xi + c_k\eta)(\eta-\xi)|^2}{||\eta-\xi||_2^2} \\
&= \lim_{R\to 0} \inf_{\substack{\eta,\xi\in\Delta_z \\ ||\xi||_2<R \\ ||\eta||_2<R}} \frac{\sum_{k\in I(z)} |D\alpha_k(z)(\eta-\xi)|^2}{||\eta-\xi||_2^2} + O(||\xi||_2^2 + ||\eta||_2^2) \\
&\geqslant \lim_{R\to 0} \inf_{\substack{w\in\Delta_z \\ ||w||_2<2R}} \frac{\sum_{k\in I(z)} |D\alpha_k(z)(w)|^2}{||w||_2^2} \\
&= \min_{\substack{w\in H_{\pi,z}(\mathbb{C}_*^{n\times k}) \\ ||w||_2=1}} \sum_{k\in I(z)} |D\alpha_k(z)(w)|^2 \\
&= \lambda_{2nk-k^2}\Big(\sum_{j\in I(z)} \frac{F_j\mu(\hat{z})\mu(\hat{z})^T F_j}{\mu(\hat{z})^T F_j \mu(\hat{z})}\Big) = \hat{A}_1^I(z)
\end{aligned} \tag{1.11.88}$$

We already had the reverse inequality $\hat{A}_2^I(z) \leqslant \hat{A}_1^I(z)$, hence $\hat{A}_2^I(z) = \hat{A}_1^I(z)$. Moreover, assuming this minimum is achieved by $w_0 \in H_{\pi,z}(\mathbb{C}_*^{n\times k})$ then if we put $x = z + \frac{1}{2}w_0$ $y = z - \frac{1}{2}w_0$ we see that the $\hat{A}_2^{I_0}(z)$ term vanishes and $\hat{A}_2^I(z)$ is achieved, hence $\hat{A}_2(z) \leqslant \hat{A}_2^I(z)$. We already had the reverse inequality, so we conclude that $\hat{A}_2(z) = \hat{A}_2^I(z) = \hat{A}_1^I(z)$ and $\hat{A}_2^{I_0}(z) = 0$. In summary

$$\begin{aligned}\hat{A}_2(z) &= \min_{\substack{W\in\mathbb{R}^{2nk}\\ W\perp\mathcal{V}_z\\ ||W||_2=1}} W^T\Big\{\sum_{j\in I(z)}\frac{F_jZZ^TF_j}{Z^TF_jZ}\Big\}W \\ &= \lambda_{2nk-k^2}\Big(\sum_{j\in I(z)}\frac{F_jZZ^TF_j}{Z^TF_jZ}\Big)\end{aligned} \tag{1.11.89}$$

Thus claims (i) and (ii) are proven. Claim (iii) follows immediately from the inequality (1.6.6). This concludes the proof of the Theorem 1.8.8. □

Remark 1.11.5. If z were not assumed full rank in (1.11.81) then $w \in \Delta_z$ would possibly have a non-zero component w_Γ in $\Gamma_z \subset V_{\pi,z}(\mathbb{C}_*^{n\times k})$. As a result, it would be possible to obtain a sequence (with the horizontal space component of w converging to zero) for which the second sum in the last line of (1.11.81) is eventually fourth order in $||w||_2$, thus $A_1(z)$ would be zero wherever α is differentiable (almost everywhere in measure). The rank constraint in the definition of $\hat{A}_1(z)$ that rank$(x) \leqslant k$ avoids this, since it allows us to assume that z is full rank and hence that Γ_z is trivial.

1.11.4 Proof of Theorem 1.8.12

Proof. The proof of (i) is essentially identical to the proof of the analogous eigenvalue formula for the lower bound a_0 in Theorem 1.8.5. One first changes coordinates to $z = \frac{1}{2}(x + y)$ and

$w = x - y$ and repeats the computation (1.11.6) to obtain

$$b_0 = \sup_{z \in \mathbb{C}^{n\times r}} \max_{\substack{W \in T_{\pi(\hat{z})}(\mathring{S}^{k,0}(\mathbb{C}^n)) \\ ||W||_2 = 1}} \sum_{j=1}^{M} |\langle W, A_j \rangle_{\mathbb{R}}|^2 \tag{1.11.90}$$

At this point we note that

$$b_0 \leqslant \sup_{W \in \mathbf{Sym}(\mathbb{C}^n)} \frac{||\mathcal{A}(W)||_2^2}{||W||_2^2} = ||\mathcal{A}||_{2\to 2}^2 \tag{1.11.91}$$

As before we observe that it suffices to take $z \in \mathbb{C}_*^{n\times r}$ since if $\hat{z} \in \mathbb{C}_*^{n\times k}$ and $\tilde{z} \in \mathbb{C}_*^{n\times(r-k)}$ and $z = [\hat{z}|\tilde{z}]$ with $\tilde{z}^*\hat{z} = 0$ then $T_{\pi(z)}(\mathring{S}^{r,0}(\mathbb{C}^n)) \supset T_{\pi(\hat{z})}(\mathring{S}^{k,0})$. One then employs the tangent space parametrization (1.11.27) and repeats the computation (1.11.28) to obtain

$$b_0 = \sup_{z \in \mathbb{C}_*^{n\times r}} \lambda_1(Q_z) = \max_{\substack{U \in U(n) \\ U = [U_1|U_2] \\ U_1 \in \mathbb{C}^{n\times r}, U_2 \in \mathbb{C}^{n\times n-r}}} \lambda_1(Q_{[U_1|U_2]}) \tag{1.11.92}$$

This concludes the proof of (i). To prove (ii) we will employ the following lemma.

Lemma 1.11.6. *Let $|||\cdot|||$ be any norm. Then*

$$||\mathcal{A}||_{1\to|||\cdot|||} = \sup_{\substack{x \in \mathbb{C}^n \\ ||x||_2 = 1}} |||\mathcal{A}(xx^*)||| \tag{1.11.93}$$

In other words the operator norm $||\mathcal{A}||_*$ of $\mathcal{A} : (Sym(\mathbb{C}^n)(\mathbb{C}^n), ||\cdot||_1) \to (\mathbb{R}^m, |||\cdot|||)$ is achieved on a matrix of rank 1.

Proof. Let $R \in \mathrm{Sym}(\mathbb{C}^n)$ be non-zero such that $||R||_1 = 1$ and $|||\mathcal{A}(R)||| = ||\mathcal{A}||_*||R||_1$. Write

$R = \sum_{j=1}^{n} r_j e_j e_j^*$ and note that $||R||_1 = 1$ implies $\sum_{j=1}^{n} |r_j| = 1$. Then

$$||\mathcal{A}||_* = ||\mathcal{A}||_* ||R||_1 = |||\sum_{j=1}^{n} r_j \mathcal{A}(e_j e_j^*)||| \leqslant (\sum_{j=1}^{n} |r_j|) \max_{j=1,\ldots,n} |||\mathcal{A}(e_j e_j^*)||| = \max_{j=1,\ldots,n} |||\mathcal{A}(e_j e_j^*)|||$$

(1.11.94)

Let $x_0 = e_{j_0}$ where j_0 is the index that achieves the maximum. Then $||x_0||_2 = 1$ and $||A||_* \leqslant |||\mathcal{A}(x_0 x_0^*)|||$, but of course this bound is achievable by just plugging in $x_0 x_0^*$ into $\mathcal{A}$. Thus the operator norm of $\mathcal{A}$ is achieved on a matrix of rank 1 and the lemma holds. □

Next note that

$$\begin{aligned} b_{0,1} &= \sup_{\substack{x,y \in \mathbb{C}^{n\times r} \\ [x]\neq[y]}} \frac{\sum_{j=1}^{m} |\langle xx^* - yy^*, A_j \rangle_{\mathbb{R}}|^2}{||xx^* - yy^*||_1^2} \\ &= \sup_{z \in \mathbb{C}_*^{n\times r}} \sup_{W \in T_{\pi(z)}(\mathring{S}^{r,0}(\mathbb{C}^n))} \frac{||\mathcal{A}(W)||_2^2}{||W||_1^2} \\ &\leqslant \sup_{\substack{W \in \mathbf{Sym}(\mathbb{C}^n) \\ ||W||_1 = 1}} ||\mathcal{A}(W)||_2^2 \\ &= ||\mathcal{A}||_{1\to 2}^2 \end{aligned} \tag{1.11.95}$$

Note that by an identical computation $b_0 \leqslant ||\mathcal{A}||_{2\to 2}$. By the Lemma $||\mathcal{A}||_{1\to 2} = \sup_{x\in\mathbb{C}^n, ||x||_2=1} ||\mathcal{A}(xx^*)||_2^2$,

hence

$$\begin{aligned} b_{0,1} &\leqslant \sup_{x\in\mathbb{C}^n} \frac{||\mathcal{A}(xx^*)||_2^2}{||xx^*||_1^2} \\ &\leqslant \sup_{x\in\mathbb{C}^{n\times r}} \frac{||\mathcal{A}(xx^*)||_2^2}{||xx^*||_1^2} \\ &= \frac{||\mathcal{A}(x_0x_0^*)||_2^2}{||x_0x_0^*||_1^2} \\ &\leqslant \sup_{\substack{U_2\in\mathbb{C}^{n\times n-k} \\ U_2^*U_2=\mathbb{I}_{n-k\times n-k} \\ k=1,\ldots,r}} \sup_{\substack{W\in\mathbf{Sym}(\mathbb{C}^n) \\ U_2^*WU_2=0}} \frac{||\mathcal{A}(W)||_2^2}{||W||_1^2} \\ &= b_0 \end{aligned} \tag{1.11.96}$$

Where in the second to last equality we note that it suffices to take U_2 such that $U_2U_2^* = \mathbb{P}_{\mathbf{Ran}(x_0)^\perp}$ and in the last equality we use the implicit parametrization of the tangent space (1.7.7). Thus

$$b_{0,1} = ||\mathcal{A}||_{1\to 2} = \sup_{x\in\mathbb{C}^n} \frac{||\mathcal{A}(xx^*)||_2^2}{||xx^*||_1^2} = \sup_{x\in\mathbb{C}^{n\times r}} \frac{||\mathcal{A}(xx^*)||_2^2}{||xx^*||_1^2} \tag{1.11.97}$$

We now seek an operator $T_r : \mathbb{C}^{n\times r} \to (\mathbb{C}^{n\times r})^m$, an integer q, and a norm $|||\cdot|||$ so that for $x \in \mathbb{C}^{n\times r}$

$$|||T_r(x)|||^q = ||\mathcal{A}(xx^*)||_2^2 \tag{1.11.98}$$

We find that if $A_j \geqslant 0$ for all j then

$$||\mathcal{A}(xx^*)||_2^2 = \sum_{j=1}^m |\langle xx^*, A_j\rangle_{\mathbb{R}}|^2 = \sum_{j=1}^m ||A_j^{\frac{1}{2}}x||_2^4 \tag{1.11.99}$$

So we let T_r be as in Definition 1.8.11, $|||X||| = |||X|||_{2,4}$ and $q = 4$ and find $b_0 = ||T_r||_{2\to(2,4)}^4 =$

$||T_1||^4_{2\to(2,4)}$. This concludes the proof of (ii). To prove (iii) note that by (1.6.5) $||(xx^*)^{\frac{1}{2}} - (yy^*)^{\frac{1}{2}}||_2 \geqslant D(x,y)$ hence

$$B_0 \leqslant \sup_{\substack{x,y\in\mathbb{C}^{n\times r}\\ [x]\neq[y]}} \frac{||\alpha(x)-\alpha(y)||_2^2}{D(x,y)^2} \tag{1.11.100}$$

Thus

$$\begin{aligned} B_0 &\leqslant \sup_{\substack{x,y\in\mathbb{C}^{n\times r}\\ [x]\neq[y]}} \frac{1}{D(x,y)^2}\sum_{j=1}^m |\langle xx^*, A_j\rangle^{\frac{1}{2}} - \langle yy^*, A_j\rangle^{\frac{1}{2}}|^2 \\ &= \sup_{\substack{x,y\in\mathbb{C}^{n\times r}\\ x^*y\geqslant 0}} \frac{1}{||x-y||_2^2}\sum_{j=1}^m \frac{|\langle xx^*-yy^*, A_j\rangle_{\mathbb{R}}|^2}{(\langle xx^*, A_j\rangle^{\frac{1}{2}} + \langle yy^*, A_j\rangle^{\frac{1}{2}})^2} \end{aligned} \tag{1.11.101}$$

We now make the change of coordinates $z = \frac{1}{2}(x+y)$, $w = x-y$ so that $x = z+\frac{1}{2}w$, $y = z-\frac{1}{2}w$. As before let $I_0(z)$ be the subset of $\{1,\dots,m\}$ for which $A_jz = 0$ and $I(z)$ its complement in $\{1,\dots,m\}$. In this case we note that if $j \in I_0(z)$ then $0\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}} = \langle xx^* - yy^*, A_j\rangle$. Thus, employing the triangle inequality via $\langle xx^*, A_j\rangle^{\frac{1}{2}} + \langle yy^*, A_j\rangle^{\frac{1}{2}} = ||A_j^{\frac{1}{2}}x||_2 + ||A_j^{\frac{1}{2}}y||_2 \geqslant 2||A_j^{\frac{1}{2}}z||_2 = 2\langle zz^*, A_j\rangle^{\frac{1}{2}}$ we find that

$$B_0 \leqslant \sup_{\substack{x,y\in\mathbb{C}^{n\times r}\\ x^*y\geqslant 0}} \frac{1}{||x-y||_2^2}\sum_{j\in I(z)}^m \frac{|\langle xx^*-yy^*, A_j\rangle_{\mathbb{R}}|^2}{(\langle xx^*, A_j\rangle^{\frac{1}{2}} + \langle yy^*, A_j\rangle^{\frac{1}{2}})^2} \tag{1.11.102}$$

$$\leqslant \sup_{\substack{z\in\mathbb{C}^{n\times r}\\ z\neq 0}} \sup_{\substack{w\in\mathbb{C}^{n\times r}\\ z^*z-\frac{1}{4}w^*w+\frac{1}{2}(w^*z-z^*w)\geqslant 0}} \frac{1}{||w||_2^2}\sum_{j\in I(z)} \frac{|\langle zw^*+wz^*, A_j\rangle_{\mathbb{R}}|^2}{4\langle zz^*, A_j\rangle} \tag{1.11.103}$$

Next note that the condition $z^*z - \frac{1}{4}w^*w + \frac{1}{2}(w^*z - z^*w) \geqslant 0$ holds if and only if $z^*w = w^*z$ and $w^*w \leqslant 4z^*z$. Moreover, since w only appears as $w/||w||_2$ we may scale w so that $\sigma_1(w) \leqslant$

$\sigma_k(z)$ (where z has rank k), thus the latter non-linear criterion becomes the linear criterion that $w\mathbb{P}_{\ker(z)} = 0$. Taken together, these these criterion hold if and only if $w \in H_z$. Thus, with reference to the computations (1.11.41) and (1.11.42) we find that

$$B_0 \leqslant \sup_{\substack{z\in\mathbb{C}^{n\times r}\\ z\neq 0}} \sup_{w\in H_z} \frac{1}{||w||_2^2} \sum_{j\in I(z)} \frac{|\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}}|^2}{4\langle zz^*, A_j\rangle} \tag{1.11.104}$$

$$= \sup_{\substack{z\in\mathbb{C}^{n\times r}\\ z\neq 0}} \max_{\substack{W\in\mathbb{R}^{2nk}\\ W\perp\mathcal{V}_Z\\ ||W||_2=1}} W^T \left(\sum_{j\in I(z)} \frac{F_j\mu(\hat{z})\mu(\hat{z})^T F_j}{\mu(\hat{z})^T F_j \mu(\hat{z})} \right) W \tag{1.11.105}$$

$$= \sup_{\substack{z\in\mathbb{C}^{n\times r}\\ z\neq 0}} \lambda_1(\hat{T}_z) \tag{1.11.106}$$

Moreover note that by setting $y = 0$ in the definition of B_0 and observing that $||(xx^*)^{\frac{1}{2}}||_2 = ||x||_2$ and that $\langle xx^*, A_j\rangle \geqslant 0$ we obtain that

$$B_0 \geqslant \sup_{x\in\mathbb{C}^{n\times r}} \frac{1}{||x||_2^2} \sum_{j=1}^{m} \langle xx^*, A_j\rangle = B \tag{1.11.107}$$

Meanwhile by Cauchy-Schwartz $\langle zw^*, A_j\rangle \leqslant ||A_j^{\frac{1}{2}} w||_2 ||A_j^{\frac{1}{2}} z||_2 = \langle ww^*, A_j\rangle^{\frac{1}{2}} \langle zz^*, A_j\rangle^{\frac{1}{2}}$ (similarly for $\langle wz^*, A_j\rangle$). Hence

$$\begin{aligned} B_0 &\leqslant \sup_{\substack{z\in\mathbb{C}^{n\times r}\\ z\neq 0}} \lambda_1(\hat{T}_z) \\ &= \sup_{\substack{z\in\mathbb{C}^{n\times r}\\ z\neq 0}} \sup_{w\in H_z} \frac{1}{||w||_2^2} \sum_{j\in I(z)} \frac{|\langle zw^* + wz^*, A_j\rangle_{\mathbb{R}}|^2}{4\langle zz^*, A_j\rangle} \\ &\leqslant \sup_{w\in H_z} \frac{1}{||w||_2^2} \sum_{j\in I(z)} \langle ww^*, A_j\rangle \\ &\leqslant \sup_{w\in\mathbb{C}^{n\times r}} \frac{1}{||w||_2^2} \sum_{j=1}^{m} \langle ww^*, A_j\rangle_{\mathbb{R}} = B \end{aligned} \tag{1.11.108}$$

Thus $B \leqslant B_0 \leqslant \sup_{\substack{z\in\mathbb{C}^{n\times r}\\ z\neq 0}} \lambda_1(\hat{T}_z) \leqslant B$ and hence all three are equal. This concludes the proof of (iii) and of Theorem 1.8.12. □

1.11.5 Proof of Theorem 1.8.13

Proof. It is shown in Proposition 5 that the map β is injective if and only if it is lower Lipschitz, that is if and only if $a_0 > 0$. This gives equivalence of (i) to (ii) immediately since we proved in Theorem 1.8.5 that

$$a_0 = \min_{\substack{U_1\in\mathbb{C}^{n\times r}\\ U_2\in\mathbb{C}^{n\times(n-r)}\\ [U_1|U_2]\in U(n)}} \lambda_{2nr-r^2}(Q_{[U_1|U_2]}) \tag{1.11.109}$$

Similarly, it is evident from (1.11.70) that $a_0 > 0$ if and only if $\hat{a}(z) > 0$ whenever $z^*z = \mathbb{I}_{r\times r}$. It is proved in Theorem 1.8.5 that $\hat{a}(z) = \lambda_{2nr-r^2}(\hat{Q}_z)$, and also that the null space of $\hat{Q}_z$ includes the r^2 dimension $\mathcal{V}_z$. Thus the frame is generalized phase retrievable if and only if the null space $\hat{Q}_z$ does not extend beyond $\mathcal{V}_z$ for any z of orthonormal columns, proving equivalence of (i) to (iii). We prove equivalence of (ii) to (iv) by noting that $Q_{[U_1|U_2]}$ is invertible if and only if

$$\text{span}_{\mathbb{R}}\{\begin{bmatrix}\tau(U_1^*A_jU_1)\\ \mu(U_2^*A_jU_1)\end{bmatrix}\}_{j=1}^m = \mathbb{R}^{2nr-r^2} \tag{1.11.110}$$

Noting that $\tau^{-1}(\mathbb{R}^{r^2}) = \text{Sym}(\mathbb{C}^r)$ and $\mu^{-1}(\mathbb{R}^{2nr-2r^2}) = \mathbb{C}^{n-r\times r}$, thus $Q_{[U_1|U_2]}$ is invertible if and only if there exist $c_1,\dots,c_m \in \mathbb{R}$ so that (1.8.39a) and (1.8.39b) are satisfied. To prove equivalence with (v) note that (1.8.39a) and (1.8.39b) both hold if and only if for all $U = [U_1|U_2]$

we have

$$\begin{aligned}\text{span}_{\mathbb{R}}\{A_jU_1\} &= \{U\begin{bmatrix} H \\ B \end{bmatrix} | H \in \text{Sym}(\mathbb{R}^n), B \in \mathbb{C}^{(n-r)\times r}\} \\ &= \{U_1K | K \in \mathbb{C}^{r\times r}, K^* = -K\}^{\perp}\end{aligned} \tag{1.11.111}$$

Finally note that while (v) trivially implies (vi) it is also the case that $\langle A_jU_1, U_1K\rangle_{\mathbb{R}} = \langle U_1^*A_jU_1, K\rangle_{\mathbb{R}} = 0$ for every U_1 and every K since $U_1^*A_jU_1$ is Hermitian and K is skew-Hermitian, hence it is automatically true that $\text{span}_{\mathbb{R}}\{A_jU_1\} \subset \{U_1K | K \in \mathbb{C}^{r\times r}, K^* = -K\}^{\perp}$. Thus we also obtain (vi) implies (v).

This concludes the proof of Theorem 1.8.13. □

1.12 Numerical experiments

The main benefit of lower Lipschitz results like Theorem 1.8.1 is that they provide quantitative control over reconstruction error in the generalized phase retrieval problem, as opposed to the topological result in Proposition 5 that the error is bounded whenever the matrix frame is generalized phase retrievable (i.e. that $a_0 > 0$). This is only true, however, if for a given frame one can make headway in computing the lower Lipschitz constant a_0. Unfortunately (1.8.18) yields a_0 as a non-convex optimization problem, so for the time being we content ourselves with examining the statistics of the local lower Lipschitz constants $\hat{a}_2(z)$ and $a(z)$. We also verify numerically the result in Theorem 1.8.8 that α is not globally lower Lipschitz (i.e. that $A_0 = 0$) by examining the statistics of the local lower Lipschitz constant $\hat{A}_2(z)$.

For each experiment we use a fixed frame set of cardinality $m = 4nk - 4k^2$, noting that

Theorem 2.1 in [9] implies that a generic frame for $\mathbb{C}^{n\times k}$ with cardinality $m \geqslant 4nk - 4k^2$ will be generalized phase retrievable when $2k \leqslant n$. The experiment shown in Figure 1.2 supports the result in Theorem 1.8.8 that $\inf_{z\in\mathbb{C}^{n\times r}\setminus\{0\}} \hat{A}_2(z) = 0$ for $r > 1$, thus that the α analysis map is not globally lower Lipschitz with respect to either $D(x, y)$ or $||(xx^*)^{\frac{1}{2}} - (yy^*)^{\frac{1}{2}}||_2$ when $r > 1$. This experiment also supports the earlier result in [23] that when $r = 1$ $\inf_{z\in\mathbb{C}^{n\times r}\setminus\{0\}} \hat{A}_2(z) > 0$. The experiment shown in Figure 1.3 supports the result noted in the proof of Theorem 1.8.5 that $\inf_{z\in\mathbb{C}^{n\times r}\setminus\{0\}} \hat{a}_2(z) = 0$ for $r > 1$, thus that the β analysis map is not globally lower Lipschitz with respect to $d(x, y)$ when $r > 1$. That this quantity is non-zero when $r = 1$ follows from the fact that for $r = 1$ we have $d(x, y) = ||xx^* - yy^*||_1$ (see Theorem 1.6.4). Finally, the experiment shown in Figure 1.4 supports the result in Theorem 1.8.5 that $a_0 = \inf_{z\in\mathbb{C}^{n\times r}\setminus\{0\}} a(z) > 0$ even when $r > 1$, thus that the β analysis map is globally lower Lipschitz with respect to $||xx^* - yy^*||_2$ whenever the frame $(A_j)_{j\geqslant 1}$ is generalized phase retrievable. Code for all numerical experiments can be found at github.com/cbartondock/LipschtizAnalysisofGenPR.

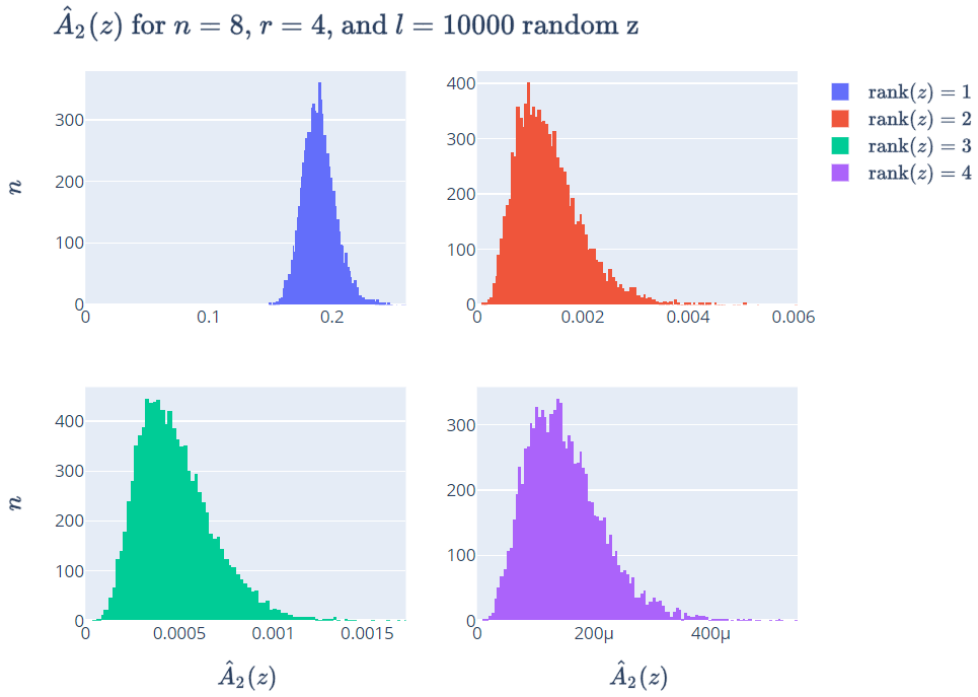


Figure 1.2: In all experiments $\hat{A}_2(z)$ is computed for a fixed frame of $4nk - 4k^2$ matrices in $\mathbb{C}^{n\times k}$ for $l = 10^4$ samples of z having rank k. The entries of both z and the frame matrices are sampled from a complex Gaussian with unit variance and zero mean. As can clearly be seen only the $k = 1$ case has a clear separation from zero.

