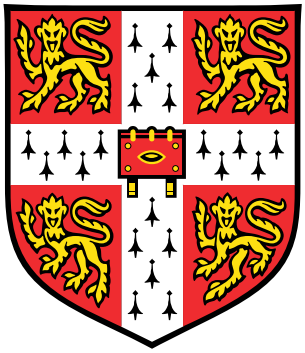
# Isospectral algorithms, Toeplitz matrices and orthogonal polynomials



**Marcus David Webb**

Jesus College
University of Cambridge

A thesis submitted for the degree of
*Doctor of Philosophy*

March 2017

# Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

Chapters 1, 2 and 3 are entirely my own work. The research for chapters 4 and 5 was done in collaboration with Dr Sheehan Olver (University of Sydney). Dr Olver and I wrote all the Julia code implementing the ideas of Chapters 4 and 5, which appears in Appendix A, in collaboration using the *github* system.

For Chapter 4, Dr Olver sugggested to me that certain similarity transformations can be used compute spectral measures of Toeplitz-plus-finite-rank Jacobi operators. Dr Olver and I found empirical evidence that the roots of the Toeplitz symbol of the connection coefficients matrix correspond to eigenvalues of the Jacobi operator together in a meeting in May 2015, but I proved it alone in September 2015.

For Chapter 5, Dr Olver suggested to me that although shifts cannot be effectively applied in the infinite dimensional QR algorithm, they should in principle be possible for an infinite dimensional QL algorithm. Together in May 2015 we worked out a basic implementation of the QL algorithm for Toeplitz-plus-finite-rank Jacobi operators.

Marcus David Webb
March 2017

# Acknowledgements

I should first thank my supervisor, Arieh, for taking me on as a student. His hands-off-but-always-available-for-advice approach to my supervision is one I am grateful to have had. Amongst many things, I enjoyed the stories, the coffee breaks and the trips to China.

I am thankful to Sheehan Olver for his hospitality and mentoring during the months I visited the University of Sydney. I look forward to fruitful future collaborations. Thank you also to others who enriched my time at Sydney: Alex Townsend, Mikael Slevinsky, Geoff Vasil and Sasha Fish. I am grateful for the Cecil King Travel Scholarship from the London Mathematical Society to fund the trip.

I'd like to thank my examiners Anders Hansen and Peter Clarkson for their comments and suggestions, which helped improve the thesis considerably.

Thank you to Reinout Quispel and others at La Trobe University for their hospitality during my visits in 2015 and 2016. I thank Wu Xinyuan, Tang Yifa, Gao Jing, Wang Bin, Liu Kai, Shi Wei and others for their hospitality during my visits to various universities in China in 2015 and 2016.

I am also grateful to many at the Cambridge Centre for Analysis. Thank you to the 2012 cohort for their friendship, and to Filip Rindler and Carola Schönlieb for the mentor roles they played in the first year of my PhD. Thank you to my scientific brother Gil Ramos for his friendship during our adventures abroad at conferences. My studies at the Cambridge Centre for Analysis were supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/H023348/1.

I am most thankful to my wife Yee Chien. What would I be without her?

# Chinese idioms

During my time at Cambridge, besides mathematics I also studied Chinese in my spare time. And so I thought it appropriate to include some Chinese idioms at the start of each chapter, each chosen to reflect an aspect of being a researcher in mathematics. They hence chronicle some of the epiphanies to hit me over these past few years as a simultaneous researcher of mathematics and learner of Chinese.

Idioms exist in all languages, but in Chinese they are known as 成语 (chéng yǔ) and maintain a special place in the culture. They usually consist of four characters; for example, 塞翁失马 (sài wēng shī mǎ), which literally means "the old man at the frontier loses his horse", but figuratively means "a blessing in disguise". This particular idiom comes from a story about an old man whose horse runs away over the northern frontier of China. Days later the horse returns with a fine stallion from across the border. Later on, the old man's son breaks his leg falling off said stallion. The moral of the story is that a seemingly bad situation can turn out to be good, and also vice versa. Such four character sayings are used frequently without explanation causing confusion for the uninitiated.