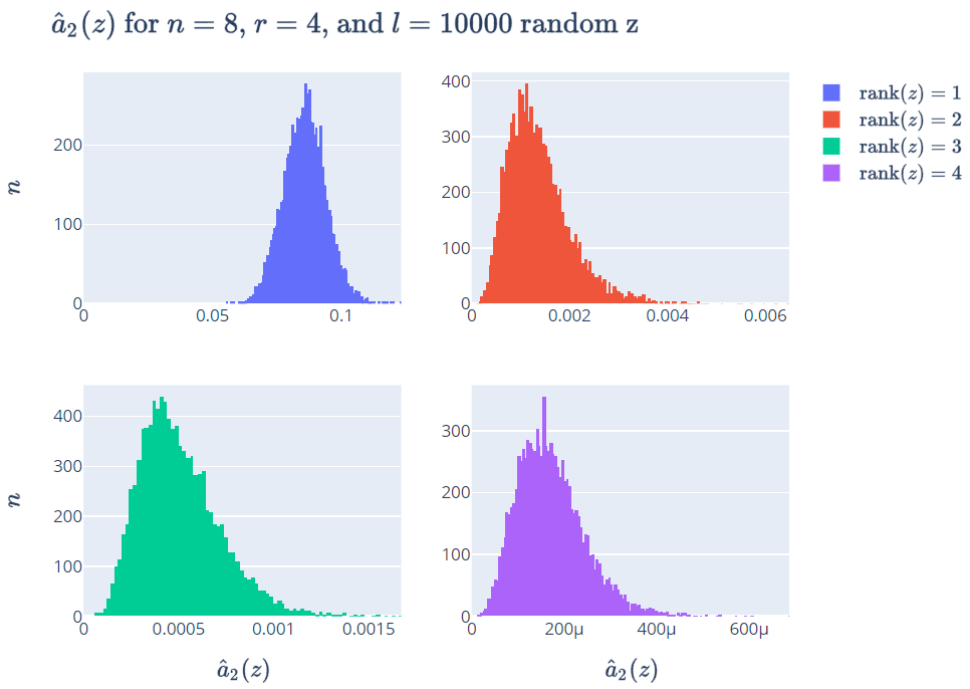


Figure 1.3: In all experiments $\hat{a}_2(z)$ is computed for a fixed frame of $4nk - 4k^2$ matrices in $\mathbb{C}^{n\times k}$ for $l = 10^4$ samples of z having rank k. The entries of both z and the frame matrices are sampled from a complex Gaussian with unit variance and zero mean. As can clearly be seen only the $k = 1$ case has a clear separation from zero.

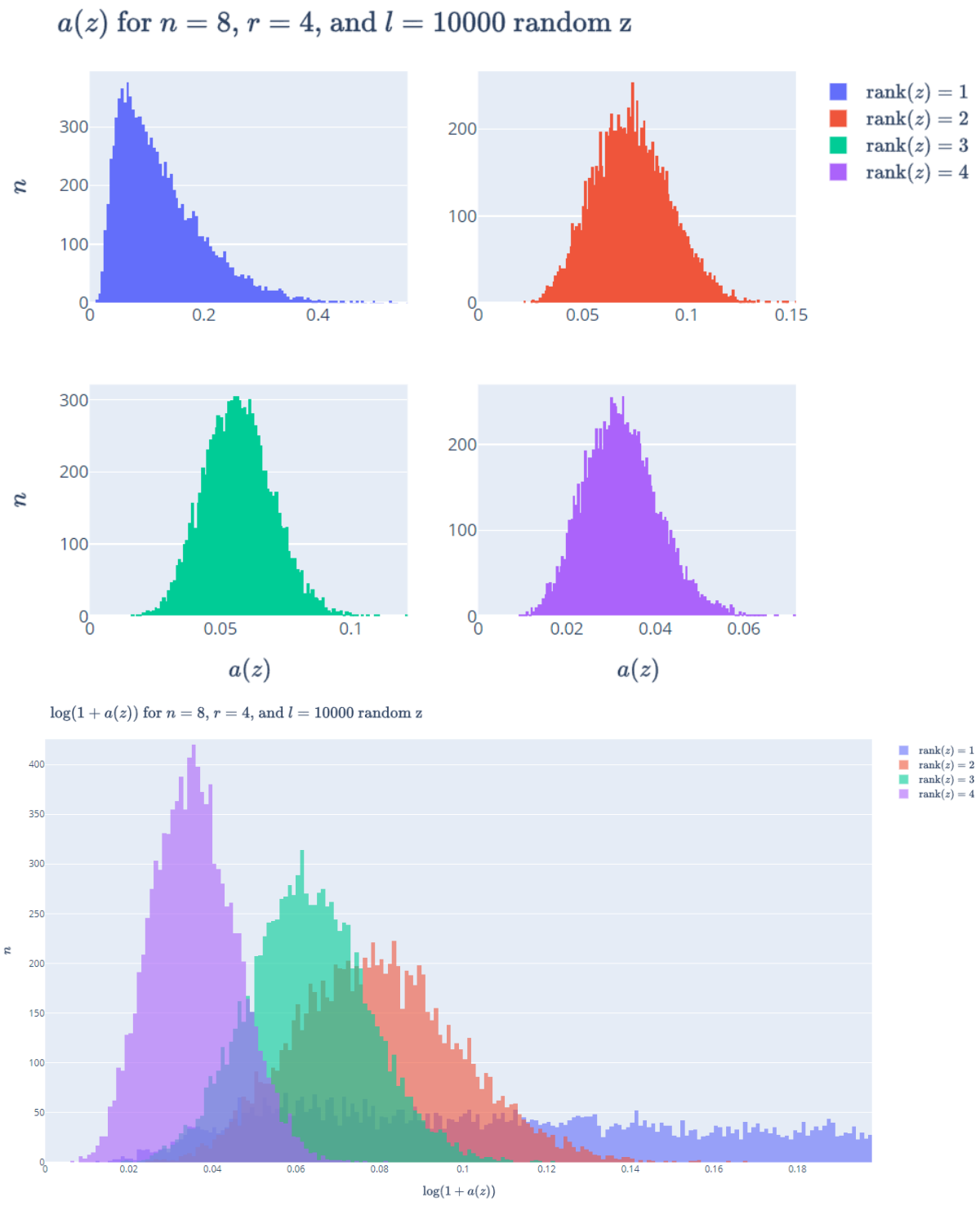


Figure 1.4: ' In all experiments $a(z) = \lambda_{2nk-k^2}(Q_{[U_1|U_2]})$ is computed for a fixed frame of $4nk - 4k^2$ matrices in $\mathbb{C}^{n\times k}$ for $l = 10^4$ samples of $U \in U(n)$ distributed according to the uniform Haar distribution on $U(n)$. $U_1 \in \mathbb{C}^{n\times k}$ is composed of the first k columns of U so that $Q_{[U_1|U_2]} \in \mathbb{C}^{2nk-k^2\times 2nk-k^2}$. The entries of the frame matrices are sampled from a complex Gaussian with unit variance and zero mean. In this case an overlapping log-plot is also included, in which clear separation from zero can be seen for $k = 1, \ldots, 4$.

1.13 Conclusion

This paper extends known results about the stability of generalized phase retrieval to the "impure state" case where the phase no longer comes from $U(1)$ but instead the non-abelian groups $U(r)$ where $r > 1$. We showed that the situation changes drastically in this case, both because $U(r)$ is non-abelian and because for $r > 1$ a sequence in $\mathbb{C}_*^{n\times r}/U(r)$ with $||x_n||_2 = 1$ can come arbitrarily close to dropping in rank. In particular, we showed that while the β analysis map remains lower Lipschitz with respect to the norm induced distance on $\text{Sym}(\mathbb{C}^n)$ (Theorem 1.8.5), the α analysis map does not (Theorem 1.8.8). Our analysis relies on several Lipschitz embeddings of $\mathbb{C}^{n\times r}/U(r)$ into the Euclidean space $\text{Sym}(\mathbb{C}^n)$ (Theorem 1.6.4) and a Whitney stratification of the positive semidefinite matrices into positive semidefinite matrices of fixed rank (Theorem 1.7.4). This investigation of the geometry of positive semidefinite matrices incidentally provided the interesting and (to the best of our knowledge) previously unknown result that the Riemannian geometry of the stratifying manifolds given by the Bures-Wasserstein metric is compatible with the stratification. In particular geodesics of positive semi-definite matrices with respect to the Bures-Wasserstein metric are rank preserving and may be approximated by geodesics of higher rank. We note that the fact that $a_0 > 0$ and can be explicitly computed as in (1.8.18) suggests that known convergent algorithms for generalized phase retrieval may be extended to the case $r > 1$. Finally, the explicit computation of the lower Lipschitz bound for the β map allowed for a novel characterization of generalized phase retrievable frames in the impure state case $r > 1$ (Theorem 1.8.13).

Chapter 2: Chart Based Normalizing Flows[1]

2.1 Introduction

Generative modeling is a machine learning paradigm that aims to learn data distributions and sample from it. If the data is drawn from a random variable $x \sim p(x)$, then one way to do this is to directly model $p(x)$ via a parameterized model so that $p_\theta(x) \approx p(x)$. Such a model can then be used to generate new samples, which are expected to be statistically indistinguishable from the observed samples. Moreover, generative models that learn $p(x)$ are useful for data augmentation, outlier detection, domain transfer [41, 42], and as priors for other downstream tasks [43–45].

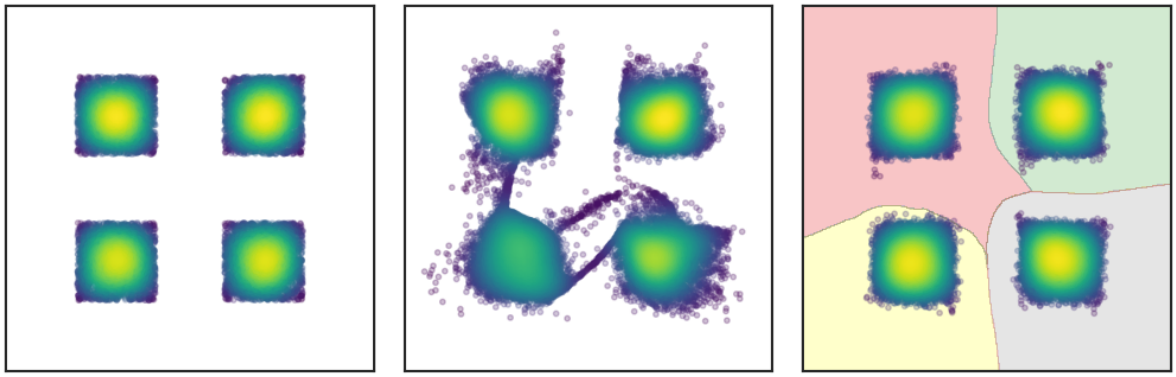
Among the most successful generative models are deep latent variable models, which assume that the latent factors of variation underlying the generative process of the data follow a simple distribution, such as a Gaussian or a uniform distribution. The non-linear function transforming this latent space to the data space (or vice-versa) is parameterized as a neural network and learned using gradient descent. Depending upon their formulation, there are three broad categories of deep latent variable models - GANs [46], VAEs [47], and normalizing flows. In this work, we focus on normalizing flows, a class of deep latent variable models introduced in [48] that support efficient sampling, exact density estimation, and inference [49]. A normalizing flow

[1] In collaboration with Radu V. Balan, Sahil Sidheekh, Tushar Jain, and Maneesh Singh. This work was submitted to the Uncertainty in Artificial Intelligence (UAI) conference. My contribution to this section was the theoretical component of the work.

maps the data space to a latent space through a series of diffeomorphisms (differentiable, bijective transformations with differentiable inverses). The data is assumed to follow an analytically computable distribution in the latent space, typically a Gaussian. Since the mapping is a diffeomorphism, the density in the data space can be obtained using the change of variables formula. To generate new samples using a flow, one can sample from the latent distribution and use the inverse transformation to map them to the data space. This makes normalizing flows powerful generative models that support exact density evaluation in contrast to GANs and VAEs.

Despite the advantages of normalizing flows over other generative models, their diffeomorphic requirement poses several restrictions. Firstly, a continuous bijective transformation with continuous inverse preserves the topology of its domain. Therefore, the data space is required to be topologically equivalent to the support of the latent distribution, typically to D dimensional Euclidean space since the latent distribution is assumed to be a Gaussian. However, real data distributions typically differ from Euclidean space in many topological respects, such as the number of connected components, the presence of holes, etc. A normalizing flow would thus fail to model such data distribution accurately.

A particularly troubling consequence of the continuous invertibility of flow transformations is that they are dimensionality preserving. However, according to the *manifold hypothesis*, high dimensional real-world data living in $\mathcal{X} \simeq \mathbb{R}^D$ is often supported on a $d << D$ manifold of the embedding space. To efficiently learn such distributions using flows, one needs to design expressive transformations that can map from a d dimensional latent space to a the D dimensional data space without making learning intractable. Recent work using stochastically invertible tall matrices [50] and dimension raising conformal embeddings [1] have paved the way in designing such transformations, however in both works expressivity is limited by the fact that the dimension



(a) Real Data (b) Classic Flow (c) VQ-Flow

Figure 2.1: Augmentation of our framework (c) enables a classic flow (b) to better model the discontinuities in the data manifold through a learned atlas of charts(shaded region).

changing operations are restricted to be linear (in [50]) or made up of Möbius transformations (in [1]).

In this work, we propose to address the above limitations by parameterizing a family of normalizing flows to compose an atlas of charts over the data manifold. As the topology of the data manifold is expected to be "locally" equivalent to Euclidean space, a local normalizing flow should be able to model the local distribution over a chart region effectively. Further, by learning a mixture of flows over well-chosen charts, our approach compensates naturally for the limited expressiveness of existing flows. We summarize the main contributions of this work below:

- We provide an understanding of the limited expressive power of existing flow-based models in modeling data distributions over complex topological spaces.

- We present a statistical framework for defining an expressive mixture of local normalizing flows that is flexible and generic enough to be used with existing approaches. We show that this framework allows for efficient sampling, inference of latent variables, and exact density evaluation while improving expressivity.

- We validate experimentally that the proposed approach improves flows for density estimation and sample generation, and is thus able to resolve many of the topological restrictions on expressivity imposed by using global diffeomorphisms.

2.2 Global Normalizing Flows

Given data $\{x_n\}_{n=1}^N \subset \mathcal{X} \simeq \mathbb{R}^D$ distributed according to an unknown distribution $p(x)$, a normalizing flow maps it through a diffeomorphism $f : \mathcal{X} \to \mathcal{Z}$ to a latent space $\mathcal{Z} \simeq \mathbb{R}^D$ such that $z = f(x)$ is simply distributed, for example $z \sim q(z)$ where $q = N(0, \mathbb{I})$. Recall that a diffeomorphism is a differentiable map that is bijective and whose inverse is also differentiable. Typically one denotes by g the inverse of f and paramaterizes the normalizing flow as $x = g_\theta(z)$, where θ is the vector of learnable model parameters. The process of going from the latent space to data space is called *generation* or *sampling* and is accomplished by the function g_θ, while the inverse procedure is termed *inference* and is accomplished by $f_\theta = g_\theta^{-1}$:

$$\begin{array}{ll} f_\theta : \mathcal{X} \to \mathcal{Z} & g_\theta : \mathcal{Z} \to \mathcal{X} \\ \underbrace{x \mapsto f_\theta(x)}_{\text{Inference}} & \underbrace{z \mapsto g_\theta(z)}_{\text{Sampling}} \end{array} \tag{2.2.1}$$

The approximation $p_\theta(x)$ to the true probability density $p(x)$ is then obtained from $q(z)$ through the change of variables formula as:

$$p_\theta(x) = q(f_\theta(x)) |\det[J f_\theta(x)]| \tag{2.2.2}$$

As compositions of diffeomorphisms are also diffeomorphisms, one can design expressive flows by composing individual transformations that have simple to compute inverses and Jacobian determinants. Suppressing the vector of model parameters θ, we will use the notation $f(x) = f^1 \circ \cdots \circ f^L(x)$ where $f^1, \ldots, f^L$ are assumed to have easily computable Jacobian determinants and inverses. Define recursively $x^{l-1} = f^l(x^l)$, $1 \leqslant l \leqslant L$, with $x^L = x$. Note that $x^l = f^{l+1} \circ \cdots \circ f^L(x)$ and $x^0 = f(x)$. One can then write the log-likelihood as:

$$\begin{aligned} \log p(x) &= \log q(z) + \log \prod_{l=1}^{L} |\det[Jf^l(x^l)]| \\ &= \log q(f(x)) + \sum_{l=1}^{L} \log |\det[Jf^l(x^l)]| \end{aligned} \tag{2.2.3}$$

A given layer f^l of the normalizing flow will depend only on a subset θ_l of the parameters of $\theta := (\theta_1, \ldots, \theta_L)$. Temporarily adding back in the θ dependence of f_θ, maximum likelihood estimation of θ then yields the following optimization problem:

$$\begin{aligned} \theta^* &= \min_{\theta=(\theta_1,\ldots,\theta_L)} \frac{1}{N} \sum_{n=1}^{N} -\log p_\theta(x_n) \\ &= \min_{\theta=(\theta_1,\ldots,\theta_L)} \frac{1}{N} \sum_{n=1}^{N} \Bigg\{ -\log q(f_\theta(x_n)) \\ &\quad - \sum_{l=1}^{L} \log |\det[Jf^l_{\theta_l}(x^l_n)]| \Bigg\} \end{aligned} \tag{2.2.4}$$

2.3 Related Work

Normalizing flows have come a long way since it was introduced in [49, 51], with much efforts focused on expanding their scalability and applicability. This has resulted in several different formulations [52–55], each with a multitude of proposed architectures [56–61], aimed at

defining expressive yet analytically invertible flow transformations with efficiently computable jacobian determinants. However, as these approaches define invertible transformations in Euclidean space, they are dimensionality preserving and less suited for modeling distributions over lower dimensional manifolds [62, 63]. Subsequent works have tried to address this challenge by building injective flows [50, 64–68]. However, they trade off the benefits of dimensionality change to intractable density estimation or stochastic inverses. The work by [1] overcomes the above limitations using conformal embeddings, but has limited expressive power, as we show in this work. One way to improve the expressivity of all the above approaches, and enable them to overcome topological constraints [69], is to relax their global diffeomorphic requirement by defining a *mixture of flows*. Prior works in this direction have looked at infinite mixtures by defining flows in a lifted space [70] or by using continuous indexing [71]. However, their added expressivity comes at the cost of tractable density computation, and one has to rely on variational approximations to train the model. On similar lines with this work, [72] proposes to use a finite mixture of flows through piecewise-invertible transformations over partitions of the data space by introducing both real and discrete valued latent variables in the flow. However, this formulation introduces discontinuities in the model density that leads to unstable training [71], necessitating the enforcement of boundary conditions through ad-hoc architectural changes. It is therefore limited in its generalizability to novel flow formulations. Our work, on the other hand, by decoupling the partition learning from the flow training, introduces a more generic and scalable framework that can aid existing flows to overcome topological constraints and learn complex data distributions efficiently.

2.4 Local Normalizing Flows

A traditional normalizing flow provides a global diffeomorphism between the latent space $\mathcal{Z}$ and the data space $\mathcal{X} \simeq \mathbb{R}^D$, and as such requires the latent space to have the same dimension as the data space. This can lead to numerical instability when the data is supported on a $d < D$ dimensional manifold $\mathcal{M} \subset \mathcal{X}$ because the learned transformation will tend to become "less and less injective" as it seeks to restrict its range to $\mathcal{M}$ [62, 63].

One way to overcome this challenge is to build transformations that map across dimensions while preserving invertibility on its image. Unfortunately, the natural approach of post-composing a d dimensional bijective normalizing flow $g : \mathcal{Z} \to \mathcal{U}$ with a dimension-raising embedding $e : \mathcal{U} \to \mathcal{X}$ leads in general to an intractable likelihood since the determinant in the change of variables formula $p(x) = q(f(x))|Det[J_g J_e^T J_e J_g]|^{-\frac{1}{2}}$ no longer separates into a product of simpler determinants. We will focus on the solution to this issue developed in [1], namely to post-compose the d dimensional bijective normalizing flow $g : \mathcal{Z} \to \mathcal{U}$ with a dimension raising *conformal embedding* $c : \mathcal{U} \to \mathcal{X}$. An alternative solution developed in [50] is to use a linear dimension raising embedding and invert it stochastically, but this approach relies on the dimension change operation being linear which is restrictive. The approach taken in [1] hinges on the fact that for every $u \in \mathcal{U}$ the Jacobian $J_c(u)$ satisfies $J_c(u)^T J_c(u) = \lambda(u)^2 \mathbb{I}$ for $\lambda : \mathcal{U} \to \mathbb{R}$, thus

$$\begin{aligned} \det[J_{c\circ g}^T J_{c\circ g}]^{\frac{1}{2}} &= \det[J_g^T J_c^T J_c J_g]^{\frac{1}{2}} \\ &= |\lambda(u)| \det[J_g^T J_g]^{\frac{1}{2}} \\ &= |\lambda(u)||\det[J_g]| \end{aligned} \tag{2.4.1}$$

This splitting keeps the likelihood computation tractable, but the requirement that $\mathcal{M}$ be the range of a conformal embedding is artificially restrictive. This issue is exacerbated by the necessity of parameterizing c. As noted in [1] the easiest way to do so is to let $c = c_J \circ \cdots \circ c_1$ where each c_j is either a trivially conformal zero padding operation or a dimension preserving conformal transformation. A dimension preserving conformal transformation $f : \mathbb{R}^d \to \mathbb{R}^d$ with $d > 2$ is restricted by Liouville's theorem to be a Möbius transformation, of the form $f(x) = (A, a, b, \alpha, \epsilon)(x) = b + \alpha(Ax - a)/||Ax - a||^\epsilon$ where $A \in O(d)$ is an orthogonal matrix, $\alpha \in \mathbb{R}$, $a, b \in \mathbb{R}^d$, and ϵ is either 0 or 2. Though it might initially appear that the composition of many such operations would give increased expressive power, the group structure of the Möbius transformations prevents this. Indeed, if $p_s : \mathbb{R}^d \to \mathbb{R}^{d+s}$ is the zero padding operation, $m_1 = (A_1, a_1, b_1, \alpha_1, \epsilon_1)$ is a d dimensional Möbius transformation and $m_2 = (A_2, a_2, b_2, \alpha_2, \epsilon_2)$ is a $d + s$ dimensional Möbius transformation then it is easily verified that for $x \in \mathbb{R}^d$

$$m_2 \circ p_s \circ m_1(x) = (m_2 \cdot \tilde{m_1})(p_s(x)) \tag{2.4.2}$$

Where $\tilde{m_1}$ is the $d + s$ dimensional Möbius transformation

$$\tilde{m_1} = (\begin{bmatrix} A_1 & 0 \\ 0 & \mathbb{I}_{m\times m} \end{bmatrix}, p_s(a_1), p_s(b_1), \alpha_1, \epsilon_1) \tag{2.4.3}$$

Thus, this parametrization yields c as a Möbius transformation of $\mathbb{R}^D$ composed with p_{D-d}. Practically speaking, if c is parameterized as above, the assumption that $\mathcal{M}$ is the image of a global conformal embedding severely limits expressiveness. The class of global conformal embeddings is not subject to Liouville's theorem and is far richer than the set of Möbius transformations, but

it is hard to parameterize.

2.4.1 Motivation: Geometry of conformally flat manifolds

A weaker and more natural assumption than $\mathcal{M}$ being the image of a conformal embedding is that $\mathcal{M}$ is *locally conformally flat*. Recall that if $f : (\mathcal{N}, \eta_1) \to (\mathcal{M}, \eta_2)$ is a map between differentiable manifolds $\mathcal{N}$ and $\mathcal{M}$ with metrics $\eta_1 : \mathcal{N} \times T\mathcal{N} \times T\mathcal{N}$ and $\eta_2 : \mathcal{M} \times T\mathcal{M} \times T\mathcal{M}$ respectively then the pullback $f^*\eta_2$ of the metric η_2 through f is defined via:

$$\begin{aligned} &f^*\eta_2 : \mathcal{N} \times T\mathcal{N} \times T\mathcal{N} \to \mathbb{R} \\ &f^*\eta_2(y, v, w) = \eta_2(f(y), Df(y)(v), Df(y)(w)) \end{aligned} \tag{2.4.4}$$

With this in mind a d dimensional manifold $\mathcal{M}$ is called *locally conformally flat* if $\eta_1 = \sum_{i=1}^{d} dy_i^2$ is the flat metric and for any $x \in \mathcal{M}$ there is a neighborhood $U \ni x$, an open set $O \subset \mathbb{R}^d$, a diffeomorphism $f : O \to U$, and a differentiable scalar function $\lambda : O \to \mathbb{R}$ such that $f^*\eta_2(y, \cdot, \cdot) = \lambda(y)\eta_1(\cdot, \cdot)$ for all $y \in O$ [73]. An alternate definition replaces $\mathbb{R}^d$ with a flat manifold (defined as having an identically vanishing Riemannian curvature tensor), but this definition is equivalent to the above since any d dimensional flat manifold is locally isometric to $\mathbb{R}^d$ (not globally isometric, for example tori are flat when equipped with appropriate coordinates) [37]. In our case the metric η_2 is assumed to be inherited from the Euclidean metric on $\mathcal{X} \simeq \mathbb{R}^D$.

The notion of local conformal flatness provides far more flexibility than its global counterpart. It is well known, for example, that every 2 dimensional Riemannian manifold is locally conformally flat, but even the sphere $S^2(\mathbb{R})$ is not globally conformally flat (by contrast an explicit local conformal equivalence of $S^d(\mathbb{R})$ to $\mathbb{R}^d$ is given by stereographic projection from the north

and south poles) [37]. In general, criteria are known for a Riemannian manifold of dimension $d > 2$ to be locally conformally flat: For $d = 3$ a pseudo-Riemanian manifold is locally conformally flat if and only if the Cotton tensor vanishes everywhere, for $d \geqslant 4$ a pseudo-Riemannian manifold is locally conformally flat if and only if the Weyl tensor vanishes everywhere [37]. The question of which manifolds are globally conformally flat is more difficult, and in applied problems this requirement is artificially restrictive.

2.4.2 A chart based probability model

We thus propose to break up the data manifold $\mathcal{X}$ into an atlas of overlapping charts $U_1, \ldots, U_K$ such that given $x \in \mathcal{X}$ there exists a neighborhood $U_k \ni x$ that may be written as $U_k = c_k(\mathcal{U})$ where c_k is a conformal dimension raising map. Because chart regions may in general overlap, we propose to choose between them probabilistically. In other words we introduce a discrete random variable k taking values in $\{1, \ldots, K\}$ that labels the chart regions and condition the normalizing flow on this quantization of the data space.

Given a collection of charts $U_1, \ldots, U_K$ that cover the data manifold $\mathcal{M}$ on which $p(x)$ is supported, we model $p(x)$ via a latent random variable z that takes values in $\mathcal{Z}$ and a "chart picking" random variable k that takes values in $\{1, \ldots, K\}$. For $k = 1, \ldots, K$ let $g_k : \mathcal{Z} \to U_k$ be a diffeomorphism with inverse $f_k : U_k \to \mathcal{Z}$. Then let the joint distribution of x, z, and k be:

$$p(x, z, k) = \delta(x - g_k(z))q(z)p_k \tag{2.4.5}$$

where $q = N(0, \mathbb{I})$ or $q = \frac{1}{vol(B_1(0))}\mathbb{1}_{B_1(0)}$ and p_k is the normalized frequency with which x occurs

in U_k, that is:

$$p_k := \frac{p(x \in U_k)}{\sum_{j=1}^{K} p(x \in U_j)} = \frac{\int_{U_k} p(x)dx}{\sum_{j=1}^{K} \int_{U_j} p(x)dx} \tag{2.4.6}$$

One may then compute the joint distribution of x and k as

$$\begin{aligned} p(x,k) &= \int_{\mathcal{Z}} p(x,z,k)dz \\ &= p_k \int_{\mathcal{Z}} \delta(x - g_k(z))q(z)dz \\ &= p_k \mathbb{1}_{U_k}(x) \int_{\mathcal{Z}} \delta(z - f_k(x))|\det[Jg_k(z)]|^{-1} q(z)dz \\ &= p_k \mathbb{1}_{U_k}(x)|\det[Jg_k(f_k(x))]|^{-1} q(f_k(x)) \\ &= p_k \mathbb{1}_{U_k}(x)|\det[Jf_k(x)]|q(f_k(x)) \end{aligned} \tag{2.4.7}$$

It is readily verified that $p(z) = q(z)$ and $p(k) = p_k$, in particular:

$$\begin{aligned} p(z) &= \sum_{k=1}^{K} \int_{\mathcal{X}} p(x,z,k)dx \\ &= \sum_{k=1}^{K} p_k \int_{\mathcal{X}} \delta(x - g_k(z))q(z)dx \\ &= q(z) \sum_{k=1}^{K} p_k = q(z) \end{aligned} \tag{2.4.8}$$

and

$$\begin{aligned} p(k) &= \int_X p(x,k)dx \\ &= p_k \int_X \mathbb{1}_{U_k}(x) |\det[Jf_k(x)]| q(f_k(x))dx \\ &= p_k \int_{U_k} |\det[Jf_k(x)]| q(f_k(x))dx \\ &= p_k \int_{\mathcal{Z}} q(z)dz = p_k \end{aligned} \tag{2.4.9}$$

Taken together, (2.4.8) and (2.4.9) yield that z and k are independent random variables since

$$p(z,k) = \int_{\mathcal{X}} p(x,z,k)dx = p_k q(z) = p(k)p(z) \tag{2.4.10}$$

Moreover simply dividing (2.4.7) by $p(k) = p_k$ we conclude that the distribution of x conditioned on a particular chart is given by

$$p(x|k) = \mathbb{1}_{U_k}(x) |\det[Jf_k(x)]| q(f_k(x)) \tag{2.4.11}$$

In particular, $p(x|k)$ is zero unless $x \in U_k$. The density $p(x)$ is then given by

$$\begin{aligned} p(x) &= \sum_{k=1}^{M} p(x|k)p(k) \\ &= \sum_{k:x\in U_k} p_k |\det[Jf_k(x)]| q(f_k(x)) \end{aligned} \tag{2.4.12}$$

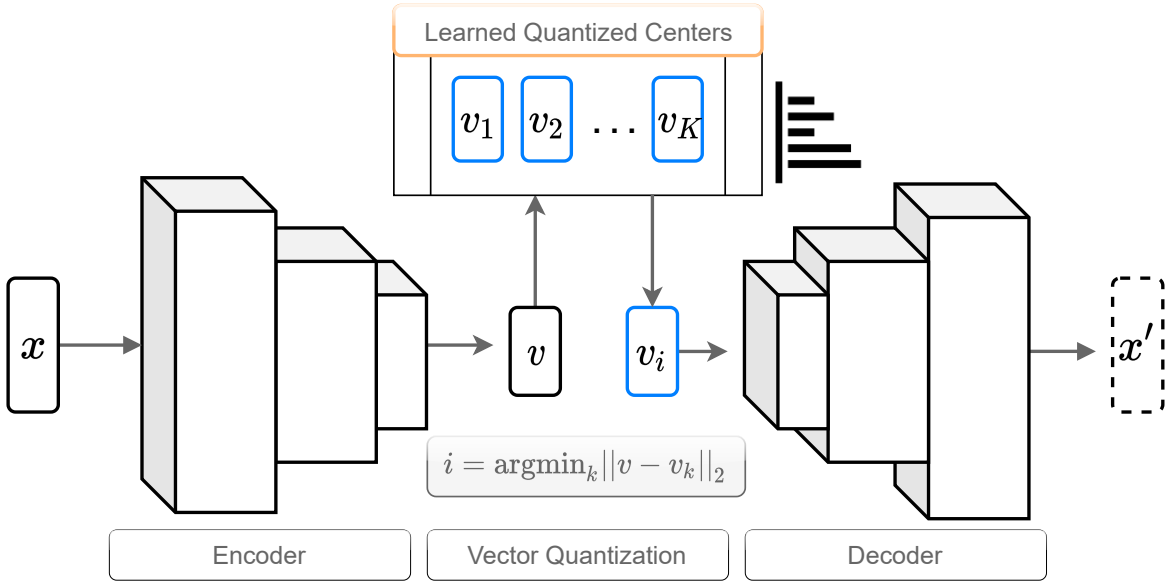


Figure 2.2: Learning quantized centers on the low dimensional data manifold using a vector quantized auto-encoder.

Meanwhile $p(k|x)$ is given by the Bayes' formula as

$$\begin{aligned} p(k|x) &= \frac{p(x|k)p(k)}{\sum_{j=1}^{K} p(x|j)p(j)} \\ &= \frac{p_k \mathbb{1}_{U_k}(x) |\det[Jf_k(x)]| q(f_k(x))}{\sum_{j:x\in U_j} p_j |\det[Jf_j(x)]| q(f_j(x))} \end{aligned} \tag{2.4.13}$$

The distribution $p(k|x)$ is thus also zero unless $x \in U_k$, a fact that will be employed during inference.

Practically speaking, it remains to learn a "good" collection of charts $U_1, \dots, U_K$, estimate $p_1, \dots, p_K$, and then to parameterize $g_1, \dots, g_K$ via normalizing flows $g_1^\theta, \dots, g_K^\theta$ and obtain a maximum likelihood estimate for θ by optimizing $-\log p_\theta(x)$ (where $p_\theta(x)$ is as in (2.4.12)), which we elaborate below.

1. We learn the charts $U_1, \dots, U_K$ via a vector-quantized auto encoder (VQ-AE) [74], as it provides an effective and scalable mechanism to learn quantized centers on lower dimensional

manifolds. The VQ-AE learns an encoder map $E : \mathcal{X} \to \mathcal{V}$, a decoder map $D : \mathcal{V} \to \mathcal{X}$, and a collection of "encoded chart centers" $Q = \{v_k\}_{k=1}^K \subset \mathcal{V}$ that minimize the reconstruction error $\mathcal{L}(D(\text{argmin}_{v \in Q} ||v - E(x)||_2), x)$. Once D, E, and Q are learned we compute $d_k(x) = ||E(x) - v_k||_2$ for $k = 1, \dots K$. With $d_1(x), \dots d_k(x)$ in hand it remains to compute our charts. We would like the charts to overlap, but we also want them to be sparse in the sense that no individual x has too many relevant charts. One possible choice is to fix $m \in \{1, \dots, K\}$ and let $\tilde{d}_1 \leqslant \cdots \leqslant \tilde{d}_K$ be the sorted permutation of $d_1, \dots, d_K$ then define $U_k = \{x : ||E(x) - v_k||_2 \leqslant \tilde{d}_m(x)\}$, so that every point x has at least m charts associated to it (those whose encoded chart centers are among the m closest to $E(x)$). With this choice, a point x will have exactly m associated charts so long as the m^{th} closest chart center is unique. Another choice would be to fix $\epsilon > 0$ and let $U_k = \{x : ||E(X) - v_k||_2 < (1 + \epsilon)\tilde{d}_m(x)\}$ (increasing ϵ enlarges each chart). For now we leave m and ϵ as hyper-parameters, and in general denote $m(x) = |\{k : x \in U_k\}|$ (one always has $m(x) \geqslant m$). Note that checking if $x \in U_k$ amounts to computing $E(x)$ and $\tilde{d}_1(x), \dots, \tilde{d}_K(x)$ and verifying that $||E(x) - v_k||_2 < (1 + \epsilon)\tilde{d}_m(x)$.

2. Once $U_1, \dots, U_K$ are fixed note that if $r_k := p(x \in U_k)$,

$$r_k = \mathbb{E}_{x \sim p(x)}[\mathbb{1}_{U_k}(x)] \tag{2.4.14}$$

The density $p(x)$ is unknown at this point, but we may estimate r_k using the empirical distribution $\rho(x) = \frac{1}{N}\sum_{n=1}^N \delta(x - x_n)$ so that $r_k \approx \mathbb{E}_{x \sim \rho(x)}[\mathbb{1}_{U_k}(x)]$. Practically speaking we thus perform a second pass over the training data and update $r_1, \dots, r_K$ (initialized as zero) via $r_k^{(n)} = \frac{n-1}{n} r_k^{(n-1)} + \frac{1}{n}\mathbb{1}_{U_k}(x_n)$, $1 \leqslant n \leqslant N$, finally setting $r_k = r_k^{(N)}$ and $p_k = r_k / \sum_{j=1}^K r_j$.

3. Once $U_1, \dots, U_K$ and $p_1, \dots, p_K$ are obtained we model $g_k : \mathcal{Z} \to U_k$ as an L lay-

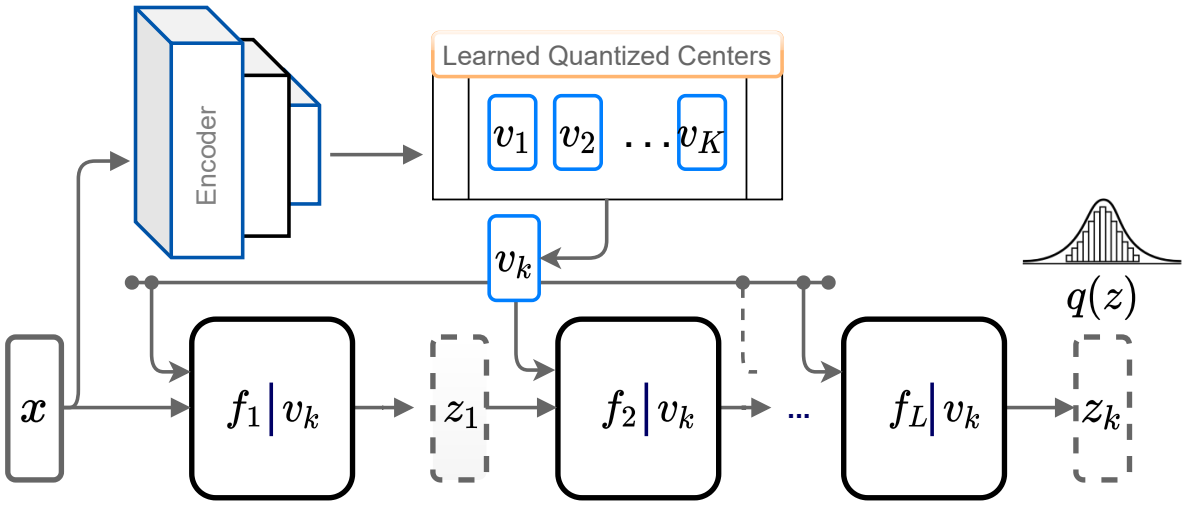


Figure 2.3: Learning the data distribution using a family of normalizing flows conditioned on the quantized centers.

ered invertible conditional normalizing flow. Where dimensionality change is required, we post-compose it with a conformal dimension raising map so that $g_k = c_k \circ g_k^L \circ \cdots \circ g_k^1$. We write the left inverse of g_k via $f_k = f_k^1 \circ \cdots \circ f_k^L \circ c_k^\dagger$ where $f_k^l = (g_k^l)^{-1}$ and $c_k^\dagger$ denotes the left inverse of the conformal map c obtained by removing the zero padding and inverting the various Möbius transformations composing c_k. In practice, we reduce the number of parameters of our model by restricting each g_k^l (and f_k^l) to depend on k only through the value of the encoded chart center v_k. With this parametrization of $f_1, \dots, f_K$ in hand (2.4.11) becomes

$$\begin{aligned} p(x|k) = \mathbb{1}_{U_k}(x) q(f_k(x)) |\lambda_k(c_k^\dagger(x))|^{-1} \\ \prod_{l=1}^{L} |\det[J f_k^l(f_k^{l+1} \circ \cdots \circ f_k^L(x))]| \end{aligned} \tag{2.4.15}$$

where $\lambda_k(u)$ is defined via $(Jc_k(u))^T (Jc_k(u)) = \lambda_k(u)^2 \mathbb{I}$.

As we'll see this approach allows for far higher expressive power than global conformal flows without sacrificing the ability to generate realistic samples, perform inference, or compute

exact densities. Indeed we may rewrite (2.4.12) via

$$\begin{aligned} p(x) &= \sum_{k:x\in U_k} p(x|k)p(k) \\ &= \mathbb{E}_{k\sim\tilde{p}_x(k)}[p(x|k)] \underbrace{\sum_{j:x\in U_j} p(j)}_{\text{piecewise constant}} \end{aligned} \tag{2.4.16}$$

Where $\tilde{p}_x(k) = p(k|p(x|k) > 0) = p(k)/\sum_{j:x\in U_j} p(j)$. Thus, during *training* of the conditional normalizing flow we may replace the expectation $\mathbb{E}_{k\sim\tilde{p}(k)}[p(x|k)]$ with the stochastic quantity $p(x|k), k \sim \tilde{p}(k)$, performing only a single gradient descent pass per data-point as opposed to $m(x)$ passes. If the exact likelihood is needed, however, it can be computed at the cost of evaluating the normalizing flow and its Jacobian $m(x)$ times:

$$\begin{aligned} p(x) &= \sum_{k:x\in U_k} p(x|k)p(k) \\ &= \sum_{k:x\in U_k} p_k q(f_k(x))|\lambda_k(c_k^\dagger(x))|^{-1} \\ &\prod_{l=1}^{L} |\det[Jf_k^l(f_k^{l+1}\circ\cdots\circ f_k^L(x))]| \end{aligned} \tag{2.4.17}$$

Since z and k are independent, one can perform the *sampling task* via first sampling $z \sim q(z)$ and $k \sim p(k)$ and then computing a single forward pass of the normalizing flow chosen by k to obtain $x = g_k(z)$.

The *inference task* is complicated slightly by the fact that z is no longer wholly determined given x, but instead takes values $(f_k(x))_{k:x\in U_k}$ with corresponding probabilities $(p(k|x))_{k:x\in U_k}$. One could perform a stochastic inference via sampling $k \sim p(k|x)$ and computing $z = f_k(x)$ (this amounts to choosing among the relevant charts for x), however if deterministic inference is pre-

ferred then of course one may always compute the expected value of z as $z = \mathbb{E}_{k\sim p(k|x)}[f_k(x)] = \sum_{k:x\in U_k} p(k|x) f_k(x)$ or the most probable value of z as $z = f_s(x)$ where $s = \operatorname{argmax}_{k:x\in U_k} p(k|x)$.

2.4.3 Hard-boundary or deterministic approximation

A particularly simple special case of the above model is the case $m = 1$ and $\epsilon = 0$, in which only a single chart is associated to a given x. This case reduces our atlas of overlapping charts to a disjoint partition of the data manifold $\mathcal{M}$. In this case U_k is exactly the subset of $\mathcal{X}$ for whom $E(x)$ is closest to the encoded chart center v_k, and thus with the exception of x lying on the chart boundaries, the random variable k can be treated as a deterministic function of the random variable x, namely $k(x) = \operatorname{argmin}_{k=1,\dots,K} ||E(x) - v_k||_2 = \sum_{k=1}^{K} k \mathbb{1}_{U_k}(x)$. Sampling in the hard-boundary case is identical to sampling in the soft-boundary case: generate samples for x by first sampling $z \sim q(z)$ and $k \sim p(k)$ and then computing $x = g_k(z)$. Inference in the hard-boundary case is unambiguous since

$$\begin{aligned} &\mathbb{E}_{k\sim p(k|x)}[f_k(x)] = f_s(x) \\ &s = \operatorname*{argmax}_{k=1,\dots,K} p(k|x) = \operatorname*{argmin}_{k=1,\dots,K} ||E(x) - v_k||_2 \end{aligned} \tag{2.4.18}$$

That is to say that one performs inference by first identifying which region R_s contains x and then computing $z = f_s(x)$. The most significant simplification in the hard-boundary case from a

Model	Spherical	Helix	Lissajous	Twisted-Eight	Knotted	Interlocked-Circles
Real NVP	3.15 $\pm$ 0.07	-3.37 $\pm$ 0.16	2.42 $\pm$ 0.07	0.94 $\pm$ 0.15	-2.17 $\pm$ 0.14	0.95 $\pm$ 0.13
VQ-RealNVP	3.55 $\pm$ 0.04	-1.66 $\pm$ 0.08	3.04 $\pm$ 0.15	2.29 $\pm$ 0.14	0.39 $\pm$ 0.18	2.42 $\pm$ 0.25
MAF	4.38 $\pm$ 0.10	-2.90 $\pm$ 0.02	2.50 $\pm$ 0.12	1.34 $\pm$ 0.22	-1.02 $\pm$ 0.14	1.07 $\pm$ 0.07
VQ-MAF	4.43 $\pm$ 0.14	-0.49 $\pm$ 0.03	3.48 $\pm$ 0.16	2.01 $\pm$ 0.10	0.62 $\pm$ 0.16	2.29 $\pm$ 0.18
CEF	0.91 $\pm$ 0.07	-3.71 $\pm$ 0.09	0.42 $\pm$ 0.15	-0.38 $\pm$ 0.21	-2.48 $\pm$ 0.26	-0.72 $\pm$ 0.11
VQ-CEF	0.98 $\pm$ 0.11	-2.90 $\pm$ 0.17	1.65 $\pm$ 0.14	-0.32 $\pm$ 0.19	-1.93 $\pm$ 0.17	1.24 $\pm$ 0.15

Table 2.1: Quantitative evaluation of **Density Estimation** in terms of the test log-likelihood in nats (higher the better) on the 3D datasets. The values are averaged across 5 independent trials, $\pm$ represents the 95% confidence interval.

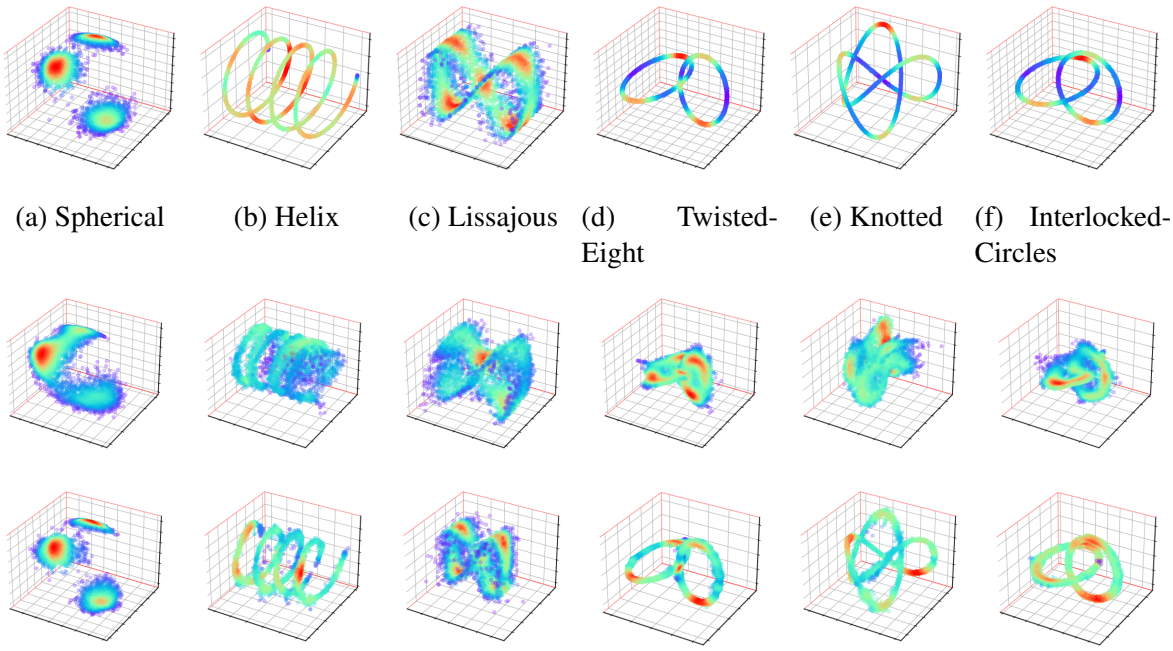
computational standpoint comes in computing the likelihood $p(x)$, since if $x \in U_k$ then

$$\begin{aligned} p(x) &= p(x,k) = p(x|k)p(k) \\ &= p(k)q(f_k(x))|\lambda_k(c_k^\dagger(x))|^{-1} \\ &\prod_{l=1}^{L} |\det[Jf_k^l(f_k^{l+1} \circ \cdots \circ f_k^L(x))]| \end{aligned} \tag{2.4.19}$$

Thus only one normalizing flow needs to be evaluated to compute the exact likelihood $p(x)$ (as opposed to $m(x)$ of them) and the normalizing flows may be trained using the exact likelihood as opposed to an unbiased estimator for it.