Chinese idioms often come from ancient stories and literature, but sometimes they don't. Many four character idioms also have eight character extended versions to clarify their meaning. The one for the above example is 塞翁失马，焉知非福, which literally means "the old man at the frontier loses his horse, who knows if it is not fortunate?". I included these long versions where possible, along with Mandarin pronunciation, a literal translation and a figurative translation.

Credit must go to my wife Yee Chien for helping me prepare the idioms and checking them for errors.

# Abstract

An isospectral algorithm is one which manipulates a matrix without changing its spectrum. In this thesis we study three interrelated examples of isospectral algorithms, all pertaining to Toeplitz matrices in some fashion, and one directly involving orthogonal polynomials.

The first set of algorithms we study come from discretising a continuous isospectral flow designed to converge to a symmetric Toeplitz matrix with prescribed eigenvalues. We analyse constrained, isospectral gradient flow approaches and an isospectral flow studied by Chu in 1993.

The second set of algorithms compute the spectral measure of a Jacobi operator, which is the weight function for the associated orthogonal polynomials and can include a singular part. The connection coefficients matrix, which converts between different bases of orthogonal polynomials, is shown to be a useful new tool in the spectral theory of Jacobi operators. When the Jacobi operator is a finite rank perturbation of Toeplitz, here called pert-Toeplitz, the connection coefficients matrix produces an explicit, computable formula for the spectral measure. Generalisation to trace class perturbations is also considered.

The third algorithm is the infinite dimensional QL algorithm. In contrast to the finite dimensional case in which the QL and QR algorithms are equivalent, we find that the QL factorisations do not always exist, but that it is possible, at least in the case of pert-Toeplitz Jacobi operators, to implement shifts to generate rapid convergence of the top left entry to an eigenvalue. A fascinating novelty here is that the infinite dimensional matrices are computed in their entirety and stored in tailor made data structures.

Lastly, the connection coefficients matrix and the orthogonal transformations computed in the QL iterations can be combined to transform these pert-Toeplitz Jacobi operators isospectrally to a canonical form. This allows us to implement a functional calculus for pert-Toeplitz Jacobi operators.

# Contents

千里之行，始于足下

(qiān lǐ zhī xíng, shǐ yú zú xià)

Lit. A journey of a thousand miles begins with a single step
Fig. Big goals are attained by many small achievements

# Chapter 1

# Introduction

The title of this dissertation is *Isospectral algorithms, Toeplitz matrices and orthogonal polynomials*. These three distinct themes come into play in the following fashion: we study interrelated examples of isospectral algorithms, all of which pertain to Toeplitz matrices in some fashion, and one of which directly involves orthogonal polynomials. The research context of the thesis is the field of Numerical Analysis, in particular Numerical Linear Algebra.

An isospectral algorithm is one which manipulates a matrix without changing its spectrum. The most famous example is the QR algorithm, which given a matrix $X_0 \in \mathbb{C}^{n\times n}$ produces a sequence of matrices $X_0, X_1, X_2, \ldots$ such that

$$X_{k+1} = Q_k^H X_k Q_k, \qquad \text{for } k = 0, 1, 2, \ldots,$$

where $Q_k$ is a unitary matrix designed to ensure that (in many cases) $X_0, X_1, X_2, \ldots$ converges to an upper triangular matrix to reveal the eigenvalues of $X_0$ [GVL12], [TBI97], [Fra61], [Wil65], [EH75].

A Toeplitz matrix is one in which the entries along each of the diagonals are constant, which means that for an $n \times n$ matrix $T = T_{ij}$ there exists a vector $(t_{1-n}, \ldots, t_0, \ldots, t_{n-1})^T$ such that $T_{ij} = t_{j-i}$. For example,

$$T = \begin{pmatrix} t_0 & t_1 & t_2 & t_3 & t_4 \\ t_{-1} & t_0 & t_1 & t_2 & t_3 \\ t_{-2} & t_{-1} & t_0 & t_1 & t_2 \\ t_{-3} & t_{-2} & t_{-1} & t_0 & t_1 \\ t_{-4} & t_{-3} & t_{-2} & t_{-1} & t_0 \end{pmatrix}. \tag{1.1}$$

Toeplitz matrices are an example of highly structured matrices. the information which describes a single Toeplitz matrix is linear in the dimension, as opposed to the usual quadratic dependence. Furthermore, their specific structure leads to some elegant and useful properties we will describe in Section 1.1.2.