2.5 Experiments

To experimentally validate the efficacy of the proposed framework, we consider six 3-dimensional data distributions over manifolds of varying complexity as shown in Figure 2.4. Each dataset consists of $10,000$ datapoints, $5,000$ of which we use for training and $2,500$ each for validation and testing. We train three different normalizing flows - RealNVP [51], Masked Au-



(a) Spherical (b) Helix (c) Lissajous (d) Twisted-Eight (e) Knotted (f) Interlocked-Circles

Figure 2.4: Qualitative visualization of the samples generated by a classical flow - RealNVP (Middle Row) and its VQ-counterpart (Bottom Row) trained on Toy 3D data distributions (Top Row).

toregressive Flows (MAF) [56] and Conformal Embedding Flows (CEF) [1] over these datasets with and without the augmentation of our framework. We refer to a base *flow* augmented with the vector quantized conditioning as VQ-*flow*. We define each model using 5 flow transformations and train them for 100 epochs using an Adam optimizer, early stopping if the validation performance does not improve over 10 epochs. For CEF, we use a 2-dimensional RealNVP as the base flow, which is then raised to the 3-dimensional space using the conformal embedding. We parameterize the VQ-AE using feedforward neural networks and use a latent dimension of 2 with $k = 32$, to learn the partitioning of the data manifold. To define the conditional normalizing flow, we use the parameterization given in [75]. We evaluate the models for density estimation and sample generation. We follow the same hyperparameters for a base flow and its VQ-counterpart without any tuning and report the performance averaged over 5 independent tri-

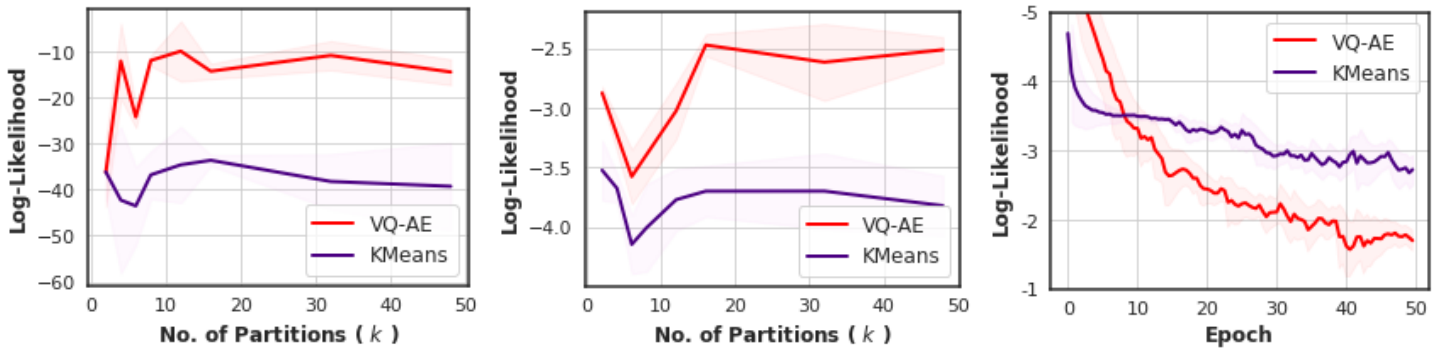
Model	Spherical	Helix	Lissajous	Twisted-Eight	Knotted	Interlocked-Circles
Real NVP	0.50 $\pm$ 0.07	-57.46 $\pm$ 2.11	0.18 $\pm$ 0.14	-2.72 $\pm$ 0.90	-8.65 $\pm$ 0.87	-2.18 $\pm$ 0.37
VQ-RealNVP	0.99 $\pm$ 0.14	-3.85 $\pm$ 0.98	0.59 $\pm$ 0.08	0.18 $\pm$ 0.17	-1.44 $\pm$ 0.37	-0.11 $\pm$ 0.12
MAF	0.65 $\pm$ 0.26	-92.83 $\pm$ 5.69	0.12 $\pm$ 0.16	-2.77 $\pm$ 0.81	-7.04 $\pm$ 0.49	-2.49 $\pm$ 0.14
VQ-MAF	1.01 $\pm$ 0.07	-4.62 $\pm$ 0.37	0.59 $\pm$ 0.07	-0.32$\pm$ 0.13	-2.44 $\pm$ 0.11	-0.15 $\pm$ 0.08
CEF	-1.17 $\pm$ 0.06	-29.90 $\pm$ 2.12	0.38 $\pm$ 0.14	-4.03 $\pm$ 0.38	-19.40 $\pm$ 1.80	-3.42 $\pm$ 0.49
VQ-CEF	0.80 $\pm$ 3.42	-20.75 $\pm$ 2.22	0.49 $\pm$ 0.03	-3.51 $\pm$ 0.73	-14.44 $\pm$ 1.57	-3.23 $\pm$ 0.19

Table 2.2: Quantitative evaluation of **Sample Generation** in terms of the log-likelihood of generated sampes in nats (higher the better) on the 3D datasets. The values are averaged across 5 independent trials, $\pm$ represents the 95% confidence interval.

als. We defer further details on data generation, implementation as well as results on additional 3D data distributions to the supplementary material.

2.5.1 Density Estimation

The ability to compute exact likelihood is one of the critical features of a normalizing flow that makes it a potential tool in solving inverse problems. Improving the expressive power of flows can thus enhance their utility as priors by better modeling the data density. Thus, we first evaluate the proposed framework's ability to enhance the expressivity of flows to perform better density estimation. Table 2.1 compares the log-likelihood (in nats) achieved by different flow models with and without the VQ-augmentation on a held-out test set. A higher value indicates a better learned density. We observe that VQ-flows are able to achieve higher test log-likelihoods than their non-VQ-counterparts consistently across the considered data distributions. Thus, our framework enables better density estimation for normalizing flows over complex manifolds.



(a) Sample Generation (b) Density Estimation (c) Upon further training

Figure 2.5: **Ablation Study** on the effect of the partitioning method and the number of partitions k on sample generation (a) and density estimation (b). (c)-The learning trajectory of the flow for a fixed k(=32), in terms of validation log-likelihood. The shaded region represents the standard deviation over 3 independent trials.

2.5.2 Sample Generation

The other key desiderata of an expressive generative model is its ability to generate high fidelity samples from the data distribution. Figure 2.4 depicts the qualitative visualizations of the samples generated by a RealNVP flow trained on the 3D data distributions with and without the VQ augmentation. We observe that while the classical flow is able to generate samples from the data manifold, it also generates data points off the manifold, resulting in a poorer fit to the real data distribution. This is, in fact, expected due to the expressivity restrictions imposed on its being a global diffeomorphism. On the other hand, VQ-flows are able to overcome these restrictions, better approximate the real data distribution, and generate samples from the data manifold. To further quantitatively establish the efficacy of our framework in improving sample generation, we evaluate the log-likelihood of the generated samples using a kernel density estimator fitted on the training data. We use a gaussian kernel, with an optimal bandwidth obtained through cross-validation for each data distribution. We observe (Table 2.2) that VQ-flows, owing to their

ability to model the topology of the data manifold better, significantly outperform their non-VQ counterparts on sample generation.

2.5.3 Ablation Study

Parameterizing the partitioning function using a VQ-AE is a design choice that we make. Further, the no. of partitions k to consider over the data manifold is a critical hyperparameter underlying the proposed framework. Thus, we conduct ablation experiments to study the sensitivity of the local normalizing flow on k and the partitioning method. We consider k-means clustering as an alternative design choice for the partitioning function. We train a RealNVP flow over the HELIX data distribution using k-means and VQ-AE, across increasing values of k. We plot the validation log-likelihood post training for 25 epochs as a function of k in Figure 2.5. We observe that VQ-AE results in better performance of the flow consistently across k, over k-means. Further, the choice of k beyond a threshold does not have any significant effect on the model, hence it is sufficient to fix it to a large enough value.

2.6 Future Work & Conclusion

Our framework is particularly well suited to high dimensional datasets (such as natural images) that obey the manifold hypothesis, an avenue we hope to explore in the sequel. One of the practical issues we encountered with our approach is that training g_k only on samples from U_k does not always restrict the learned $p(x|k)$ to be supported only on U_k. In such cases, the sum over k such that $x \in U_k$ in (2.4.17) yields an underestimate for $p(x)$, and the total sum $k = 1, \ldots, K$ must be used instead during testing. In the future, we hope to address this issue by explicitly discouraging the generation of samples outside U_k.

To summarize, motivated by differential and conformal geometry, we have developed a novel probabilistic framework for "local" flows. We have demonstrated experimentally on toy data distributions with various topological features that this framework outperforms global flows - both dimension preserving (bijective flows) and dimension raising (embedding flows). Our framework is agnostic to the type of flow transformation employed and retains the key feature of normalizing flows: exact density evaluation. As such, we argue that using local flows as probabilistic chart maps over the data manifold is a natural way to overcome limited expressivity in the presence of dimension change or other topological impediments.

Chapter 3: Higher Order Fourier Transforms[1]

3.1 Introduction

Given a measurable space $(\mathcal{X}, \sigma_{\mathcal{X}}, \mu)$ and $1 \leqslant p < \infty$ denote by $L^p(\mathcal{X}, d\mu)$ the set of measurable functions modulo equivalence almost everywhere (with respect to measure μ) $f : \mathcal{X} \to \mathbb{C}$ such that

$$||f||_{L^p_X} := \left(\int_{\mathcal{X}} |f(x)|^p d\mu(x) \right)^{\frac{1}{p}} < \infty \tag{3.1.1}$$

When there is no confusion about the ambient measure space we will simply write $||f||_{L^p_X}$ as $||f||_p$. Given an exponent $p \in (1, \infty)$ the dual exponent p' is defined so that $\frac{1}{p} + \frac{1}{p'} = 1$, so that the dual of $L^p(\mathcal{X}, d\mu)$ can be identified with $L^{p'}(\mathcal{X}, \mu)$. As usual denote by $L^\infty(\mathcal{X}, d\mu)$ the set of measurable functions such that $||f||_\infty = \text{essup}_{x \in \mathcal{X}} |f(x)| < \infty$ and define $p' = 1$ when $p = \infty$ (we caution that in this case it is not true that $L^p(\mathcal{X}, d\mu)^* = L^{p'}(\mathcal{X}, d\mu)$). Given a second measure space $(\mathcal{Y}, \sigma_{\mathcal{Y}}, \nu)$ define the mixed space $L^p_x L^q_y(\mathcal{X} \times \mathcal{Y}, d\mu d\nu)$ as the set of measurable functions modulo equivalence almost everywhere (with respect to the product measure) $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{C}$

[1] In collaboration with Radu V. Balan and Yonina C. Eldar.

such that

$$||f||_{p,q} := \left(\int_{\mathcal{X}} \left(\int_{\mathcal{Y}} |f(x,y)|^q d\mu(x) \right)^{\frac{p}{q}} d\nu(y) \right)^{\frac{1}{p}} < \infty \tag{3.1.2}$$

Denote by $\mathcal{F}$ the Fourier transform on $L^2(\mathbb{R}^n, dx)$:

$$\begin{aligned} &\mathcal{F} : L^2(\mathbb{R}^n, dx) \ \rightarrow \ L^2(\mathbb{R}^n, dx) \\ &\mathcal{F}[f](k) = \int_{\mathbb{R}^n} e^{-2\pi i \langle k, x \rangle} f(x) dx \end{aligned} \tag{3.1.3}$$

With these definitions in mind, define the Chirp Fourier Transform (CFT) via

$$\begin{aligned} &T : L^2(\mathbb{R}^n, dx) \ \rightarrow \ L^\infty_A L^2_b(\text{Sym}(\mathbb{R}^n) \times \mathbb{R}^n, dAdb) \\ &T[f](A,b) = \mathcal{F}[e^{-2\pi i \langle \cdot, A \cdot \rangle} f](b) = \int_{\mathbb{R}^n} e^{-2\pi i (\langle x, Ax \rangle + \langle b, x \rangle)} f(x) dx \end{aligned} \tag{3.1.4}$$

This definition concurs with that in [76] when $n = 1$. The transform in [76] is not studied directly but is instead employed for the purpose of chirp rate estimation to obtain good chirp parameters to be used in the chirplet transform [77] [78]. The transform thus given is invertible since for any fixed A and almost every $x \in \mathbb{R}^n$

$$f(x) = e^{2\pi i \langle x, Ax \rangle} \mathcal{F}^{-1}[T[f](A, \cdot)](x) \tag{3.1.5}$$

Note further that it is indeed true that for $f \in L^2(\mathbb{R}^n)$ we have $||Tf||_{L^\infty_A L^2_b} < \infty$, since for A fixed Parseval gives

$$||T[f](A, \cdot)||_{L^2_b} = ||e^{-2\pi i \langle \cdot, A \cdot \rangle} f||_2 = ||f||_2 \tag{3.1.6}$$

. As we will see, this transform arises naturally in nuclear magnetic resonance imaging when multiple frequency gradients are used, as well as in connection to an interesting symmetric matrix variant of the free Schrödinger equation. In the context of matrix Schrödinger, it will often be more natural to work with

$$\tilde{T} := T \circ \mathcal{F}^{-1} \tag{3.1.7}$$

Also of interest for applications will be the discrete variant of the above transform. If $\mathbb{Z}_d := \mathbb{Z}/d\mathbb{Z}$ then define

$$\begin{aligned} &T_D : l^2(\mathbb{Z}_d^n) \to l^2(\text{Sym}(\mathbb{Z}_d^n) \times \mathbb{Z}_d^n) \\ &T_D[A, b] := d^{-\frac{n}{2}} \sum_{j \in \mathbb{Z}_d^n} z[j] e^{-2\pi i(\langle j, Aj \rangle + \langle b, j \rangle)/d} \end{aligned} \tag{3.1.8}$$

When $n = 1$ can be T_D may be written as

$$\begin{aligned} &T_D : \mathbb{C}^d \to \mathbb{C}^{d \times d} \\ &T_D z[k, l] = d^{-\frac{1}{2}} \sum_{j=0}^{d-1} z[j] e^{-2\pi i(kj^2 + lj)/d} \end{aligned} \tag{3.1.9}$$

The discrete CFT above is introduced in the case $n = 1$ in [79] in the context of chirp rate estimation. As noted in [79] and in complete analogy with the continuous case, for fixed $A \in \text{Sym}(\mathbb{Z}_d^n)$ the discrete CFT is precisely the multidimensional discrete fourier transform of the

signal $(e^{-2\pi i\langle j,Aj\rangle/d}z[j])_{j\in\mathbb{Z}_d^n}$, and as such is invertible via

$$z[j] = d^{-\frac{n}{2}}e^{2\pi i\langle j,Aj\rangle/d}\sum_{b\in\mathbb{Z}_d^n} Z[A,b]e^{2\pi i\langle b,j\rangle/d} \tag{3.1.10}$$

For any $A \in \mathrm{Sym}(\mathbb{Z}_d^n)$. As in the continuous case the discrete CFT is thus highly redundant.

3.2 Application: Nuclear Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) works by first aligning the spins of atomic nuclei with a strong external magnetic field B_{ext}, then hitting the nuclei with a weaker perpendicular radio-frequency (RF) oscillating magnetic field, then finally measuring the voltage induced by the precession of the nuclear spins caused by the RF field to infer the distribution of various elements in the body [80]. A variety of nuclei may be targeted by MRI, but for the sake of example we may consider commonly targeted spin $1/2$ nuclei such as ${}^1H$. In this case (fixing any axis of measurement) the two available magnetic quantum numbers (the component of the spin magnetic moment along the measurement axis) $m = \frac{1}{2}$ and $m = -\frac{1}{2}$ have equal energy in the absence of an external magnetic field, and so will be approximately equally common. Once the field B_{ext} is switched on, however, the energy associated to a magnetic moment μ becomes $E = -B_{ext}\cdot\mu = B_0\mu_z$ where we have taken (without loss of generality) $B_{ext} = B_0 e_z$ to be along the z axis. Here $\mu_z = \gamma m\hbar$ where γ is the gyromagnetic ratio and $\hbar$ the reduced Planck's constant. Thus there is now an energy difference between the $m = \frac{1}{2}$ (aligned) and $m = -\frac{1}{2}$ (anti-aligned) states of $\Delta E = \gamma\hbar H$, thus the nuclear magnetic moments will thermally align with the external magnetic field in proportion to the strength of the applied field, eventually resulting in an overall parallel net magnetization of the nuclei [80].

The energy difference $\Delta E = \gamma\hbar B_0$ also induces a characteristic Larmor frequency of precession of the spin magnetic moments around the z axis $\omega_0 = \Delta E/\hbar = \gamma B_0$ (for typical field strengths this ω_0 is a radio frequency [81]). This precession occurs on much smaller time-scales than the thermal alignment, and as such can be assumed to have ceased by the time the spins have aligned with the applied field. The intrinsic precession frequency can, however, be exploited via the application of a transverse field (perpendicular to B_{ext}) oscillating at the Larmor frequency: $B_{RF} = B_1 \cos(\omega_0 t) e_x$ (in reality of course, B_{RF} is applied as a pulse and is thus not monochromatic but supported over a thin band of frequencies). In this case, solving the Schrodinger equation induced by the Hamiltonian $\hat{H} = B \cdot \mu = \gamma(B_0 \hat{S}_z + B_1 \cos(\omega_0 t)\hat{S}_x)$ yields the wave function

$$|\psi(t)\rangle = \cos(\frac{\omega_1 t}{2})|0\rangle + e^{i(\omega_0+\pi)} \sin(\frac{\omega_1 t}{2})|1\rangle \tag{3.2.1}$$

Where $\omega_1 = \frac{\gamma}{2} B_1$, $|0\rangle$ is the low energy (aligned) magnetization, and $|1\rangle$ is the high energy (anti-aligned) magnetization (see [82]). This fully describes the dynamics on the Bloch sphere, geometrically (in physical space) (3.2.1) yields that the magnetic moment now has a transverse component and is precessing about the z axis at the Larmor frequency ω_0, and that the polar angle of this precession is also oscillating (more slowly) with frequency ω_1 (incidentally, the reason it is oscillating with ω_1 and not $\frac{\omega_1}{2}$ is precisely the so-called "Dirac belt-trick," in which two full polar rotations on the Bloch sphere are equivalent to a single rotation in physical space). The probability of measuring the nuclear spin to be aligned (or anti-aligned) thus varies proportionally to $|\cos(\frac{\omega_1 t}{2})|^2$. This flipping between low and high energy states periodically "sucks energy" out of the RF field and can be easily detected as it induces an AC current in the receiver coil [3],

yielding the NMR signal.

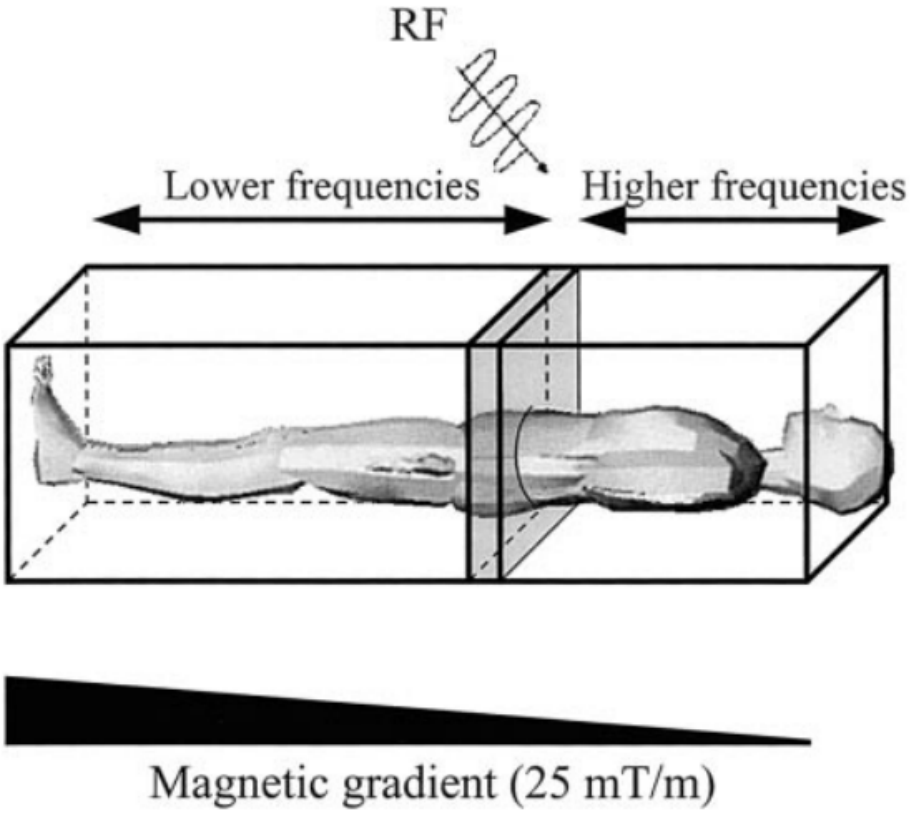


Figure 3.1: From [3]. The strength of the RF pulse varies in space.

While in this setup an RF sinc pulse of bandwidth ω_b allows for the determination of the elements present, it does not give information on their spatial distribution [80]. In order to do obtain spatial information, one approach is to replace the constant field B_{ext} with a magnetic gradient $B_{ext} = (B_0 + gz)e_z$ (see Figure 3.1). Because the magnetic field gradient varies spatially and the Larmor frequency varies proportionally to the strength of the field, only a thin slice of width $\delta \propto \omega_b/g$ resonates with the RF pulse [81]. An issue with this approach is that the resolution of the image is constrained by the bandwidth of the RF pulse (assuming a fixed gradient g), and that as a result the peak power of the RF pulse grows *quadratically* in the resolution [83]. A solution to this problem is to use a frequency modulated RF signal as in Figure 3.2 in which

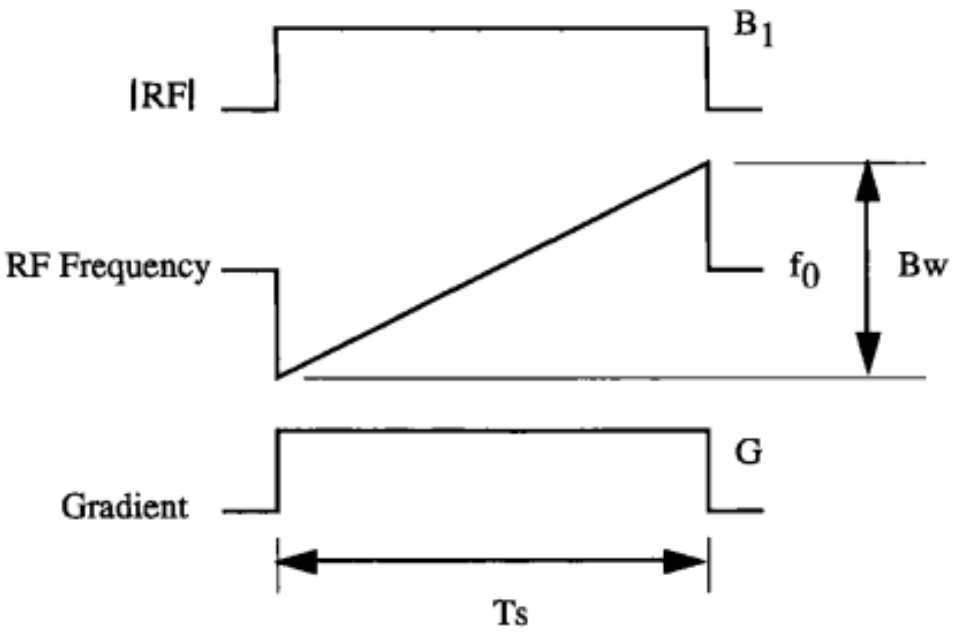


Figure 3.2: From [4]. A frequency modulated RF pulse.

the frequency varies linearly in time. In this case the peak power of the RF pulse grows only linearly in the resolution, allowing for cheaper and simpler RF amplifiers to be used [83]. In the frequency modulated case, the spatial information is encoded in the *quadratic phase* of the NMR signal. In particular, we note that the spatial dependence of the phase of the transverse excitation in the presence of a constant magnetic gradient $B_{ext} = (B_0 + gz)e_z$ will be given by

$$\phi_g = 2\pi\gamma g z T \tag{3.2.2}$$

where T is the duration of the RF pulse [4] (the local Larmor frequency here is simply $\gamma B_0 + \gamma g z$).

In the case of a variable frequency RF pulse, as in Figure 3.2 where

$$\begin{cases} f(t) = f_0 + \frac{Bw}{T_s}(t - \frac{T_s}{2}) & 0 \leqslant t \leqslant T_s \\ f(t) = f_0 & \text{otherwise} \end{cases} \tag{3.2.3}$$

the time of excitation T will depend on the position of the spin along the direction of the applied gradient [4]. In particular, for the pulse shown in Figure 3.2 we will have $T(z) = T_s(\frac{1}{2} - \frac{z-z_0}{D_s})$ where z_0 is the position of the spin that has Larmor frequency f_0 and D_s is the (spatial) width of the RF pulse: $D_s = \frac{Bw}{\gamma g}$ [4]. Thus the phase shift of the transverse excitation (which corresponds to a phase shift in the measured AC current) depends quadratically on the position as:

$$\phi'_{RF} = 2\pi\gamma g T_s(\frac{1}{2} - \frac{z - z_0}{D_s})z \tag{3.2.4}$$

An additional spatially varying contribution to the phase arises due to the constant phase change of the excitation pulse [4]:

$$\phi''_{RF} = 2\pi \int_{T(z)}^{T_s} f(t)dt = \frac{\pi Bw}{T_s}(T_s - T(z))^2 + 2\pi(f_0 - \frac{Bw}{2})(T(z) - T_s) \tag{3.2.5}$$

Combining these two contributions, simplifying, and dropping constant terms one finds a spatially varying phase contribution:

$$\phi_{RF} = \pi\gamma g T_s(x + \frac{(x - x_0)^2}{D_s}) \tag{3.2.6}$$

The linear term may in fact be removed by a gradient pulse with the appropriate area [4] [3], thus

measuring at multiple center locations $x_0^{(n)} = x_0 + n\delta$ one would like to be able to reconstruct a signal $\phi(x)$ (the true NMR profile) from measurements of the form

$$s(Q, n) = \int \phi(x) e^{-2\pi i Q(x-n\delta)^2} dx \tag{3.2.7}$$

Where we take $x_0 = 0$ without loss of generality and $Q = \frac{\gamma g T_s}{2D_s}$ is known as the "second order coefficent" arising from the RF pulse. Up to an overall constant phase, we can write $s(Q, n)$ as a multiple of the one dimensional CFT

$$T[\phi](A, b) = \int \phi(x) e^{-2\pi i(Ax^2+bx)} dx \tag{3.2.8}$$

In particular, if $A = Q$ and $b = -2Q\delta n$ then

$$s(Q, n) = e^{-2\pi i \frac{b^2}{4A}} T(A, b) \tag{3.2.9}$$

Thus if we would like to be able to reconstruct the NMR profile from measurements of different central locations (x_0 or correspondingly f_0) and different frequency gradients Bw/T_s, it behooves us to study the properties of the Chirp Fourier Transform (CFT).