Let $\mu$ be a probability measure on the real line, and $P_0, P_1, P_2 \dots$ be a sequence of polynomials such that the exact degree of $P_k$ is $k$ for each integer $k$. We say that these polynomials are orthogonal polynomials for $\mu$ if

$$\int_{\mathbb{R}} P_k(s)P_j(s)\,\mathrm{d}\mu(s) \begin{cases} = 0 & \text{if } j \neq k, \\ \neq 0 & \text{if } j = k. \end{cases} \tag{1.2}$$

Orthogonal polynomials are most often thought of in connection with approximation theory and numerical integration, chiefly because they are useful for computing orthogonal projections in the Hilbert space $L^2_\mu(\mathbb{R})$ [Gau04], [SM03], but as will be explained in Section 1.1.3 there are also connections to linear algebra. We use these connections for the solution of some numerical linear algebra problems in Chapter 4.

## 1.1 Background and motivation

In this Section we will discuss some of topics which provide a good foundation, background and motivation for what we cover in the thesis proper.

### 1.1.1 Eigenvalues and spectra

For a matrix $A \in \mathbb{C}^{n\times n}$, its eigenvalues are the solutions $\lambda \in \mathbb{C}$ to the problem

$$Av = \lambda v, \qquad \|v\|_2 = 1. \tag{1.3}$$

More generally, for a closed linear operator $A : X \to X$ where $X$ is a Banach space, its spectrum is the set of all $\lambda \in \mathbb{C}$ such that $A - \lambda I$ does not have a bounded inverse defined everywhere on $X$.

Let us describe some of the basic aspects of eigenvalues which make them so applicable and interesting. When $A$ has a complete linearly independent set of eigenvectors $v_1, v_2, \dots, v_n$ whose eigenvalues are $\lambda_1, \dots, \lambda_n$ respectively, then there exists a diagonalisation of $A$,

$$A = V\Lambda V^{-1}, \tag{1.4}$$

where the $k$th column of $V$ is $v_k$, and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. This diagonalisation elucidates the action of any entire function $f : \mathbb{C} \to \mathbb{C}$ when applied to the matrix $A$ [Hig08]:

$$f(A) = V \begin{pmatrix} f(\lambda_1) & & & \\ & f(\lambda_2) & & \\ & & \ddots & \\ & & & f(\lambda_n) \end{pmatrix} V^{-1}. \tag{1.5}$$

Equation (1.5) shows how the eigenvectors and eigenvalues offer a separation of the variables which are important and the variables which are unimportant, for applying a scalar function $f$ to a matrix. If the function $f$ changes, only the effect of this change on the values $f(\lambda_1), \ldots, f(\lambda_n)$ determines the behaviour of $f(A)$. If for example the function $f$ changed but its values at $\lambda_1, \ldots, \lambda_n$ remained the same, the value of $f(A)$ would be unaffected.

When $A$ does not have a complete linearly independent set of eigenvectors, the corresponding results are not quite as clean. It is still the case that the eigendecomposition determines a simple way in which functions $f$ applied to the matrix behave [Hig08], but it does not provide the full story and it can be more informative to consider results from the beautiful theory of pseudospectra [TE05].

We will describe five aspects and applications of eigenvalues and spectra, but note that we could never hope to do justice to the breadth and depth of this fundamental mathematical topic.

### Vibrations at fundamental frequencies

To see a concrete example of matrix functions in action, consider the harmonic oscillator,

$$\ddot{x}(t) + Ax(t) = 0, \qquad x(t) \in \mathbb{R}^n. \tag{1.6}$$

As long as $A$ is positive definite, this models vibrations and oscillation in classical mechanics [Mar13]. There exist solutions of the form

$$\begin{aligned} x(t) &= \cos(A^{1/2}t)x(0) \\ &= V \begin{pmatrix} \cos(\sqrt{\lambda_1}t) & & \\ & \ddots & \\ & & \cos(\sqrt{\lambda_n}t) \end{pmatrix} V^{-1}x(0), \end{aligned}$$

The solution (in this specific case and for the general solution) are oscillatory with fundamental frequencies given by the square roots of the eigenvalues of $A$: $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}$.

In the infinite dimensional setting in which $x(t) \in L^2(\Omega)$ for a suitably regular domain $\Omega$ in $\mathbb{R}$, $\mathbb{R}^2$ or $\mathbb{R}^3$, and setting $A$ to be the negative Laplacian operator on $L^2(\Omega)$, equation (1.6) becomes the wave equation on $\Omega$. The domain $\Omega$ may represent a vibrating drum or wind instrument (depending on the boundary conditions), and the eigenvalues of the Laplacian on $\Omega$ correspond to the fundamental frequencies of said instrument.

In his book *The Symmetric Eigenvalue Problem*, Parlett said "Vibrations are everywhere, and so too are the eigenvalues associated with them" [Par80]. Equations of the form (1.6) are fundamental and eigenvalues are fundamental too as a consequence.

### Dynamics about equilibria

Now consider the first order linear system of ordinary differential equations,

$$\dot{x}(t) = Ax(t), \qquad x(0) = x^{(0)} \in \mathbb{C}^{n\times n}, \tag{1.7}$$

which can describe (or at least approximate) the behaviour of a continuous dynamical system in equilibrium and then perturbed [Str14]. The solution vector is

$$x(t) = \exp(tA)x^{(0)} = V \begin{pmatrix} \exp(t\lambda_1) & & & \\ & \exp(t\lambda_2) & & \\ & & \ddots & \\ & & & \exp(t\lambda_n) \end{pmatrix} V^{-1}x^{(0)}. \tag{1.8}$$

The asymptotic stability of the system in equation (1.7) is determined by the signs of $\mathrm{Re}(\lambda_k)$ for each $k$. If $\mathrm{Re}(\lambda_k) < 0$ for all $k$ then $x(t) \to 0$ for all initial data $x(0)$ — the effect of the perturbation tends to zero and so the system is said to be (linearly) stable.

If as in Subsection 1.1.1 for equation (1.7) we have initial data $x^{(0)} \in L^2(\Omega)$ for a suitably regular domain $\Omega$ in $\mathbb{R}$, $\mathbb{R}^2$ or $\mathbb{R}^3$, and $A$ is the Laplacian operator on $L^2(\Omega)$, then we obtain the heat equation, or the diffusion equation, on $\Omega$. Fourier's famous solution to the heat equation on a bounded domain via a Fourier series expansion was one of the early examples of eigenvalue analysis, long before the invention of the matrix.

### Dynamics of matrix iterations

Suppose the functions $f$ are the monomials $f_k(z) = z^k$, so that $f_k(A)$ corresponds to $k$ iterations of $A$. Then for any $x \in C^{n\times n}$,

$$f_k(A)x = V \begin{pmatrix} \lambda_1^k & & & \\ & \lambda_2^k & & \\ & & \ddots & \\ & & & \lambda_n^k \end{pmatrix} V^{-1}x. \tag{1.9}$$

If $|\lambda_j| > 1$ for some $j$ then the input $x = v_j$ gives $f_k(A)x = \lambda_j^k x$, which diverges as $k \to \infty$ — *resonance* — and if $|\lambda_j| < 1$ for all $j$ then all inputs will converge to 0 as $k \to \infty$ — *stability*.

The eigendecomposition of $A$ in such situations is applicable for example when $A$ represents the adjacency matrix of a network or graph [BH11], which may represent for example the internet [BP98], or the transition matrix of a Markov chain [Nor98]. Alternatively, consider iterative methods for solving a linear system $Ax = b$. The basic methods such as the Jacobi, Gauss-Seidel and successive over-relaxation methods converge if and only if all the eigenvalues of a certain matrix related to $A$ have absolute value less than 1 [Saa03]. More advanced methods such as Krylov subspace methods, in particular Conjugate Gradients and GMRES, have convergence behaviour which depends intimately on functions of the eigenvalues of $A$. Preconditioning of linear systems utilises the existence a direct relationship between the eigenvalues of $A$ and the rates of convergence of iterative methods [Saa03].