3.3 Connection to the Linear Canonical Transform

An additional motivation for introducing the Chirp Fourier Transform is that it provides a novel perspective on the much celebrated Linear Canonical Transform (LCT). The LCT is a generalization of the Fourier transform that appears in applications such as paraxial wave optics

and accelerator physics [84] [85]. Given $L_{oo}, L_{ii} \in \text{Sym}(\mathbb{R}^n)$ and $L_{io} \in \text{GL}(\mathbb{R}^n)$ the LCT is defined as

$$\begin{aligned} &LCT : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n) \\ &LCT[f](k) = (i^{-1} \det L_{io})^{\frac{1}{2}} \int_{\mathbb{R}^n} e^{\pi i (k^T L_{oo} k - 2k^T L_{io} x - x L_{ii} x)} f(x) dx \end{aligned} \tag{3.3.1}$$

Like the Fourier transform, the LCT is unitary in the sense that

$$\langle LCT[f], LCT[g] \rangle_{L^2(\mathbb{R}^n)} = \langle f, g \rangle_{L^2(\mathbb{R}^n)} \tag{3.3.2}$$

While the Fourier transform can be thought of as a $\frac{\pi}{2}$ rotation in phase space, there is an LCT for each metaplectic transformation of phase space. In particular, since $\text{Mp}(2n, \mathbb{R})$ is a double cover of the symplectic group $\text{Sp}(2n, \mathbb{R})$ there are two LCTs for a given element of the symplectic group [84]. Recall that

$$\text{Sp}(2n, \mathbb{R}) = \{M \in \mathbb{R}^{2n \times 2n} \mid M^T \begin{bmatrix} 0 & -\mathbb{I}_{n \times n} \\ \mathbb{I}_{n \times n} & 0 \end{bmatrix} = \mathbb{I}_{2n \times 2n}\} \tag{3.3.3}$$

To obtain the phase space deformation associated with a given linear canonical transform, one can compute the Wigner distribution (the uncertainty principle precludes the existence of an instantaneous time-frequency distribution, however for a large class of signals the Wigner distribution provides the highest possible time-frequency resolution allowable within the bounds of the un-

certainty principle [86]):

$$W_f(x,k) := \int_{\mathbb{R}^n} f(x+\frac{1}{2}x')\overline{f(x-\frac{1}{2}x')}e^{-2\pi i\langle k,x'\rangle}dx' \tag{3.3.4}$$

In this case we can compute the effect of the LCT on the Wigner distribution via

$$\begin{aligned} W_{LCT[f]}(x,k) &= \int_{\mathbb{R}^n} LCT[f](x+\frac{1}{2}x')\overline{LCT[f](x-\frac{1}{2}x')}e^{-2\pi i\langle k,x'\rangle}dx' \\ &= \int_{\mathbb{R}^n} f(A_{11}x+A_{12}k+\frac{1}{2}x')\overline{f(A_{11}x+A_{12}k-\frac{1}{2}x')}e^{-2\pi i\langle (A_{21}x+A_{22}k),x'\rangle}dx' \\ &= W_f(Ax+Bk,Cx+Dk) \end{aligned} \tag{3.3.5}$$

Where $A_{11} = L_{io}^{-1}L_{ii}$, $A_{12} = L_{i0}^{-1}$, $A_{21} = L_{oo}L_{io}^{-1}L_{ii} - L_{io}^T$, and $A_{22} = L_{oo}L_{io}^{-1}$ (see the appendix of [84] for the full computation). Thus the LCT is associated to the transformation M of phase space:

$$\begin{bmatrix} x \\ k \end{bmatrix} \rightarrow \underbrace{\begin{bmatrix} L_{io}^{-1}L_{ii} & L_{i0}^{-1} \\ L_{oo}L_{io}^{-1}L_{ii} - L_{io}^T & L_{oo}L_{io}^{-1} \end{bmatrix}}_{M} \begin{bmatrix} x \\ k \end{bmatrix} \tag{3.3.6}$$

If we allow the triple (L_{ii}, L_{io}, L_{oo}) to vary and write $LCT[f] = LCT(L_{ii}, L_{io}, L_{oo})[f]$ then, returning the to the CFT:

$$T[f](A,b) = LCT(0,\mathbb{I},2A)[f](b) \tag{3.3.7}$$

Thus for fixed A, the CFT is associated to the transformation of phase space

$$M = \begin{bmatrix} 2A & \mathbb{I} \\ -\mathbb{I} & 0 \end{bmatrix} \tag{3.3.8}$$

And for fixed A, the transformation $\tilde{T} = T \circ \mathcal{F}^{-1}$ is associated to a shear transformation of phase space:

$$M = \begin{bmatrix} \mathbb{I} & 2A \\ 0 & \mathbb{I} \end{bmatrix} \tag{3.3.9}$$

Thus the exact nature of the redundancy present in the CFT is that it corresponds to all possible shearings of phase space (only one of which is required to recover the signal f in the absence of noise, owing to the unitarity of the LCT).

3.4 Connection to Matrix Schrödinger

Choosing units so that $\frac{\hbar}{m} = \frac{1}{\pi}$, and using time variable a and space variable b (for reasons that will be obvious) the free Schrödinger equation becomes

$$i\partial_a \psi(a,b) = -\frac{1}{2\pi}\nabla_b^2 \psi(a,b) \tag{3.4.1}$$

It is easily verified that for sufficiently well behaved initial data, this equation has as its general solution:

$$\psi(a,b) = \int_{\mathbb{R}^n} e^{-2\pi i(a||k||^2 + \langle b,k\rangle)} \check{\psi}(0,k) dk \tag{3.4.2}$$

Thus the behavior of the wave function for the free particle is determined at all times by its initial momentum space wave function integrated against the kernel $K(a,b) = e^{-2\pi i(a||k||^2+\langle b,k\rangle)}$, which is precisely the kernel for the CFT when the chirp matrix A is the multiple of identity $A = a\mathbb{I}$. Evidently,

$$\psi(a,b) = T[\check{f}(0,\cdot)](a\mathbb{I}, b) \tag{3.4.3}$$

This clarifies exactly the nature of the redundancy of the CFT when $A = a\mathbb{I}$, since of course the behavior of the wave function is determined at all future (and past) times given its value at any particular time. We might hope, therefore, that the the redundancy present in the general CFT (for abitrary symmetric A) arises exactly from an underlying PDE, as it does when $A = a\mathbb{I}$. And indeed, we define *the free matrix Schrödinger equation* by

$$iD_A^{\text{Sym}}\psi(A,b) = -\frac{1}{2\pi} H_b \psi(A,b) \tag{3.4.4}$$

Where D_A^{Sym} is the gradient operator on the vector space of symmetric matrices, namely $(D_A^{\text{Sym}}\psi)_{ij} = \frac{1}{2}(\frac{\partial}{\partial A_{ij}} + \frac{\partial}{\partial A_{ji}})\psi$ and H_b is the Hessian operator with respect to b, specifically $(H_b\psi)_{ij} = \frac{\partial^2}{\partial b_i \partial b_j}\psi$.

In this case let

$$\psi(A,b) = \tilde{T}[\psi(0,\cdot)](A,b) = T[\check{\psi}(0,\cdot)](A,b) = \int_{\mathbb{R}^n} e^{-2\pi i(\langle k,Ak\rangle + \langle b,k\rangle)}\check{\psi}(0,k)dk \tag{3.4.5}$$

Then, assuming every entry in $\check{\psi}(0,k)(\mathbb{I} + kk^T)$ is an L^1 function we obtain by Lebesgue dominated convergence:

$$\begin{aligned} iD_A^{\text{Sym}}\psi(A,b) &= \int_{\mathbb{R}^n} i(-2\pi i)kk^T e^{-2\pi i(\langle k,Ak\rangle + \langle b,k\rangle)}\check{\psi}(0,k)dk \\ &= -\frac{1}{2\pi}H_b\psi(A,b) \end{aligned} \tag{3.4.6}$$

And moreover ψ satisfies the boundary condition $\tilde{T}[\psi(0,\cdot)](0,b) = \psi(0,b)$. Thus as expected the transform coordinates A and b are not independent but are related by the free matrix Schrödinger partial differential equation. This connection to Schrödinger also yields an important representation formula for Tf vis a vis the matrix analog of the Schrodinger propagator. In particular,

$$\begin{aligned} T[\check{\psi}](A,b) &= \mathcal{F}[e^{-2\pi i\langle\cdot,A\cdot\rangle}\check{\psi}](b) \\ &= \underbrace{\widehat{e^{-2\pi i\langle\cdot,A\cdot\rangle}}}_{:=K(A,b)} * \psi(b) \end{aligned} \tag{3.4.7}$$

The propagator $K(A,b)$ can be computed via a contour integral when A is invertible:

$$K(A,b) = \int_{\mathbb{R}^n} e^{-2\pi i\langle b,k\rangle}e^{-2\pi i\langle k,Ak\rangle} \tag{3.4.8}$$

Write $A = U\Lambda U^T$ where $U \in O(n)$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is diagonal, then make the change of coordinates $\tilde{k} = U^T k$ and let $\gamma_i = (U^T b)_i$ so that

$$K(A,b) = \int_{\mathbb{R}^n} e^{-2\pi i \sum_{j=1}^n \gamma_j k_j + \lambda_j k_j^2} dk \tag{3.4.9}$$

$$= \prod_{j=1}^n \int_{-\infty}^{\infty} e^{-2\pi i(\gamma_i s_i + \lambda_i s_i^2)} \tag{3.4.10}$$

We then note that

$$\int_{-\infty}^{\infty} e^{-2\pi i(\gamma s + \lambda s^2)} ds = \frac{2}{\sqrt{2\pi|\lambda|}} e^{\frac{\pi i}{2}\gamma^2/\lambda} \int_0^{\infty} e^{-iz^2} dz = \frac{1}{\sqrt{2i|\lambda|}} e^{\frac{\pi i}{2}\gamma^2/\lambda} \tag{3.4.11}$$

Where in the last step an arc contour with angle $-\pi/4$ yields $\int_0^{\infty} e^{-z^2} dz = \sqrt{\frac{\pi}{i}}$. Multiplying the n copies of this integral together one obtains

$$K(A,b) = \frac{1}{\sqrt{(2i)^n |A|}} e^{\frac{\pi i}{2}\langle b, A^{-1} b\rangle} \tag{3.4.12}$$

Where $|A| := |\det A|$. Thus an alternate form of the CFT is $\tilde{T}[f] = K * f$ or

$$Tf(A,b) = (K(A,\cdot) * \hat{f})(b) \tag{3.4.13}$$

3.5 Strichartz Estimates for Matrix Schrodinger

The connection between the transform T and the free matrix Schrödinger PDE gives us hope that we should be able to use a variant of the homogeneous Strichartz estimate to further constrain the range of T. In particular a pair of exponents (q,p) is called admissible if $2 \leqslant q, p \leqslant$

∞, $(q,p,n) \neq (2,\infty,2)$, and $\frac{2}{q} + \frac{1}{p} = \frac{1}{2}$. Then when $A = a\mathbb{I}$ for any admissible exponents (q,p) the homogeneous Strichartz estimate states that there exists constant $C_{q,r,n}$ such that

$$||Tf||_{L^q_a L^r_b(\mathbb{R}\times\mathbb{R}^n)} \leqslant C_{q,r,n}||f||_{L^2(\mathbb{R})} \tag{3.5.1}$$

In other words, T is a bounded operator from $L^2(\mathbb{R})$ to $L^q_a L^p_b(\mathbb{R}\times\mathbb{R}^n)$ [87]. We would like to prove an analogous result when A is not assumed to be a multiple of identity. It will be instructive to first consider the proof of the homogeneous Strichartz estimate in the usual setting where $A = a\mathbb{I}$. In particular we would like to show for (q,p,n) admissible that

$$||\int_{\mathbb{R}^n} e^{-2\pi i(a||x||^2+\langle b,x\rangle)} f(x)dx||_{L^q_a L^p_b(\mathbb{R}\times\mathbb{R}^n)} \leqslant C_{q,p,n}||f||_2 \tag{3.5.2}$$

It will be convenient to work with $\tilde{T} = T \circ \mathcal{F}^{-1}$ instead of T, replacing f with its inverse Fourier transform on the left hand side and instead showing

$$||\int_{\mathbb{R}^n} e^{-2\pi i(a||x||^2+\langle b,x\rangle)} \check{f}(x)dx||_{L^q_a L^p_b(\mathbb{R}\times\mathbb{R}^n)} \leqslant C_{q,p,n}||f||_2 \tag{3.5.3}$$

This latter estimate is equivalent to the first, owing to the unitarity of the Fourier transform. We will follow [87] in proving the homogeneous Strichartz estimate vis a vis the dual homogeneous Strichartz estimate. In particular we compute $\tilde{T}^* : L^{q'}_a L^{p'}_b(\mathbb{R}\times\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ via the property

that $\langle F, \tilde{T} f\rangle = \langle \tilde{T}^* F, f\rangle$ for any $F \in L^q_a L^p_b(\mathbb{R}\times\mathbb{R}^n)$ and $f \in L^2(\mathbb{R}^n)$:

$$\begin{aligned}\langle F, \tilde{T} f\rangle &= \int_{\mathbb{R}}\int_{\mathbb{R}^n} F(a,b)\overline{(\mathcal{F}e^{-2\pi i a||\cdot||^2}\mathcal{F}^{-1}f)(b)}dbda \\ &= \int_{\mathbb{R}}\langle F(a,\cdot), \mathcal{F}e^{-2\pi i a||\cdot||^2}\mathcal{F}^{-1}f\rangle_{L^2(\mathbb{R}^n)}da \\ &= \int_{\mathbb{R}}\langle \mathcal{F}e^{2\pi i a||\cdot||^2}\mathcal{F}^{-1}F(a,\cdot), f\rangle_{L^2(\mathbb{R}^n)}da \\ &= \Big\langle \int_{\mathbb{R}} \mathcal{F}e^{2\pi i a||\cdot||^2}\mathcal{F}^{-1}F(a,\cdot)da, f\Big\rangle_{L^2(\mathbb{R}^n)}\end{aligned} \tag{3.5.4}$$

Where we employ Fubini in the last step. Since this holds for all $f \in L^2(\mathbb{R}^n)$ we find that $\tilde{T}^*$ is given by

$$\begin{aligned}&\tilde{T}^* : L^{q'}_a L^{p'}_b(\mathbb{R}\times\mathbb{R}^n) \to L^2(\mathbb{R}^n) \\ &\tilde{T}F(y) = \int_{\mathbb{R}}\int_{\mathbb{R}^n} e^{-2\pi i(-a||k||^2+\langle y,k\rangle)}\check{F}(a,k)dkda\end{aligned} \tag{3.5.5}$$

Employing the fact that $||\tilde{T}||_{L^2(\mathbb{R})\to L^q_a L^r_b(\mathbb{R}\times\mathbb{R}^n)} = ||\tilde{T}^*||_{L^{q'}_a L^{p'}_b(\mathbb{R}\times\mathbb{R}^n)\to L^2(\mathbb{R}^n)}$ it thus suffices to prove the *dual homogeneous Strichartz estimate*

$$||\tilde{T}^* F||_{L^2(\mathbb{R}^n)} \leqslant C_{p,q,n}||F||_{L^{q'}_a L^{p'}_b(\mathbb{R}\times\mathbb{R}^n)} \tag{3.5.6}$$

in order to show (3.5.1). In fact, we will seek an estimate of the form

$$||\tilde{T}\tilde{T}^* F||_{L^q_a L^p_b(\mathbb{R}\times\mathbb{R}^n)} \lesssim_{p,q,n} ||F||_{L^{q'}_a L^{p'}_b(\mathbb{R}\times\mathbb{R}^n)} \tag{3.5.7}$$

since in this case

$$\begin{aligned} ||\tilde{T}^* F||^2_{L^2(\mathbb{R}^n)} &= \langle \tilde{T}^* F, \tilde{T}^* F \rangle_{L^2(\mathbb{R}^n)} \\ &= \langle F, \tilde{T}\tilde{T}^* F \rangle \\ &\leqslant ||F||_{L^{q'}_a L^{p'}_b(\mathbb{R}\times\mathbb{R}^n)} ||\tilde{T}\tilde{T}^* F||_{L^q_a L^p_b(\mathbb{R}\times\mathbb{R}^n)} \\ &\lesssim_{p,q,n} ||F||^2_{L^{q'}_a L^{p'}_b} \end{aligned} \tag{3.5.8}$$

And we will have proved the homogeneous Strichartz estimate. In order to show (3.5.7) we will interpolate between two "fixed time" estimates. The first is an equality rather than an inequality, namely that that $\tilde{T}$ is an isometry: $||\tilde{T}f(a,\cdot)||_{L^2(\mathbb{R}^n)} = ||f||_{L^2(\mathbb{R}^n)}$. The second estimate comes from the propagator form $\tilde{T}f = K * f$. In the case $A = a\mathbb{I}$ this form yields

$$\begin{aligned} |\tilde{T}f(a,b)| &= |\int_{\mathbb{R}^n} (2i|a|)^{-\frac{n}{2}} e^{\frac{\pi i}{2a}||y||^2} f(x-y)dy| \\ &\leqslant (2|a|)^{-\frac{n}{2}} ||f||_{L^1_b(\mathbb{R}^n)} \end{aligned} \tag{3.5.9}$$

Thus we have the additional fixed time estimate $||\tilde{T}f||_{L^\infty_b(\mathbb{R}^n)} \leqslant (2|a|)^{-\frac{n}{2}} ||f||_{L^1_b(\mathbb{R}^n)}$. We can combine these two estimates via Marcinkiewicz interpolation to obtain the following family of fixed time estimates for $r \in [1,2]$:

$$||\tilde{T}f||_{L^{r'}_b(\mathbb{R}^n)} \leqslant (2|a|)^{\frac{n}{2}-\frac{n}{r}} ||f||_{L^r(\mathbb{R}^n)} \tag{3.5.10}$$

In addition to this fixed time estimate we will need the Hardy-Littlewood-Sobolev fractional integration estimate:

$$||||\cdot||^{-\alpha} * f||_{L^s(\mathbb{R}^d)} \lesssim_{s,v,d,\alpha} ||f||_{L^v(\mathbb{R}^d)} \tag{3.5.11}$$

Where $\frac{1}{v} = \frac{1}{s} + \frac{d-\alpha}{d}$. With this in mind we can compute

$$\begin{aligned} ||\tilde{T}\tilde{T}^* F||_{L_a^q L_b^p(\mathbb{R}\times\mathbb{R}^n)} =& \left|\left| \int_{\mathbb{R}} \int_{\mathbb{R}^n} e^{-2\pi i[(a-s)||k||^2 + \langle b,k\rangle]} \check{F}(s,k) dk ds \right|\right|_{L_a^q L_b^p(\mathbb{R}\times\mathbb{R}^n)} \\ \leqslant& \left|\left| \int_{\mathbb{R}} \left|\left| \int_{\mathbb{R}^n} e^{-2\pi i[(a-s)||k||^2 + \langle b,k\rangle]} \check{F}(s,k) dk \right|\right|_{L_b^p(\mathbb{R}^n)} ds \right|\right|_{L_a^q(\mathbb{R})} \end{aligned} \tag{3.5.12}$$

The innermost integral here is of non other than $(\tilde{T}F(s,\cdot))(a-s,b)$, thus we can employ our fixed time estimate and conclude

$$||\tilde{T}\tilde{T}^* F||_{L_a^q L_b^p(\mathbb{R}\times\mathbb{R}^n)} \leqslant \left|\left| \int_{\mathbb{R}} (2|a-s|)^{n(\frac{1}{2}-\frac{1}{p'})} ||F(s,\cdot)||_{L_b^{p'}(\mathbb{R}^n)} ds \right|\right|_{L_a^q(\mathbb{R})} \tag{3.5.13}$$

Finally, we note that this is a convolution with a kernel of the form $|\cdot|^{-\alpha}$, and thus employ Hardy-Littlewood-Sobolev with $\alpha = \frac{n}{p'} - \frac{n}{2}$, $d = 1$, $s = q$, and $v = q'$ (it is easy to check that this choice for v results in the required identity $\frac{1}{v} = \frac{1}{s} + \frac{d-\alpha}{d}$):

$$||\tilde{T}\tilde{T}^* F||_{L_a^q L_b^p(\mathbb{R}\times\mathbb{R}^n)} \lesssim_{p,q,n} ||F||_{L_a^{q'} L_b^{p'}(\mathbb{R}\times\mathbb{R}^n)} \tag{3.5.14}$$

This concludes the proof of the homogeneous Strichartz estimate, in other words the proof that when the chirp matrix $A = a\mathbb{I}$ is restricted to be a multiple of identity T is a bounded operator from $L^2(\mathbb{R})$ to $L_a^q L_b^p(\mathbb{R}\times\mathbb{R}^n)$. Towards a generalization of this result to arbitrary chirp matrices,

we will prove the following the following:

Theorem 3.5.1. *(Dual homogeneous Strichartz for matrix Schrödinger)*

Let $2 \leqslant p, q \leqslant \infty$ such that $\frac{2}{q} + \frac{1}{p} = \frac{1}{2}$. Define the null aliasing operator τ on $L_A^{q'}(Sym(\mathbb{R}^n))$ via

$$\tau g(A,t) = \sum_{\epsilon \in \{-1,1\}^n} \int_{Z(\det)} g(A + \sigma_\epsilon(\sqrt{Z^2 + \lambda_Z(t)\mathbb{I}}))dZ \tag{3.5.15}$$

where for $\epsilon \in \{-1,1\}^n$

$$\begin{aligned} &\sigma_\epsilon : \mathbb{P}(n) \to Sym(\mathbb{R}^n) \\ &\sigma_\epsilon(U\Lambda U^T) = U \operatorname{diag}(\epsilon_1, \ldots, \epsilon_n)\Lambda U^T \end{aligned} \tag{3.5.16}$$

and $\lambda_Z(t)$ is the unique positive increasing function such that

$$(\mu_1 + \lambda_Z(t)) \cdots (\mu_n + \lambda_Z(t)) = t^2 \tag{3.5.17}$$

where $\mu_1, \ldots, \mu_n$ are the eigenvalues of Z^2. Then if $F \in L_A^{q'} L_b^{p'}(Sym(\mathbb{R}^n) \times \mathbb{R}^n)$ satisfies the growth condition

$$\tau||F||_{L_b^{p'}(\mathbb{R}^n)}(A,t) \lesssim ||F(A + t\mathbb{I}, \cdot)||_{L_b^{p'}(\mathbb{R}^n)} \tag{3.5.18}$$

We have:

$$||\tilde{T}^* F||^2_{L^2(\mathbb{R}^n)} \lesssim_{p,q,n} ||F||_{L_A^{q'} L_b^{p'}(Sym(\mathbb{R}^n)\times\mathbb{R}^n)}|| \cdot ||F||_{L_A^q L_t^{q'} L_b^{p'}(V_I \times V_T \times \mathbb{R}^n)} \tag{3.5.19}$$

Where $V_I = \{\lambda \mathbb{I} \mid \lambda \in \mathbb{R}\}$ *is the identity subspace of Sym*$(\mathbb{R}^n)$ *and* $V_T = \{A \in$ *Sym*$(\mathbb{R}^n) \mid$ *tr*$\{A\} = 0\}$ *the traceless subspace.*

Proof. The proof proceeds in the same manner as in the multiple of identity case, up until it comes to computing fixed time estimates. The first estimate, namely unitarity of $\tilde{T}$, is identical. For the second estimate, however, we employ the fact that $\tilde{T}f = K * f$ where K is as in (3.4.12) and find that for all A invertible and all $b \in \mathbb{R}^n$ we have

$$|\tilde{T}F(A,b)| \leqslant (2^n|A|)^{-\frac{1}{2}}||f||_{L^1(\mathbb{R}^n)} \tag{3.5.20}$$

In this case we can again employ Marcinkiewicz to obtain for $r \in [1,2]$ the estimate

$$||\tilde{T}f||_{L_b^{r'}(\mathbb{R}^n)} \leqslant (2^n|A|)^{\frac{1}{2}-\frac{1}{r}}||f||_{L_b^r(\mathbb{R}^n)} \tag{3.5.21}$$

We continue as before with the $\tilde{T}\tilde{T}^*$ estimate until we arrive at

$$||\tilde{T}\tilde{T}^*F||_{L_A^q L_b^p(\text{Sym}(\mathbb{R}^n)\times\mathbb{R}^n)} \leqslant \left|\left| \int_{\text{Sym}(\mathbb{R}^n)} (2^n|A-S|)^{\frac{1}{2}-\frac{1}{p'}}||F(S,\cdot)||_{L_b^{p'}(\mathbb{R}^n)} dS \right|\right|_{L_A^q(\text{Sym}(\mathbb{R}^n))} \tag{3.5.22}$$

Here difficulties occur, since the kernel $|A|^{-\alpha}$ is singular not only at zero but on the entire determinental variety $Z(\det) = \{A \in \text{Sym}(\mathbb{R}^n) \mid \det A = 0\}$, and as such Hardy-Littlewood-Sobolev cannot be directly applied. If we had a "determinental coordinate system" in which $t = \det A$ was a coordinate, then $|t|^{-\alpha}$ could be factored out of the inner most integral and perhaps Hardy-Littlewood-Sobolev could be used. In particular we will seek $\phi : Z(\det) \times \mathbb{R} \to \text{Sym}(\mathbb{R}^n)$

satisfying

$$\phi(Z,0) = Z \tag{3.5.23}$$

$$\det[\phi(Z,t)] = t \tag{3.5.24}$$

If we compute the t derivative of (3.5.24) we obtain

$$1 = \langle \nabla \det[\phi(Z,t)], \partial_t \phi(Z,t) \rangle \tag{3.5.25}$$

If we further impose the requirement that the coordinate curves for t be orthogonal to the level sets of the determinant, then we obtain the following first order ODE in $\frac{1}{2}n(n+1)$ variables:

$$\begin{aligned} \frac{d\phi}{dt} &= \frac{1}{||\nabla \det[\phi(Z,t)]||_2^2} \nabla \det[\phi(Z,t)] \\ \phi(Z,0) &= Z \end{aligned} \tag{3.5.26}$$

The trick is that in general the gradient of the determinant, put into matrix form, is the matrix of signed cofactors (this follows from Laplace expansion):

$$\nabla \det[\phi] = \text{cof}(\phi) = \text{adj}(\phi)^T \tag{3.5.27}$$

But ϕ is symmetric, which means its adjugate matrix is symmetric too! Thus $\nabla \det[\phi] = \text{adj}(\phi) = \det[\phi]\phi^{-1}$ and our ODE becomes

$$\frac{d\phi}{dt} = \frac{1}{\det \phi ||\phi^{-1}||_2^2} \phi^{-1} \tag{3.5.28}$$

Next note that $\det \phi = t$ and multiply $\frac{d\phi}{dt}$ on the left and the right by ϕ and add the two together to obtain

$$\frac{d}{dt}\phi^2 = \phi\frac{d\phi}{dt} + \frac{d\phi}{dt}\phi = \frac{2}{t||\phi^{-1}||_2^2}\mathbb{I} \tag{3.5.29}$$

Thus if $\psi = \phi^2$ then ψ obeys the ODE

$$\frac{d\psi}{dt} = \frac{2}{t||\psi^{-1}||_1}\mathbb{I} \tag{3.5.30}$$

Only the diagonal entries of ψ change under this flow, and their time derivative is a multiple of identity. Thus we may assume that $\psi(t) = \psi(0) + \lambda(t)\mathbb{I} = Z^2 + \lambda(t)\mathbb{I}$, in which case we obtain a one dimensional ODE for $\lambda(t)$ when $t > 0$

$$\begin{aligned} &\frac{d\lambda}{dt} = \frac{2}{t||(Z^2 + \lambda(t)\mathbb{I})^{-1}||_1} \\ &\lambda(0) = 0 \end{aligned} \tag{3.5.31}$$

Thus if $\mu_1, \ldots, \mu_n$ are the eigenvalues of $\psi(0) = Z^2$ then $||(\psi(0) + \lambda(t)\mathbb{I})^{-1}||_1 = (\mu_1 + \lambda)^{-1} + \cdots + (\mu_n + \lambda)^{-1}$ thus we may explicitly integrate to obtain the solution

$$|\mu_1 + \lambda| \cdots |\mu_n + \lambda| = At^2 \tag{3.5.32}$$

The initial condition $\lambda(0) = 0$ taken together with the fact that $|\mu_1 \ldots \mu_n| = (\det \phi)^2 = t^2$ sets the integration constant $A = 1$. Thus λ is a positive, non-decreasing function depending only on $\mu_1, \ldots, \mu_n$ defined such that $\det[Z^2 + \lambda(t)\mathbb{I}] = t^2$. If $\phi(Z, t)$ is positive definite then

$\phi(Z,t) = \sqrt{Z^2 + \lambda(t)\mathbb{I}}$, otherwise we have to account for the signature $\epsilon \in \{-1,1\}^n$ of $\phi(Z,t)$ via $\phi(Z,t) = \sigma_\epsilon(\sqrt{Z^2 + \lambda(t)\mathbb{I}})$ where

$$\begin{aligned} &\sigma_\epsilon : \mathbb{P}(n) \to \mathring{S}^{p,q}(\mathbb{C}^n) \\ &\sigma_\epsilon(U\Lambda U^T) = U\,\mathrm{diag}(\epsilon_1,\dots,\epsilon_n)\Lambda U^T \end{aligned} \tag{3.5.33}$$

Thus, using ϕ as a change of variables to evaluate the troublesome integral in the upper bound of 3.5.22 we obtain

$$||\tilde{T}\tilde{T}^* F||_{L^q_A L^p_b(\mathrm{Sym}(\mathbb{R}^n)\times\mathbb{R}^n)} \leqslant \left|\left| \int_{\mathrm{Sym}(\mathbb{R}^n)} (2^n|A|)^{\frac{1}{2}-\frac{1}{p'}} ||F(A+S,\cdot)||_{L^{p'}_b(\mathbb{R}^n)} dS \right|\right|_{L^q_A(\mathrm{Sym}(\mathbb{R}^n))} \tag{3.5.34}$$

$$= \left|\left| \int_{\mathbb{R}} (2^n|t|)^{\frac{1}{2}-\frac{1}{p'}} \sum_{\epsilon\in\{-1,1\}^n} \int_{Z(\det)} ||F(A+\sigma_\epsilon(\sqrt{Z^2+\lambda_Z(t)\mathbb{I}}),\cdot)||_{L^{p'}_b(\mathbb{R}^n)} dZ \right|\right|_{L^q_A(\mathrm{Sym}(\mathbb{R}^n))} \tag{3.5.35}$$

$$= \left|\left| \int_{\mathbb{R}} (2^n|t|)^{\frac{1}{2}-\frac{1}{p'}} \tau||F||_{L^{p'}_b(\mathbb{R}^n)}(A,t) dt \right|\right|_{L^q_A(\mathrm{Sym}(\mathbb{R}^n))} \tag{3.5.36}$$

$$\lesssim \left|\left| \int_{\mathbb{R}} (2^n|t|)^{\frac{1}{2}-\frac{1}{p'}} ||F(A+t\mathbb{I})||_{L^{p'}_b(\mathbb{R}^n)} dt \right|\right|_{L^q_A(\mathrm{Sym}(\mathbb{R}^n))} \tag{3.5.37}$$

Where in the last line the assumption (3.5.18) is used. We now decompose the symmetric matrices into two parts: the identity component $V_I = \{\lambda\mathbb{I} \mid \lambda \in \mathbb{R}\}$ and the traceless component $V_T = \{A \in \mathrm{Sym}(\mathbb{R}^n) \mid \mathrm{tr}\{A\} = 0\}$. We then note that for $g \in L^q_A(\mathrm{Sym}(\mathbb{R}^n))$ we have $||g||_{L^q_A(\mathrm{Sym}(\mathbb{R}^n))} =$

$||g||_{L^q_A(V_T)L^q_A(V_I)}$, thus:

$$\begin{aligned} ||\tilde{T}\tilde{T}^* F||_{L^q_A L^p_b(\text{Sym}(\mathbb{R}^n)\times\mathbb{R}^n)} &\leqslant \left|\left| \int_{\mathbb{R}} (2^n|t|)^{\frac{1}{2}-\frac{1}{p'}} ||F(A+t\mathbb{I})||_{L^{p'}_b(\mathbb{R}^n)} dt \right|\right|_{L^q_A(V_T)L^q_A(V_I)} \\ &\lesssim ||F||_{L^q_A L^{q'}_t L^{p'}_b(V_I\times V_T\times\mathbb{R}^n)} \end{aligned} \tag{3.5.38}$$

Where in the first inequality Hardy-Littlewood-Sobolev is used with $\alpha = \frac{1}{p'} - \frac{1}{2}$, $d = 1$, $s = q$, and $v = q'$ (the assumption $\frac{2}{q} + \frac{1}{p} = \frac{1}{2}$ then implies the requirement Hardy-Littlewood-Sobolev $\frac{1}{v} = \frac{1}{s} + \frac{d-\alpha}{d}$). This concludes the proof of the theorem since in this case

$$||\tilde{T}^* F||^2_{L^2(\mathbb{R}^n)} \lesssim_{p,q,n} ||F||_{L^{q'}_A L^{p'}_b(\text{Sym}(\mathbb{R}^n)\times\mathbb{R}^n)} || \cdot ||F||_{L^q_A L^{q'}_t L^{p'}_b(V_I\times V_T\times\mathbb{R}^n)} \tag{3.5.39}$$

3.6 A Convolution Identity for the CFT

An alternate approach is to enlarge the range of T and $\tilde{T}$ by using a measure different from the Haar measure (the Lebesgue measure) on $\text{Sym}(\mathbb{R}^n) \times \mathbb{R}^n$. Instead, we consider a measure $dW = W(A,b)dAdb$ with $W(A,b) > 0$ so that

$$||W||_{1,\infty} := \sup_{b\in\mathbb{R}^n} \int_{\text{Sym}(\mathbb{R}^n)} W(A,b)dA < \infty \tag{3.6.1}$$

and define our transform as

$$\begin{aligned} &T : L^2(\mathbb{R}^n) \to L^2(\text{Sym}(\mathbb{R}^n)\times\mathbb{R}^n, W(A,b)dAdb) \\ &T(f)(A,b) = \int_{\mathbb{R}^n} e^{-2\pi i(\frac{1}{2}\langle x,Ax\rangle + \langle b,x\rangle)} f(x)dx \end{aligned} \tag{3.6.2}$$

The added factor of $\frac{1}{2}$ in the phase we will be convenient in this context. We will prove the following theorem:

Theorem 3.6.1. *Let T be as in (3.6.2). Then T is a bounded linear operator with closed range, and moreover if $W(A,b) = W(A)$, that is if the weight depends only on the symmetric chirp matrix we have:*

(i) *T^*T is a multiple of identity on $L^2(\mathbb{R}^n)$.*

(ii) *$F \in L^2(Sym(\mathbb{R}^n) \times \mathbb{R}^n, W(A)dAdb)$ is in Ran(T) if and only if $F = K * (WF)$ where*

$$K(A,b) = \frac{1}{\sqrt{i|A|}} e^{\pi i ||A^{-\frac{1}{2}} b||^2} \tag{3.6.3}$$

(iii) *Ran(T) is not closed under multiplication, however if $TfTg \in$ Ran(T) then $TfTg = T(f \star g)$ where*

$$f \star g(z) := \int_{\mathbb{R}^n} \hat{W}(xx^T + \frac{1}{2}(xz^T + zx^T)) f(x) g(z-x) dx \tag{3.6.4}$$

In general for any $f, g \in L^2(\mathbb{R}^n)$ we have

$$TfTg = T(f \star g) + H \tag{3.6.5}$$

where for almost every $x \in \mathbb{R}^n$

$$0 = \int_{Sym(\mathbb{R}^n)} \int_{\mathbb{R}} e^{2\pi i (\frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle)} H(A,b) W(A,b) dbdA \tag{3.6.6}$$

Proof.

Define $W(A) := \sup_{b\in\mathbb{R}^n} W(A,b)$ then

$$
\begin{aligned}
||Tf||^2_{L^2(G,dW)} &= \int_{\text{Sym}(\mathbb{R}^n)} \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} e^{-2\pi i(\frac{1}{2}\langle x,Ax\rangle + \langle b,x\rangle)} f(x)dx \right|^2 W(A,b)dbdA \\
&= \int_{\text{Sym}(\mathbb{R}^n)} \int_{\mathbb{R}^n} \left| \widehat{e^{-2\pi i\frac{1}{2}\langle\cdot,A\cdot\rangle} f(b)} \right|^2 W(A,b)dbdA \\
&\leqslant \int_{\text{Sym}(\mathbb{R}^n)} \int_{\mathbb{R}^n} \left| \widehat{e^{-2\pi i\frac{1}{2}\langle\cdot,A\cdot\rangle} f(b)} \right|^2 dbW(A)dA \\
&= \int_{\mathbb{R}^n} |f(x)e^{-2\pi i\frac{1}{2}\langle x,Ax\rangle}|^2 dx \int_{\text{Sym}(\mathbb{R}^n)} W(A)dA \\
&= ||f||^2_{L^2(\mathbb{R}^n)} ||W||_{1,\infty}
\end{aligned}
\tag{3.6.7}
$$

This proves that T is a bounded operator with $||T||_* \leqslant \sqrt{||W||_{1,\infty}}$. Note that this holds for any operator of the form

$$T_\phi : L^2(\mathbb{R}^n) \to L^2(X \times \mathbb{R}^n, dW) \tag{3.6.8}$$

$$T_\phi f(A,b) = \int_{\mathbb{R}}^n e^{-2\pi i(\phi(A,x)+\langle b,x\rangle)} f(x)dx \tag{3.6.9}$$

Where $(X \times \mathbb{R}^n, dW)$ is a measure space and $\phi : X \to \mathbb{R}$. Finally we observe for convenience that if W depends only on A then $||W||_{1,\infty} = \hat{W}(0)$.

Our strategy will be to seek a left inverse transform in terms of the adjoint T^* which is given

by

$$T^*F(x) = \int_{\text{Sym}(\mathbb{R}^n)} \int_{\mathbb{R}^n} e^{2\pi i(\frac{1}{2}\langle x,Ax\rangle + \langle b,x\rangle)} F(A,b)W(A,b)dbdA \tag{3.6.10}$$

We observe that

$$\begin{aligned} T^*Tf(x) &= \int_{\text{Sym}(\mathbb{R}^n)} \int_{\mathbb{R}^n} e^{2\pi i(\frac{1}{2}\langle x,Ax\rangle + \langle b,x\rangle)} \int_{\mathbb{R}^n} e^{-2\pi i(\frac{1}{2}\langle y,Ay\rangle + \langle b,y\rangle)} f(y)dydbdA \\ &= \int_{\mathbb{R}^n} f(y) \int_{\text{Sym}(\mathbb{R}^n)} \int_{\mathbb{R}^n} e^{-2\pi i(\frac{1}{2}\langle A,yy^T - xx^T\rangle + \langle b,y-x\rangle)} W(A,b)dbdAdy \\ &= \int_{\mathbb{R}^n} \hat{W}(yy^T - xx^T, y-x)f(y)dy \end{aligned} \tag{3.6.11}$$

At first glance this "quadratic convolution" doesn't look terribly promising, however we note that if $W(A,b) = W(A)$ depends only on A then

$$\begin{aligned} \hat{W}(yy^T - xx^T, y-x) &= \int_{\text{Sym}(\mathbb{R}^n)} \int_{\mathbb{R}^n} e^{-2\pi i(\frac{1}{2}\langle A,yy^T - xx^T\rangle + \langle b,y-x\rangle)} W(A)dbdA \\ &= \int_{\text{Sym}(\mathbb{R}^n)} W(A)e^{2\pi i\frac{1}{2}\langle A,yy^T - xx^T\rangle}\delta(y-x)dA \\ &= \hat{W}(0)\delta(x-y) \end{aligned} \tag{3.6.12}$$

From now on we thus consider W depending only on A and normalize $\hat{W}(0) = dW(\text{Sym}(\mathbb{R}^n)) = 1$ so that T^*T is the identity on $L^2(\mathbb{R}^n)$ (this proves (i)). Two related questions about this left-invertible transform are how to characterize $\text{Ran}(T) \subset L^2(G, dW)$ and whether there exists a "convolution like" operation such that T is an algebra homomorphism. We may say with confidence that $\text{Ran}(T)$ is a strict subspace of $L^2(G, dW)$ since if $F \in \text{Ran}(T)$ then there exists

f such that $F = \widehat{e^{-2\pi i \frac{1}{2}\langle \cdot, A\cdot \rangle} f}$, thus one characterization of $\text{Ran}(T)$ is

$$\text{Ran}(T) = \left\{ F \in L^2(G, dW) | \exists f \in L^2(\mathbb{R}^n) \widetilde{F(A, \cdot)}(x) = e^{-2\pi i \frac{1}{2}\langle x, Ax \rangle} f(x) \right\} \tag{3.6.13}$$

This characterization also provides incidentally that $\text{Ran}(T)$ is closed. A second characterization is given by computing $TT^* = \mathbb{P}_{\text{Ran}(T)}$ and noting that $F \in \text{Ran}(T) \iff TT^* F = F$.

$$\begin{aligned}
TT^* F(\tilde{A}, \tilde{b}) &= T(\int_{\text{Sym}(\mathbb{R}^n)} \int_{\mathbb{R}^n} e^{2\pi i (\frac{1}{2}\langle x, Ax \rangle + \langle b, x \rangle)} F(A, b) W(A, b) db dA) \\
&= \int_{\mathbb{R}^n} e^{-2\pi i (\frac{1}{2}\langle y, \tilde{A} y \rangle + \langle \tilde{b}, y \rangle)} \int_{\text{Sym}(\mathbb{R}^n)} \int_{\mathbb{R}^n} e^{2\pi i (\frac{1}{2}\langle x, Ax \rangle + \langle b, x \rangle)} F(A, b) W(A, b) db dA dy \\
&= \int_{\text{Sym}(\mathbb{R}^n)} W(A) \int_{\mathbb{R}^n} F(A, b) \int_{\mathbb{R}^n} e^{-2\pi i (\frac{1}{2}\langle y, (\tilde{A} - A) y \rangle + \langle \tilde{b} - b, y \rangle)} dy db dA \\
&= \int_{\text{Sym}(\mathbb{R}^n)} W(A) \int_{\mathbb{R}^n} F(A, b) \frac{1}{\sqrt{i} |\tilde{A} - A|^{\frac{1}{2}}} e^{\pi i || (\tilde{A} - A)^{-\frac{1}{2}} (\tilde{b} - b) ||^2} db dA \\
&= (K * (WF))(\tilde{A}, \tilde{b})
\end{aligned} \tag{3.6.14}$$

Where

$$K(A, b) = \frac{1}{\sqrt{i|A|}} e^{\pi i || A^{-\frac{1}{2}} b ||^2} \tag{3.6.15}$$

Thus a second characterization of of Ran(T) is

$$\text{Ran}(T) = \{F \in L^2(G, dW) | TT^*F = F\} = \left\{ F \in L^2(G, dW) | K * (WF) = F \right\} \tag{3.6.16}$$

Proving (ii). Thus if $TfTg \in \text{Ran}(T)$ then $TfTg = (TT^*)TfTg = T(T^*(TfTg))$, hence $f \star g := T^*(TfTg)$ satisfies the convolution like identity $T(f \star g) = TfTg$ so long as $TfTg \in$ Ran(T). Moreover, $TfTg \notin$ Ran(T) then $T(f \star g) = (TT^*)(TfTg) = K * (WTfTg)$. As one might expect given the final expression in (3.6.11) the operation $f \star g$ involves a kind of nonlinear convolution with kernel $\hat{W}$. Namely

$$\begin{aligned}
f \star g(z) &= T^*(TfTg)(z) \\
&= \int_{\text{Sym}(\mathbb{R}^n)} \int_{\mathbb{R}^n} e^{2\pi i(\frac{1}{2}\langle z, Az\rangle + \langle b, z\rangle)} \\
&\quad \left(\int_{\mathbb{R}^n} e^{-2\pi i(\frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle)} f(x) dx \right) \left(\int_{\mathbb{R}^n} e^{-2\pi i(\frac{1}{2}\langle y, Ay\rangle + \langle b, y\rangle)} g(y) dy \right) dbW(A)dA \\
&= \int_{\text{Sym}(\mathbb{R}^n)} e^{2\pi i \frac{1}{2}\langle z, Az\rangle} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} e^{-2\pi i(\frac{1}{2}\langle x, Ax\rangle + \langle y, Ay\rangle)} f(x)g(y)\delta(z - x - y) dxdyW(A)dA \\
&= \int_{\text{Sym}(\mathbb{R}^n)} \int_{\mathbb{R}^n} e^{-2\pi i(\langle x, Ax\rangle + \langle z, Ax\rangle)} W(A) f(x)g(z - x) dxdA \\
&= \int_{\mathbb{R}^n} \hat{W}(xx^T + \frac{1}{2}(xz^T + zx^T)) f(x)g(z - x) dx
\end{aligned} \tag{3.6.17}$$

This proves (iii) since $TfTg \in L^2(\text{Sym}(\mathbb{R}^n) \times \mathbb{R}^n, W(A), dAdb)$ and may thus always be decomposed as $TfTg = TT^*(TfTg) + H$ where $T^*H = 0$. This concludes the proof of the theorem.

3.7 Sampling and Reconstruction for Discrete CFT

In this section we will concern ourselves with the numerical invertibility of the discrete Chirp Fourier Transform T_D in the case when $n = 1$. Let $I = \{(a_1, b_1), \ldots, (a_m, b_m)\} \subset \mathbb{Z}_d \times \mathbb{Z}_d$ be of cardinality $|I| = M$. In this case define the sampling operator

$$\begin{aligned} S_I &: \mathbb{C}^{d\times d} \to \mathbb{C}^M \\ S_I[X]_k &= X_{a_k b_k} \end{aligned} \tag{3.7.1}$$

Then $S_I \circ T_D$, the down-sampling of T_D corresponding to the collection I, can be written according to 3.1.9 in terms of the $m \times d$ matrix

$$\mathcal{T}[I]_{k,j} = (S_I T_D e_j)_k = \frac{1}{\sqrt{d}} e^{-2\pi i (a_k j^2 + b_k j)/d} \tag{3.7.2}$$

We will examine the conditioning of this matrix for various collections of M polynomials $(p_k)_{k=1}^M$ where $p_k(j) = a_k j^2 + b_k j \bmod d$ on $\mathbb{Z}_d$. Without taking into account the fact that two different polynomials may take equal values for all $j \in \mathbb{Z}_d$, there are $\binom{d^2}{M}$ choices for the collection I. If $d = 2\tilde{d}$ is even, however, and we consider the polynomials p and $\tilde{p}$ corresponding to (a, b) and $(a + \tilde{d}, b + \tilde{d})$ then for all $j \in \mathbb{Z}_d$

$$\tilde{p}(j) - p(j) = \tilde{d}(j^2 + j) = \tilde{d} j(j + 1) = 0 \bmod d \tag{3.7.3}$$

since $j(j+1)$ is even. Moreover, if $\delta a = \tilde{a} - a$ and $\delta b = \tilde{b} - b$ then

$$\begin{aligned} & \tilde{p}(j) - p(j) = 0 \bmod d \; \forall j \in \mathbb{Z}_d \\ \iff & \delta a j^2 + \delta b j = 0 \bmod d \; \forall j \in \mathbb{Z}_d \\ \iff & j(\delta a j + \delta b) = 0 \bmod d \; \forall j \in \mathbb{Z}_d \end{aligned} \tag{3.7.4}$$

The only way for $j(\delta a + j\delta b)$ to have a common divisor $h > 1$ for all $j \in \mathbb{Z}_d$ is if h divides δa and δb. Thus let $h = \gcd(\delta a, \delta b)$ and x, y be such that $\gcd(x, y) = 1$, $\delta a = xh$, and $\delta b = yh$. In this case $\tilde{p}(j) - p(j) = hj(xj + y)$. Now unless h also divides d, the only choice of x, y such that $hj(xj + y) = 0 \bmod d$ for all $j \in \mathbb{Z}_d$ is $x = 0, y = 0$. Thus assume $d = hc$, then we wish to find x and y such that $j(xj + y) = 0 \bmod c$ for all $j \in \mathbb{Z}_c$. Since this polynomial has at most 2 distinct roots, we must take $c = 2$ in which case $x = y = 1$ is the only non-trivial choice. Thus there are no two quadratic polynomials that take equal values over $\mathbb{Z}_d$ unless $d = 2\tilde{d}$ is even, in which case this occurs when $\delta b = \delta a = \tilde{d}$. Thus, up to equivalence of values there are

$$\tau[d; M] := \text{\# Choices for } I = \begin{cases} \binom{\frac{1}{2}d^2}{M} & d \text{ even} \\ \binom{d^2}{M} & d \text{ odd} \end{cases} \tag{3.7.5}$$

As such, for small values of d we will be able to "brute force" the number theoretic problem of which I give rise to full rank matrices $\mathcal{T}[I]$ (see Table 3.1 and Figure 3.3). Before doing so, however, we note that there is another trivial failure mode for which $\mathcal{T}[I]$ will not be full rank. In particular, if $p_k(j_1) - p_k(j_2) \equiv \text{const} \bmod d$ for all k then the j_1th and j_2th columns of I will be proportional. Unfortunately other types of linear dependencies between the columns of $\mathcal{T}[I]$ do not readily lift to relations between the polynomials $p_1, \ldots, p_M$, so in general it is a

d	$\tau[d;d]$	# w/ proportional columns	# Invertible	# Not Invertible
3	84	9	75	0
4	70	2	36	32
5	53125	25	48005	5100
6	18564	177	5625	12762

Table 3.1: Results from brute force counting the number of invertible $\mathcal{T}[I]$ when $M = d$ and $d = 3, \ldots, 6$. Unfortunately the problem quickly becomes intractable for larger d, indeed for $d = 7$ we have $\tau[d;d] = 85900584$.

difficult problem to obtain necessary and sufficient criteria for a given collection of M distinct polynomials to give rise to $\mathcal{T}[I]$ full rank (simpler criteria than just constructing the matrix $\mathcal{T}[I]$ and verifying that it is full rank). After removing the case of proportional columns we will be forced to manually check if the remaining $\mathcal{T}[I]$ are full rank by computing $\sigma_d(\mathcal{T}[I])$.