### Quantum theory

In quantum theory, physically observable quantities such as energy, momentum and position may be interpreted as spectra of a Hermitian operator on a Hilbert space call the *Schrödinger operator*, and the quantum states (of matter) which give these observable quantities are given by the associated eigenfunctions [FLSL66].

Besides being pretty fundamental to our understanding of the universe, the spectra of Schrödinger operators have some real world applications. The atomic absorption lines observed in the light emitted from excited atoms and molecules is the basis of the field of spectroscopy. Using absorption spectra, the composition of gases in a laboratory or on distant objects such as stars, planets, or stellar dust clouds can be determined simply by precisely observing the wavelengths of light they emit, because

these wavelengths correspond to (differences in) energy levels of electrons in their constituent atoms. Precise measurements of redshift of distant galaxies are also enabled by finding their absorption spectra (this led to the discovery that the universe is expanding, the first hint of the Big Bang Theory). In computational chemistry, the approximation of the eigenvalues of Schrödinger operators assists in understanding the experimental spectroscopy data [Fra99].

#### Data analysis

In the modern, data-driven world, vast amounts of information can be collected as vectors, in turn forming the columns of a large matrix $X$. One of the most basic forms of data analysis is Principal Component Analysis (PCA), which involves computing the eigenvalues and eigenvectors of the covariance matrix $X^T X$. The eigenvectors with the largest eigenvalues represent salient and mutually, linearly independent aspects of the data set. To use an explicit example, for facial recognition, one forms vectors from digital photographs of faces, and using PCA finds the eigenvectors which in this setting are known as *eigenfaces* [TP91]. The eigenfaces with large eigenvalues represent salient features of a typical face from the data set. A face can then be approximated using a linear combination of a few eigenfaces. Completely ignoring eigenfaces with very small eigenvalues is a form of dimensionality reduction called low rank approximation.

### 1.1.2 Toeplitz matrices and their relatives

The most fascinating aspect of Toeplitz matrices is their connection to functions defined on the complex unit circle. Using the notation $T_{ij} = t_{j-i}$ as in equation (1.1), we can define the function

$$f(z) = \sum_{k=1-n}^{n-1} t_k z^k, \tag{1.10}$$

which is called the *symbol* of the Toeplitz matrix. For a given function $f$ of the form in equation (1.10), can use the notation $T_n(f)$ to denote the $n \times n$ Toeplitz matrix with symbol $f$. The symbol can also be defined in infinite dimensional cases where we have

semi-infinite Toeplitz matrices (Toeplitz operators),

$$T(f) = \begin{pmatrix} t_0 & t_1 & t_2 & t_3 & t_4 & \cdots \\ t_{-1} & t_0 & t_1 & t_2 & t_3 & \ddots \\ t_{-2} & t_{-1} & t_0 & t_1 & t_2 & \ddots \\ t_{-3} & t_{-2} & t_{-1} & t_0 & t_1 & \ddots \\ t_{-4} & t_{-3} & t_{-2} & t_{-1} & t_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{1.11}$$

and doubly-infinite Toeplitz matrices (Laurent operators),

$$L(f) = \begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & t_0 & t_1 & t_2 & t_3 & t_4 & \ddots \\ \ddots & t_{-1} & t_0 & t_1 & t_2 & t_3 & \ddots \\ \ddots & t_{-2} & t_{-1} & \underline{t_0} & t_1 & t_2 & \ddots \\ \ddots & t_{-3} & t_{-2} & t_{-1} & t_0 & t_1 & \ddots \\ \ddots & t_{-4} & t_{-3} & t_{-2} & t_{-1} & t_0 & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{1.12}$$

but we must be careful regarding the regularity of the sequence $(t_k)_{k\in\mathbb{Z}}$ and the symbol $f(z)$, and which spaces the operators act upon. Assume for the sake of argument that $(t_k)_{k\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$. Such coefficients are precisely those such that $T(f)$ is a bounded operator on $\ell^2(\mathbb{N}_0)$ and the symbol $f$ extends almost everywhere to an $L^2(\mathbb{T})$ function, where $\mathbb{T}$ is the complex unit circle ($\mathbb{T} = \partial\mathbb{D}$) [BS13]. The Hilbert space $L^2(\mathbb{T})$ is endowed with the inner product $\langle f, g\rangle = \frac{1}{2\pi}\int_{\mathbb{T}} \overline{f(z)}g(z)\,\mathrm{d}S$, where $\mathrm{d}S$ is the Lebesgue measure on $\mathbb{T}$.