We will focus first on the case when $M = d$, in the hopes of finding the sparsest possible sampling from which to reconstruct the signal $z \in \mathbb{C}^d$. Obviously the collection $I_d = \{(a,k)\}_{k=0}^{d-1}$ corresponds to the discrete Fourier transform of $(e^{-2\pi i a j^2/d} z_j)_{j=0}^{d-1}$ and is invertible with condition number $\kappa = 1$, thus in order to make the problem "hard" it is of interest to consider restricting the number of linear frequencies available and to compensate using well chosen chirp frequencies. This setup lends itself to applications in which sampling many linear frequencies is expensive. Specifically, given a family of strict subsets of linear frequencies $(B_d)_{d \geqslant 2}$ with $B_d \subsetneq \{0, \ldots, d-1\}$ we would like to find a family of polynomials $(q_j)_{j \geqslant 1} = (a_j x^2 + b_j x)_{j \geqslant 1}$ with $b_j \in B_d$ whenever $j \leqslant d$ so that for every $d \geqslant 2$ the d polynomials $(q_j|_{\mathbb{Z}_d})_{j=1}^{d}$ give rise to an invertible matrix $\mathcal{T}[I_d] = \mathcal{T}[\{(a_1,b_1), \ldots, (a_d,b_d)\}]$. Consider the simple case in which we disallow the highest linear frequency, that is $B_d = \{0, \ldots, d-2\}$. In this case numerical experiments show that the choice $I_d = \{(0,j) | j = 0, \ldots, d-2\} \cup \{(1,0)\}$ yields $\mathcal{T}[I_d]$ invertible except when d is

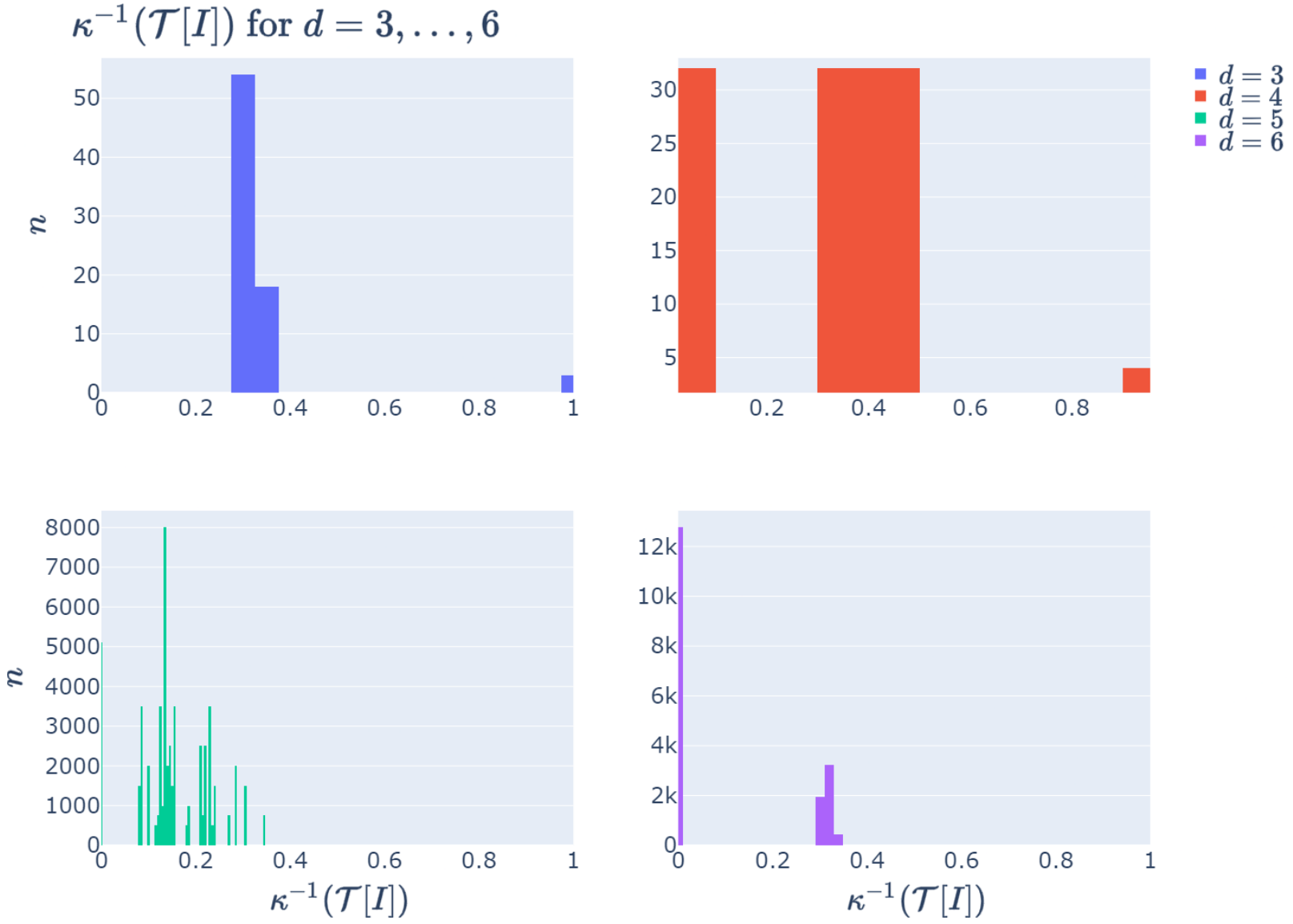


Figure 3.3: Histogram of the inverse condition number of $\mathcal{T}[I]$ when $M = d = 3, \ldots, 6$ over possible choices of I (excluding choices for I that yield proportional columns for $\mathcal{T}[I]$).

an integer multiple of 4. Indeed, when $d = 0 \bmod 4$ we can show that

$$\ker[\mathcal{T}[(0,0),\ldots,(0,d-2),(1,0)]] \ni \begin{bmatrix} e^{2\pi i(d-1)/d} \\ \vdots \\ e^{2\pi i/d} \\ 1 \end{bmatrix} := \omega_d \tag{3.7.6}$$

Here ω_d is precisely the (reversed) vector whose entries are the d dth roots of unity. Indeed if $k < d-1$ and we denote $X = \mathcal{T}[(0,0),\ldots,(0,d-2),(1,0)]$ then

$$\begin{aligned} \sum_{l=0}^{d-1} X_{kl}(\omega_d)_l &= \sum_{l=0}^{d-1} e^{-2\pi ikl/d} e^{2\pi i(d-1-l)/d} \\ &= e^{2\pi i(d-1)/d} \sum_{l=0}^{d-1} e^{-2\pi i(k+1)l/d} \\ &= e^{2\pi i(d-1)/d} \frac{1 - e^{-2\pi i(k+1)}}{1 - e^{-2\pi i(k+1)/d}} \\ &= 0 \end{aligned} \tag{3.7.7}$$

Meanwhile for $k = d-1$ we have

$$\begin{aligned} \sum_{l=0}^{d-1} X_{kl}(\omega_d)_l &= \sum_{l=0}^{d-1} e^{-2\pi il^2/d} e^{2\pi i(d-1-l)/d} \\ &= \sum_{l=0}^{d-1} e^{-2\pi i(l^2+l+1)/d} \end{aligned} \tag{3.7.8}$$

Up to a constant factor of $e^{-2\pi i/d}$ this last sum is the (complex conjugate of) the generalized

quadratic Gauss sum $G(1,1,d)$, defined via

$$G(a,b,c) = \sum_{l=0}^{c-1} e^{2\pi i(al^2+bl)/c} \tag{3.7.9}$$

Such sums are not easy to evaluate in general. The celebrated result of Gauss is that

$$G(s,0,k) = \begin{cases} (1+i^s)\left(\frac{k}{s}\right)\sqrt{k} & k = 0 \mod 4 \\ \left(\frac{s}{k}\right)\sqrt{k} & k = 1 \mod 4 \\ 0 & k = 2 \mod 4 \\ i\left(\frac{s}{k}\right)\sqrt{k} & k = 3 \mod 4 \end{cases} \tag{3.7.10}$$

One can however show using Hensel's lemma and the multiplicative property of quadratic Gauss sums that when $d = 0 \mod 4$ we have $G(1,1,d) = 0$. The multiplicative property says that if $\gcd(c,d) = 1$ then

$$G(a,b,cd) = G(ac,b,d)G(ad,b,c) \tag{3.7.11}$$

See Chapter 6 of [88] for a derivation. In this case if $d = 0 \mod 4$ then we may write $d = 2^k q$ with $k > 1$ and q odd. Thus

$$G(1,1,d) = G(2^k,1,q)G(q,1,2^k) \tag{3.7.12}$$

We will show that $G(q,1,2^k) = 0$. First note that $ql^2 + l \mod 2^k$ takes only even values (if l is even then it is the sum of two even numbers, and if l is odd then it is the sum of two odd

numbers). We will then need the following form of Hensel's lemma (Hensel's lemma has plentiful generalizations, but the following version will suffice for our purposes):

Lemma 3.7.1. *Hensel's lemma [89]. If $p \geqslant 2$ is a natural number, $f(x) \in \mathbb{Z}[x]$, and $a_1, \ldots, a_L \in \mathbb{Z}_p$ are such that $f(a_l) = 0$ mod p and $f'(a_l)$ is coprime to p for $l = 1, \ldots L$, then for any $k > 1$ there exist at least L distinct solutions $b_1, \ldots, b_L \in \mathbb{Z}_{p^k}$ such that $f(b_l) = 0$ mod p^k.*

We apply this lemma to the polynomials $ql^2 + l - 2s$ where $s = 0, \ldots, d/2$. Considered mod 2 we have that

$$ql^2 + l = 0 \mod 2 \tag{3.7.13}$$

has two solutions, namely $l = 0$ and $l = 1$ are both solutions. By the lemma, for each value of s both solutions extend uniquely to solutions of $ql^2 + l - 2s = 0 \mod 2^k$, thus we have determined that each even number $0, \ldots, 2^k - 2$ occurs as a residue of $ql^2 + l$ mod 2^k at least twice. Thus, including multiplicities, we have determined $2(2^k/2) = 2^k$ residues of $ql^2 + l$ mod 2^k. But of course this is all of them! So Hensel's lemma tells us that each even number occurs exactly twice as a residue of $ql^2 + l$ mod 2^k. Thus

$$\begin{aligned} G(q, 1, 2^k) &= 2 \sum_{\substack{s \text{ even} \\ 0 \leqslant s < 2^k}} e^{2\pi i s/2^k} \\ &= 2 \sum_{n=0}^{2^{k-1}-1} e^{4\pi i n/2^k} \\ &= 2 \frac{1 - (e^{4\pi i/2^k})^{2^{k-1}}}{1 - e^{4\pi i/2^k}} \\ &= 0 \end{aligned} \tag{3.7.14}$$

Note that the second to last equality requires $k > 1$, hence why $G(1,1,d) = 0$ when $d = 0 \bmod 4$ but not in general for $d = 0 \bmod 2$. Thus we have shown that the sum in (3.7.8) is zero, and hence that $\mathcal{T}[(0,0),\ldots,(0,d-2),(1,0)]\omega_d = 0$.

Meanwhile, the choice $I_d = \{(0,j)|j = 0,\ldots,d-2\} \cup \{(1,1)\}$ yields $\mathcal{T}[I_d]$ invertible except when $d-2$ is an integer multiple of 4. Indeed, an essentially identical proof to the above yields that

$$\ker[\mathcal{T}[(0,0),\ldots,(0,d-2),(1,1)] \ni \omega_d \tag{3.7.15}$$

When $d = 2 \bmod 4$. Therefore a good strategy to guarantee the invertibility of $\mathcal{T}[I_d]$ is to take $I_d = \{(0,j)|j = 0,\ldots,d-2\} \cup \{(1,\epsilon_d)\}$ with

$$\epsilon_d := \begin{cases} 1 & d = 0 \mod 4 \\ 0 & d \neq 0 \mod 4 \end{cases} \tag{3.7.16}$$

The resulting $\kappa^{-1}(\mathcal{T}[I_d])$ for this strategy are shown in Figure 3.4. As seen in this figure choosing the final chirp frequency pair to be $(1,\epsilon_d)$ always yields the larger of the two inverse condition numbers corresponding to the final chirp frequency pair being $(1,0)$ or $(1,1)$, suggesting that $\kappa^{-1}(\mathcal{T}[I_d]) > 0$. Indeed, we note that the nullity of $\mathcal{T}[I_d]$ is at most 1 since the first $d-1$ columns of $\mathcal{T}[I_d]$ are the first $d-1$ columns of the DFT matrix and are independent. Thus $\mathcal{T}[I_d]$ will fail to be invertible if and only if the last column is in the span of the first $d-1$ columns, that

is:

$$(e^{-2\pi i(j^2+\epsilon_d j)/d})_{j=0}^{d-1} \in \text{span}\{f_0, \ldots, f_{d-2}\} = \text{span}\{f_{d-1}\}^{\perp} \tag{3.7.17}$$

Thus we compute

$$\begin{aligned} 0 &= \langle e^{-2\pi i(j^2+\epsilon_d j)/d})_{j=0}^{d-1}, f_{d-1} \rangle \\ &= \sum_{j=0}^{d-1} e^{2\pi i(j^2+(\epsilon_d-(d-1))j)/d} \\ &= G(1, \epsilon_d + 1, d) \end{aligned} \tag{3.7.18}$$

At this point note that if $\alpha = \beta \bmod 2$ we have the useful identity

$$|G(1, \alpha, d)| = |G(1, \beta, d)| \tag{3.7.19}$$

Indeed, if $\alpha = \beta + 2s$ then completing the square yields:

$$\begin{aligned} |G(1, \alpha, d)| &= |\sum_{j=0}^{d-1} e^{2\pi i(j^2+(\beta+2s)j)/d}| \\ &= |e^{-2\pi i(s^2+\beta s)/d} \sum_{j=0}^{d-1} e^{2\pi i((j+s)^2+\beta(j+s))/d}| \\ &= |\sum_{j=0}^{d-1} e^{2\pi i(j^2+\beta j)/d}| \\ &= |G(1, \beta, d)| \end{aligned} \tag{3.7.20}$$

Thus there are only two cases to concern ourselves with: $|G(1, 0, d)|$ when ϵ_d is odd and $|G(1, 1, d)|$ when ϵ_d is even. In this case (3.7.10) tells us immediately that $|G(1, 0, d)| \neq 0$ when $d \neq 2 \bmod 4$

Figure 3.4: Plotted above is the inverse condition number $\kappa^{-1}(\mathcal{T}[I]) = \sigma_d(\mathcal{T}[I])/\sigma_1(\mathcal{T}[I])$ for $I = \{(0,0,), \ldots, (0, d-2), (a_d, b_d)\}$. As is shown, the choice $(a_d, b_d) = (1, \epsilon_d)$ gives the larger of the two singular values for the choices $(a_d, b_d) = (1, 0)$ and $(a_d, b_d) = (1, 1)$ and as such it is always the case that $\kappa^{-1}(\mathcal{T}[\{(0,0), \ldots, (0, d-2), (1, \epsilon_d)\}]) > 0$.

hence $\mathcal{T}[(0,0), \ldots, (0, d-2), (1, \epsilon_d)]$ is invertible when ϵ_d is odd and when $d \neq 2 \bmod 4$. It remains to show that $|G(1,1,d)| \neq 0$ when $d \neq 0 \bmod 4$. Fortunately, it was shown in [90] that

$$G(1,1,d) = \sqrt{d}\frac{1 - i^{-d}}{1 + i^{-1}} \tag{3.7.21}$$

Thus $\mathcal{T}[(0,0), \ldots, (0, d-2), (1, \epsilon_d)]$ is invertible when ϵ_d is even and when $d \neq 0 \bmod 4$.

We would of course like to do better than removing a single linear frequency. The previous

example suggests that it should suffice to consider the inclusion of a single chirp frequency, so for simplicity we will consider sampling schemes of the form $I_d = \{(0, b_1), \ldots, (0, b_t)\} \cup \{(1, b_1), \ldots, (1, b_s)\}$ where $t = \lceil d/2 \rceil$ and $s = \lfloor d/2 \rfloor$ so that $t + s = d$. In the case where d is odd we will thus have one additional chirp zero sample, and we are restricting to a set of linear frequencies $B_d = \{1, \ldots, b_t\}$ of size $|B_d| = \lceil d/2 \rceil$. A useful family of such schemes is given by:

$$S(d, m) = \begin{cases} \{(l, 2mk + j)\}_{\substack{l=0,1 \\ k=0,\ldots,\lfloor d/2m \rfloor - 1 \\ j=0,\ldots,m-1}} & d = 0 \bmod 2m \\ \{(l, 2mk + j)\}_{\substack{l=0,1 \\ k=0,\ldots,\lfloor d/2m \rfloor \\ j=0,\ldots,m-1-\eta(d,m,l,k)}} & d \neq 0 \bmod 2m \end{cases} \tag{3.7.22}$$

Where if $d = 2mq + r$ for $0 \leqslant r < 2m$ and $2m - r = 2u + v$ for v either 0 or 1 then $\eta(d, m, l, k)$ is given by

$$\eta(d, m, l, k) = \begin{cases} 0 & k < \lfloor d/2m \rfloor \\ u & k = \lfloor d/2m \rfloor, l = 0 \\ u + v & k = \lfloor d/2m \rfloor, l = 1 \end{cases} \tag{3.7.23}$$

This somewhat complicated definition arises from wanting to sample exactly d points. In particular, if we allow 2 choices for l (0 or 1), $\lceil d/2m \rceil$ choices for k ($|\{0, \ldots \lfloor d/2m \rfloor - 1\}| = \lceil d/2m \rceil$ when $d = 0 \bmod 2m$ and $|\{0, \ldots \lfloor d/2m \rfloor\}| = \lceil d/2m \rceil$ when $d \neq 0 \bmod 2m$), and m choices for j ($j = 0, \ldots, m - 1$) then the total number of samples is $2m\lceil d/2m \rceil$. If $d = 2mq + r$ then the number of extra samples is 0 when $r = 0$ and $2m\lceil d/2m \rceil - d = 2m(q + 1) - 2mq - r = 2m - r$ when $r > 0$. If $2m - r = 2u + v$ then we remove u samples from $k = \lfloor d/2m \rfloor$ and $l = 0$ and $u + v$ samples from $k = \lfloor d/2m \rfloor$ and $l = 1$, thus removing a total of $2u + v = 2m - r$ samples to

obtain a total of d samples. The sampling schemes $S(d, m)$ are shown in Figure 3.5 for $d = 20$ and $m = 1, 2, 4, 8$.

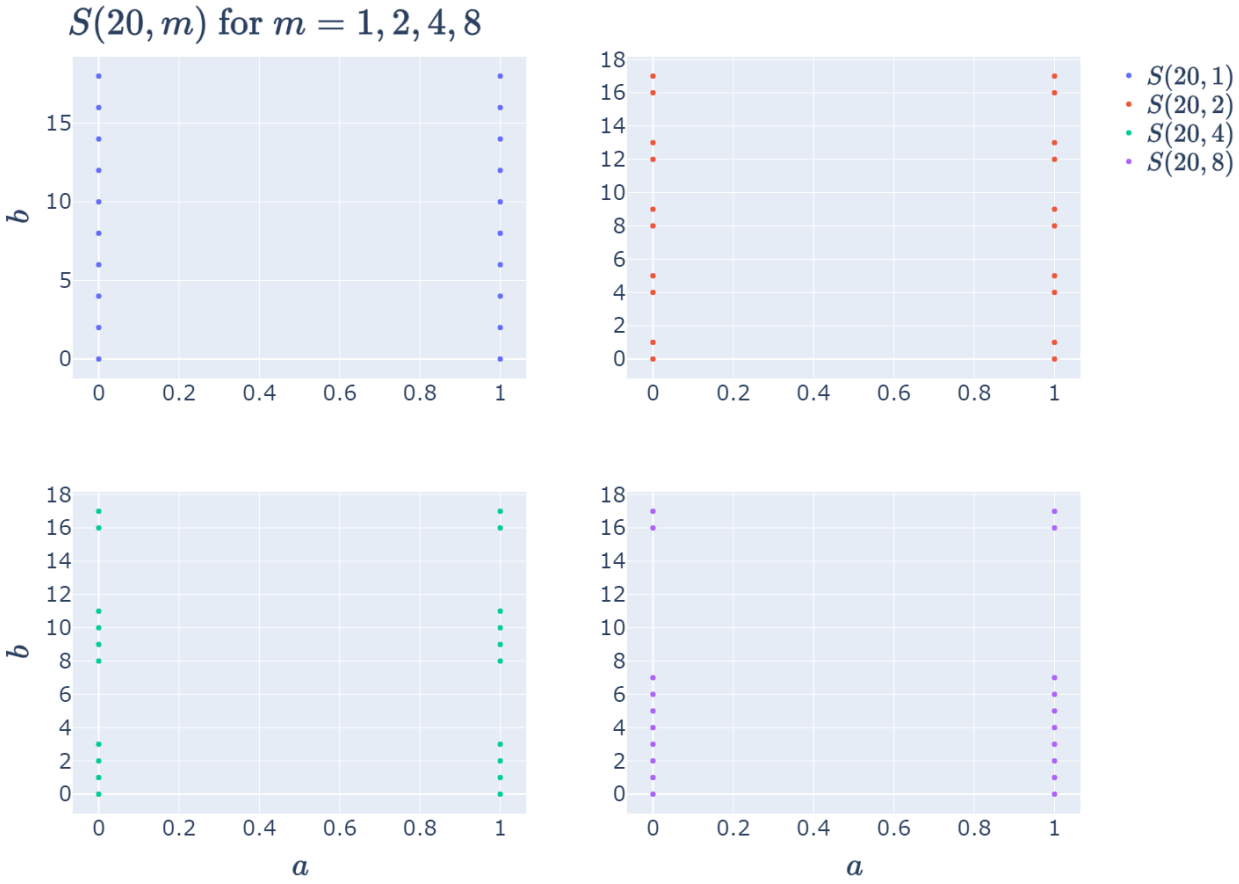


Figure 3.5: The sampling scheme $S(d, m)$ samples $l = 0$ and $l = 1$ equally when d is even, with one additional sample granted to $l = 0$ when d is odd. When m is 1 only even frequencies are sampled, when m is 2 only frequencies that are equal to 0 or 1 modulo 4 are sampled, etc.

The inverse condition numbers resulting from the sampling schemes $S(d, m)$ are shown in Figure 3.6. Interestingly, as m increases the largest value of $\kappa^{-1}(\mathcal{T}[S(d, m)])$ decreases but the period with which $\kappa^{-1}(\mathcal{T}[S(d, m)])$ vanishes increases. In particular when $m = 1$ we find that $\kappa^{-1}(\mathcal{T}[S(d, m)])$ vanishes for $d = 0 \mod 4 = 0$, for $d = 0 \mod 8$ when $m = 2$, for $d = 0 \mod 32$ when $m = 4$, and for $d = 0 \mod 128$ when $m = 8$. A reasonable assumption is therefore that the sampling scheme $S(d, d) = \{(0, 1) \ldots, (0, \lceil d/2 \rceil)\} \cup \{(1, 1) \ldots, (1, \lfloor d/2 \rfloor)\}$

will always have $\kappa^{-1}(\mathcal{T}[S(d,d)]) > 0$. While this appears to be the case (see Figure 3.7 for a log plot of $\kappa^{-1}(\mathcal{T}[S(d,d)])$) the inverse condition number decays exponentially, too quickly for this scheme to be useful for large d (for example $\kappa^{-1}(\mathcal{T}[S(100,100)]) = 1.04743 \cdot 10^{-12}$).

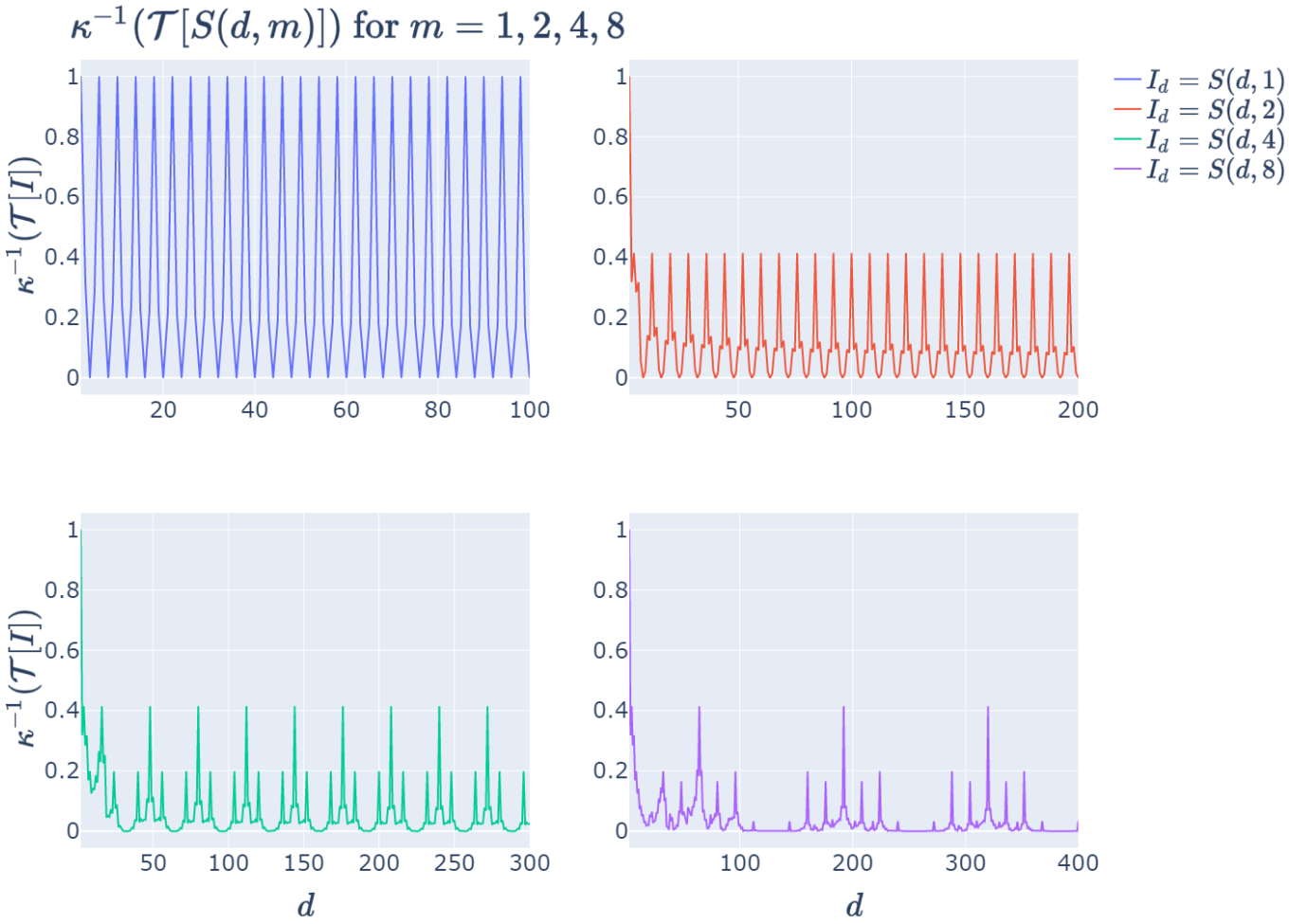


Figure 3.6: Plotted above is $\kappa^{-1}(\mathcal{T}[S(d,m)])$ for $m = 1, 2, 4, 8$.

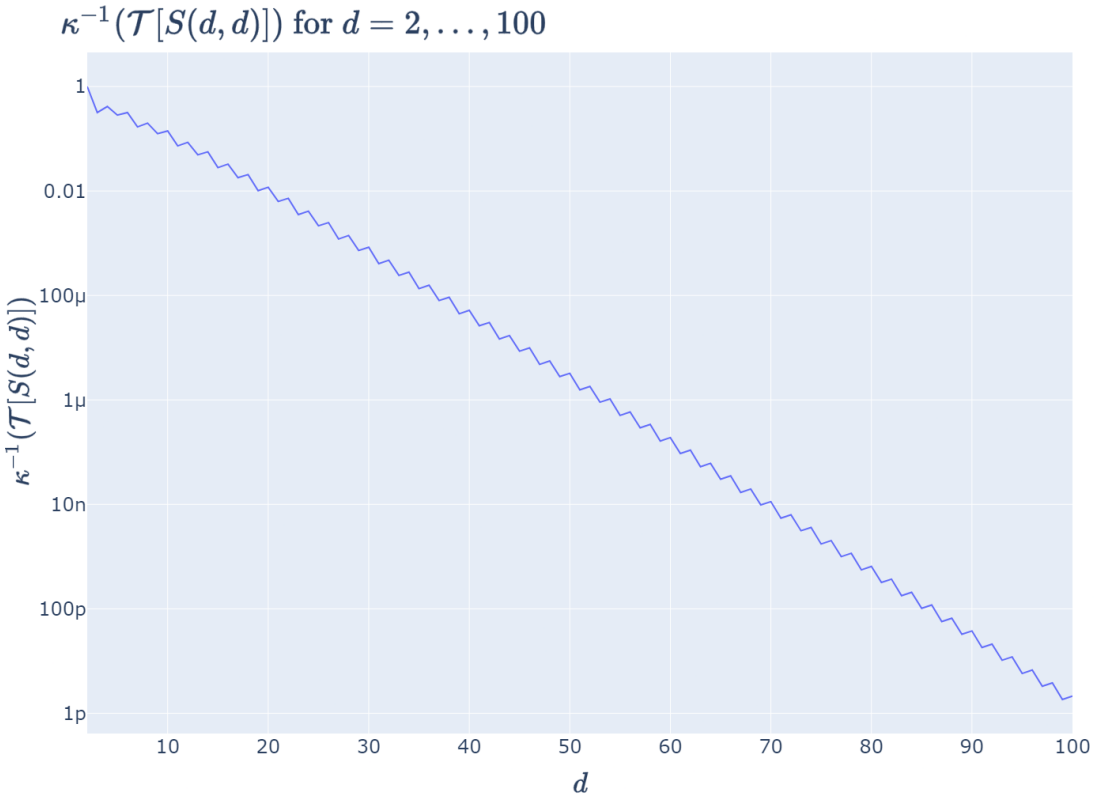


Figure 3.7: Log plot of $\kappa^{-1}(\mathcal{T}[S(d,d)])$ demonstrating that $\mathcal{T}[S(d,d)]$ is always invertible but with an exponentially increasing condition number.

At this point we will relax the requirement that $M = d$ and consider the family of sampling schemes:

$$S(d,m,q) = \begin{cases} \{(l, 2mk+j)\}_{\substack{l=0,\ldots,q-1 \\ k=0,\ldots,\lfloor d/2m \rfloor - 1 \\ j=0,\ldots,m-1}} & d = 0 \bmod 2m \\ \{(l, 2mk+j)\}_{\substack{l=0,\ldots,q-1 \\ k=0,\ldots,\lfloor d/2m \rfloor \\ j=0,\ldots,m-1}} & d \neq 0 \bmod 2m \end{cases} \tag{3.7.24}$$

Note that $S(d,m,2)$ is only equal to $S(d,m)$ when $2m$ divides d, if $d \neq 0 \bmod 2m$ and $d = 2mq + r$ for $1 \leqslant r < 2m$ then $S(d,m,2)$ contains an extra $2m - r$ sample points in addition to those of $S(d,m)$. In general $|S(d,m,q)| = qm\lceil d/2m \rceil$.

As one would expected and as is shown in Figure 3.8, increasing q increases $\kappa^{-1}(\mathcal{T}[S(d,d,q)])$.

Moreover, Figure 3.8 shows that increasing the number of chirp frequencies both delays the decay of and reduces the decay rate of $\kappa^{-1}(\mathcal{T}[S(d,d,q)])$ as d increases. Indeed, as we can see from Figure 3.9 $\kappa^{-1}(\mathcal{T}[S(d,d,d)])$ does not decay at all as d increases, but instead oscillates between 1 when $d = 0 \mod 2$ and a value that is approximately 0.471 when $d = 1 \mod 2$. Note that $|S(d,d,d)| = O(d^2)$, thus it remains an interesting open problem to obtain a sampling scheme that grows as $O(d)$ and whose associated inverse condition number does not decay. Nevertheless, it is thus possible to sample fewer linear frequencies and compensate by sampling a greater number of chirp frequencies and obtain a stably invertible sub-sampling of the discrete chirp Fourier transform for any value of d.

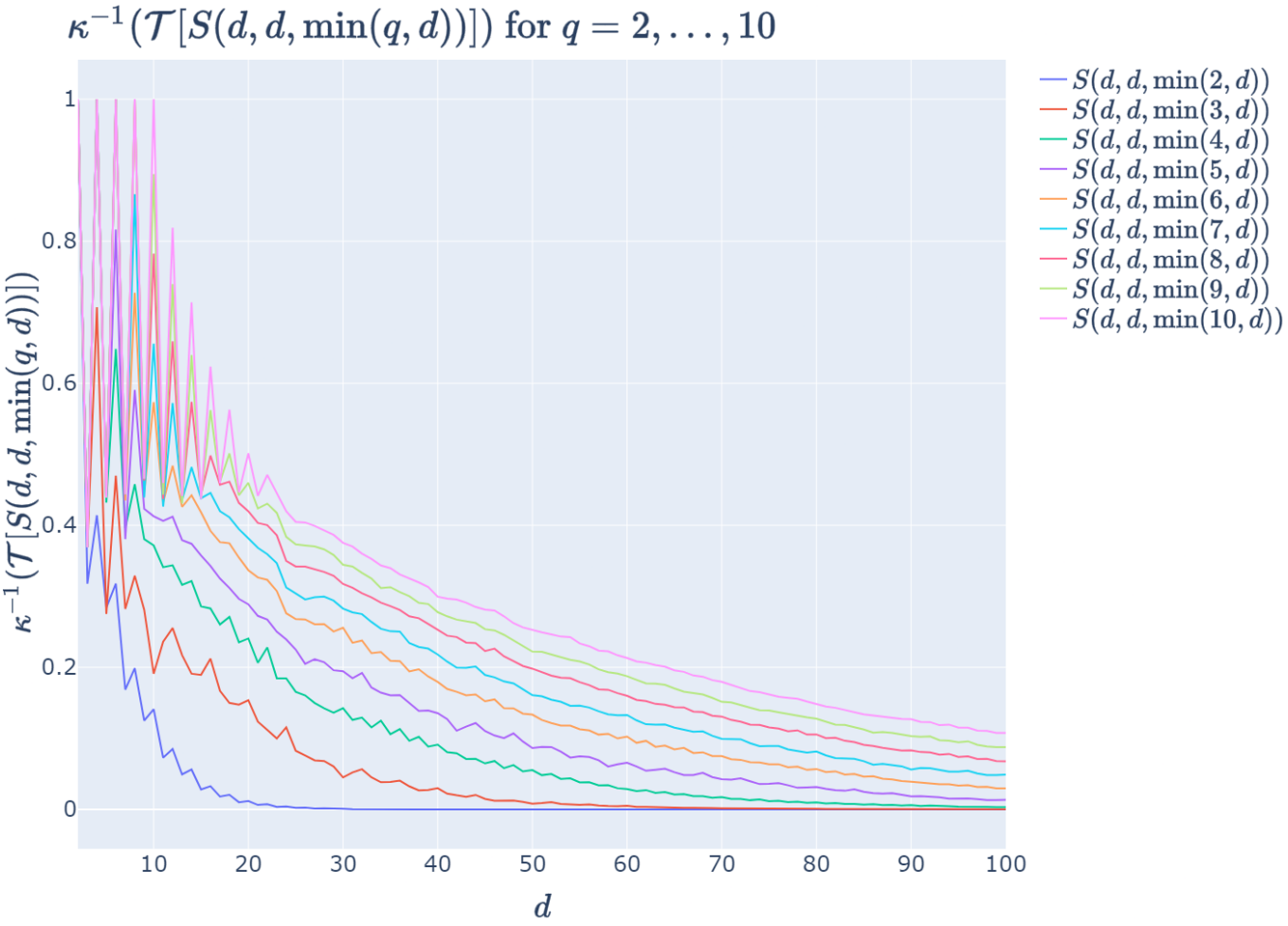


Figure 3.8: Plotted here is the inverse condition number $\kappa^{-1}(\mathcal{T}[S(d,d,\min(q,d))]$ for $q = 2,\ldots,10$ and $d = 2,\ldots,100$. Note that q cannot exceed d since only d chirp frequencies are available.

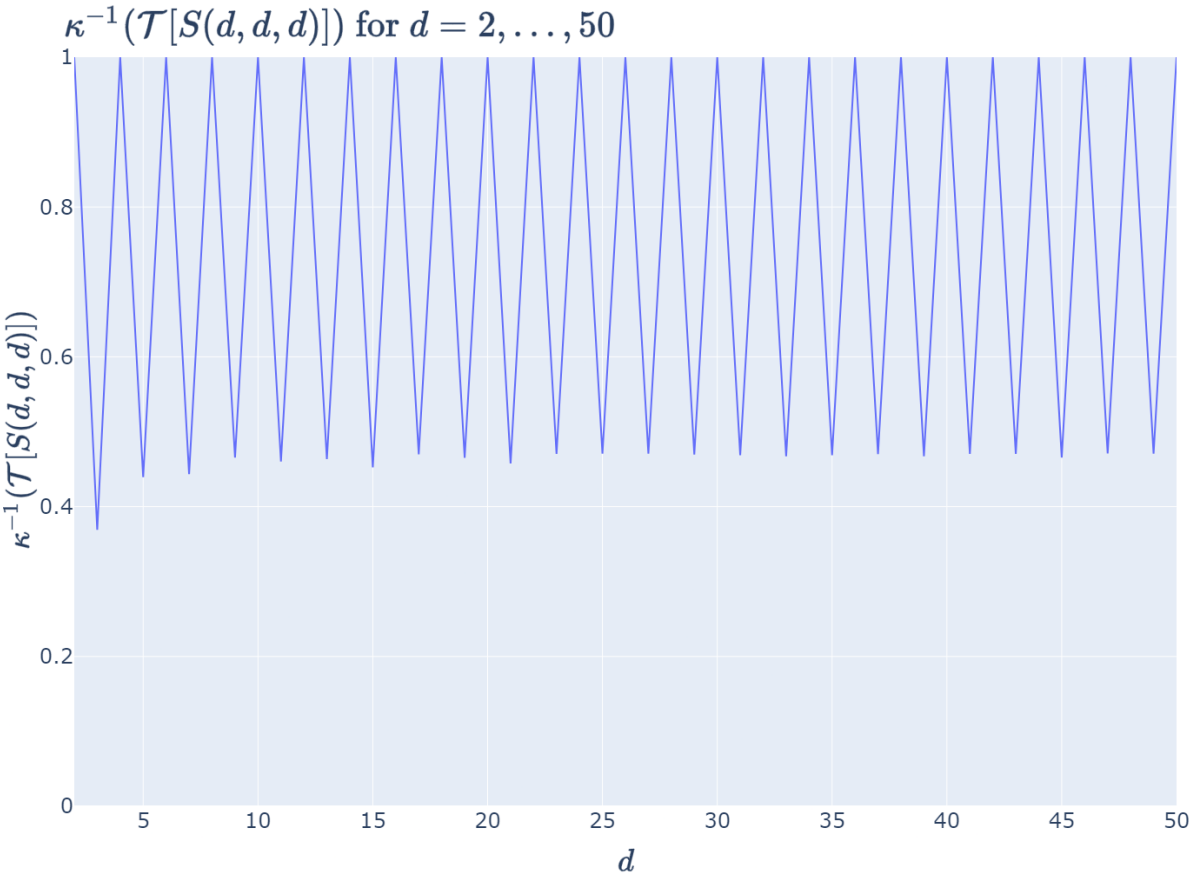


Figure 3.9: Plotted here is the inverse condition number $\kappa^{-1}(\mathcal{T}[S(d,d,d)])$. Evidently $\kappa^{-1}(\mathcal{T}[S(d,d,d)]) = 1$ when d is even. The value of $\kappa^{-1}(\mathcal{T}[S(d,d,d)])$ for d odd appears to approach a limit close to 0.471.

Bibliography

[1] Brendan Leigh Ross and Jesse C Cresswell. Tractable density estimation on learned manifolds with conformal embedding flows. In *Advances in Neural Information Processing Systems*, 2021.

[2] Max Born and Emil Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.

[3] Robert-Jan M Van Geuns, Piotr A Wielopolski, Hein G de Bruin, Benno J Rensing, Peter MA van Ooijen, Marc Hulshoff, Matthijs Oudkerk, and Pim J de Feyter. Basic principles of magnetic resonance imaging. *Progress in cardiovascular diseases*, 42(2):149–156, 1999.

[4] James G Pipe. Spatial encoding and reconstruction in mri with quadratic phase profiles. *Magnetic resonance in medicine*, 33(1):24–33, 1995.

[5] Yoav Shechtman, Yonina C Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 32(3):87–109, 2015.

[6] John R Deller Jr. Discrete-time processing of speech signals. In *Discrete-time processing of speech signals*, pages 908–908. Wiley Online Library, 1993.

[7] Michael V Klibanov, Paul E Sacks, and Alexander V Tikhonravov. The phase retrieval problem. *Inverse problems*, 11(1):1, 1995.

[8] Jameson Cahill, Peter Casazza, and Ingrid Daubechies. Phase retrieval in infinite-dimensional hilbert spaces. *Transactions of the American Mathematical Society, Series B*, 3(3):63–76, 2016.

[9] Zhiqiang Xu. The minimal measurement number for low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 44(2):497–508, 2018.

[10] Radu Balan, Pete Casazza, and Dan Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006.

[11] Afonso S Bandeira, Jameson Cahill, Dustin G Mixon, and Aaron A Nelson. Saving phase: Injectivity and stability for phase retrieval. *Applied and Computational Harmonic Analysis*, 37(1):106–125, 2014.

[12] Yang Wang and Zhiqiang Xu. Generalized phase retrieval: measurement number, matrix recovery and beyond. *Applied and Computational Harmonic Analysis*, 47(2):423–446, 2019.

[13] Radu Balan. Stability of frames which give phase retrieval. *Houston Journal of Mathematics*, 2015.

[14] Yonina C Eldar and Shahar Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[15] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[16] Xuemei Chen, Douglas P Hardin, and Edward B Saff. On the search for tight frames of low coherence. *Journal of Fourier Analysis and Applications*, 27(1):1–27, 2021.

[17] Palina Salanevich. Stability of phase retrieval problem. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–4. IEEE, 2019.

[18] Felix Krahmer and Yi-Kai Liu. Phase retrieval without small-ball probability assumptions. *IEEE Transactions on Information Theory*, 64(1):485–500, 2017.

[19] Zhitao Zhuang. On stability of generalized phase retrieval and generalized affine phase retrieval. *Journal of Inequalities and Applications*, 2019(1):1–13, 2019.

[20] Ji Li and Tie Zhou. On gradient descent algorithm for generalized phase retrieval problem. *arXiv preprint arXiv:1607.01121*, 2016.

[21] Ji Li, Tie Zhou, and Chao Wang. On global convergence of gradient descent algorithms for generalized phase retrieval problem. *Journal of Computational and Applied Mathematics*, 329:202–222, 2018.

[22] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

[23] Radu Balan and Dongmian Zou. On lipschitz analysis and lipschitz synthesis for the phase retrieval problem. *Linear Algebra and its Applications*, 496:152–181, 2016.

[24] Radu Balan and Yang Wang. Invertibility and robustness of phaseless reconstruction. *Applied and Computational Harmonic Analysis*, 38(3):469–488, 2015.

[25] B Bogert, J Healy, and J Tukey. The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocorrelation, cross-cepstrum and saphe cracking. In *Proceedings of symposium on time series analysis*, 1963.

[26] Alan Oppenheim and Ronald Schafer. Homomorphic analysis of speech. *IEEE Transactions on Audio and Electroacoustics*, 16(2):221–226, 1968.

[27] Steven T Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012.

[28] Yang Wang and Zhiqiang Xu. Generalized phase retrieval: Measurement number, matrix recovery and beyond. *Applied and Computational Harmonic Analysis*, 47(2):423–446, 2019.

[29] EE Esser and FJH Herrmann. Application of a convex phase retrieval method to blind seismic deconvolution. In *76th EAGE Conference and Exhibition 2014*, pages 1–5. European Association of Geoscientists & Engineers, 2014.

[30] E. Candés, Y. Eldar, T Strohmer, and V Voroninski. Phase retrieval via matrix completion problem. *SIAM J. Imag. Sci.*, 6(1):199–225, 2013.

[31] Mohammad Ali Hasankhani Fard and Saeedeh Moazeni. Signal reconstruction without phase by norm retrievable frames. *Linear and Multilinear Algebra*, 69(8):1484–1499, 2021.

[32] Vadim Kaloshin. A geometric proof of the existence of whitney stratifications. *arXiv preprint math/0010144*, 2000.

[33] Rajendra Bhatia and Fuad Kittaneh. Notes on matrix arithmetic–geometric mean inequalities. *Linear Algebra and Its Applications*, 308(1-3):203–211, 2000.

[34] Bart Vandereycken, P-A Absil, and Stefan Vandewalle. Embedded geometry of the set of symmetric positive semidefinite matrices of fixed rank. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 389–392. IEEE, 2009.

[35] Christopher G. Gibson. *Singular points of smooth mappings*, volume 105. Pitman London, 1979.

[36] Radu Balan. Reconstruction of signals from magnitudes of redundant representations: The complex case. *Foundations of Computational Mathematics*, 16(3):677–721, 2016.

[37] Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian geometry*, volume 2. Springer, 1990.

[38] John Mather. Notes on topological stability. *Bulletin of the American Mathematical Society*, 49(4):475–506, 2012.

[39] Hassler Whitney. Local properties of analytic varieties. In *Hassler Whitney Collected Papers*, pages 497–536. Springer, 1992.

[40] Radu Balan. Frames and phaseless reconstruction. *Finite Frame Theory: A Complete Introduction to Overcompleteness*, 93:175, 2016.

[41] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

[42] Guanglei Yang, Haifeng Xia, Mingli Ding, and Zhengming Ding. Bi-directional generation for unsupervised domain adaptation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 6615–6622. AAAI Press, 2020.

[43] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *European Conference on Computer Vision (ECCV)*, 2020.

[44] Jay Whang, Erik Lindgren, and Alex Dimakis. Composing normalizing flows for inverse problems. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11158–11169. Proceedings of Machine Learning Research, 18–24 Jul 2021.

[45] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. Density modeling of images using a generalized normalization transformation. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR*, 2016.

[46] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[47] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR*, 2014.

[48] Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[49] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. Proceedings of Machine Learning Research, 2015.

[50] Edmond Cunningham and Madalina Fiterau. A change of variables method for rectangular matrix-vector products. In *International Conference on Artificial Intelligence and Statistics*, pages 2755–2763. Proceedings of Machine Learning Research, 2021.

[51] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR ,*, 2017.

[52] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron C. Courville. Neural autoregressive flows. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2083–2092. PMLR, 2018.

[53] Priyank Jaini, Kira A. Selby, and Yaoliang Yu. Sum-of-squares polynomial flow. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 3009–3018. PMLR, 2019.

[54] Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 573–582. PMLR, 2019.

[55] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pages 6572–6583, 2018.

[56] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[57] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR , Workshop Track Proceedings*, 2015.

[58] Diederik P. Kingma, Tim Salimans, Rafal Józefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational autoencoders with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4736–4744, 2016.

[59] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10236–10245, 2018.

[60] Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 2722–2730. PMLR, 2019.

[61] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, pages 7509–7520, 2019.

[62] Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019.

[63] Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger B. Grosse, and Jörn-Henrik Jacobsen. Understanding and mitigating exploding inverses in invertible neural networks. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pages 1792–1800. PMLR, 2021.

[64] Johann Brehmer and Kyle Cranmer. Flows for simultaneous manifold learning and density estimation. In *Advances in Neural Information Processing Systems*, 2020.

[65] Edmond Cunningham, Renos Zabounidis, Abhinav Agrawal, Ina Fiterau, and Daniel Sheldon. Normalizing flows across dimensions, 2020.

[66] Konik Kothari, AmirEhsan Khorashadizadeh, Maarten V. de Hoop, and Ivan Dokmanic. Trumpets: Injective flows for inference and inverse problems. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI*, volume 161 of *Proceedings of Machine Learning Research*, pages 1269–1278. AUAI Press, 2021.

[67] Abhishek Kumar, Ben Poole, and Kevin Murphy. Regularized autoencoders via relaxed injective probability flow. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 4292–4301. PMLR, 2020.

[68] Anthony L. Caterini, Gabriel Loaiza-Ganem, Geoff Pleiss, and John Patrick Cunningham. Rectangular flows for manifold learning. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[69] George Papamakarios, Eric T. Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22:57:1–57:64, 2021.

[70] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. In *Advances in Neural Information Processing Systems*, pages 3134–3144, 2019.

[71] Robert Cornish, Anthony L. Caterini, George Deligiannidis, and Arnaud Doucet. Relaxing bijectivity constraints with continuously indexed normalising flows. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 2133–2143. PMLR, 2020.

[72] Laurent Dinh, Jascha Sohl-Dickstein, Razvan Pascanu, and Hugo Larochelle. A RAD approach to deep mixture models. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*, 2019.

[73] John M Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.

[74] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[75] You Lu and Bert Huang. Structured output learning with conditional generative flows. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 5005–5012. AAAI Press, 2020.

[76] Xiangxiang Zhu, Bei Li, Zhuosheng Zhang, Wenting Li, and Jinghuai Gao. High-resolution chirplet transform: from parameters analysis to parameters combination. *arXiv preprint arXiv:2108.00572*, 2021.

[77] Steve Mann and Simon Haykin. The chirplet transform: A generalization of gabor's logon transform. In *Vision interface*, volume 91, pages 205–212. Citeseer, 1991.

[78] Steve Mann and Simon Haykin. The chirplet transform: Physical considerations. *IEEE Transactions on Signal Processing*, 43(11):2745–2761, 1995.

[79] Xiang-Gen Xia. Discrete chirp-fourier transform and its application to chirp rate estimation. *IEEE Transactions on Signal processing*, 48(11):3122–3133, 2000.

[80] Brian M Tress and Michael Brant-Zawadski. Nuclear magnetic resonance imaging: basic principles. *Medical Journal of Australia*, 142(1):21–24, 1985.

[81] Vadim Kuperman. *Magnetic resonance imaging: physical principles and applications*. Elsevier, 2000.

[82] Felix Bloch. Nuclear induction. *Physical review*, 70(7-8):460, 1946.

[83] Dietmar Kunz. Use of frequency-modulated radiofrequency pulses in mr imaging experiments. *Magnetic resonance in medicine*, 3(3):377–384, 1986.

[84] John J Healy, M Alper Kutay, Haldun M Ozaktas, and John T Sheridan. *Linear canonical transforms: Theory and applications*, volume 198. Springer, 2015.

[85] Kurt Wolf. *Integral transforms in science and engineering*, volume 11. Springer Science & Business Media, 2013.

[86] Leon Cohen. Time-frequency distributions-a review. *Proceedings of the IEEE*, 77(7):941–981, 1989.

[87] Terence Tao. *Nonlinear dispersive equations: local and global analysis*. Number 106 in Regional Conference Series in Mathematics. American Mathematical Soc., 2006.

[88] K Ireland and M Rosen. A classical introduction to modern number theory. *Grad. Texts in Math*, 84, 1982.

[89] Keith Conrad. Hensel's lemma. *Unpublished. Available at* `https://kconrad.math.uconn.edu/blurbs/gradnumthy/hensel.pdf`, 2015.

[90] Kh M Saliba and Vladimir Nikolaevich Chubarikov. A generalization of the gauss sum. *Moscow University Mathematics Bulletin*, 2(64):92–94, 2009.