The notion of a Toeplitz symbol is surprisingly fruitful from a mathematical and computational perspective, as we hope to convey in Subsections 1.1.2 to 1.1.2. What is more, the elegant analysis turns out to not only be beautiful, but useful because Toeplitz matrices occur in a multitude of applications [Gra06].

A prime application of Toeplitz matrices is in signal processing, which is a broad field encompassing areas spanning from audio and speech processing, to economic and financial modelling, to control theory [Smi07], [Hay08], [Pro96]. A (real) *digital signal* is a sequence of real numbers $\mathbf{X} = (X_k)_{k\in\mathbb{Z}}$ (which may come from regular measurements of a continuous analogue signal). Denote the space of all signals by $\mathcal{S}$. A (real) *digital filter* takes one digital signal and converts it to another, which we

write $\mathcal{T}\mathbf{X} = (\mathcal{T}_k(\mathbf{X}))_{k\in\mathbb{Z}} = \mathbf{Y}$. If possible for the purpose, the following assumptions are made about a filter $\mathcal{T}$.

- $\mathcal{T}$ acts linearly.
- $\mathcal{T}$ is *shift-invariant*, or *time independent*. This means that $\mathcal{T}_j((X_k)_{k\in\mathcal{Z}}) = \mathcal{T}_{j-s}((X_{k-s})_{k\in\mathcal{Z}})$ for all shifts $s \in \mathbb{Z}$.

Such filters are called Linear Time Invariant (LTI) filters. It is also often assumed that there exists an integer $m$ such that $\mathcal{T}_k(\mathbf{X})$ depends only on $X_k, X_{k-1}, \ldots, X_{k-m}$ — we say the filter is causal and is of finite impulse response. A filter with these assumptions can be defined by a single vector $(b_0, b_1, \ldots, b_m)^T$ and acts upon a signal in a simple manner,

$$Y_k = b_0X_k + b_1X_{k-1} + b_2X_{k-2} + \cdots + b_mX_{k-m}, \tag{1.13}$$

for every $k \in \mathbb{Z}$. This is expressed diagrammatically in Figure 1.1. The signal flow graph shows (with different notation) how with very simple apparatus involving $m + 1$ signal amplifiers, such a filter can physically constructed in hardware.

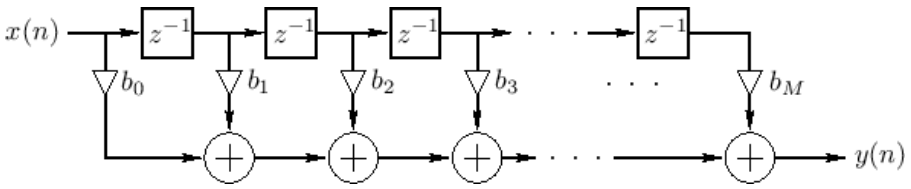



Fig. 1.1 Signal flow graph of a Linear Time Independent filter taken from [Smi07]. The input signal $x(n)$ is sent through the filter and passed through various parts to become signal $y(n)$ at the end. The square labelled $z^{-1}$ delays the signal by 1 discrete time interval, the triangle with the number $b_k$ beside it amplifies the signal by a factor of $b_k$, and the circle labelled $+$ sums the signals from the arrows leading into it to produce the arrow coming out of it. The end effect is that of equation (1.13) (with different notation)

The sequence $b_0, \ldots, b_m$ is called the *impulse response*. This is because if $\mathbf{X}$ were an impulse (i.e. $X_0 = 1$ and $X_k = 0$ for all $k \neq 0$) then the output would be the signal $Y_k = b_k$.

When considering finitely many entries of the output signal simultaneously, equation (1.13) becomes

$$\begin{pmatrix} Y_0 \\ Y_{-1} \\ \vdots \\ Y_{-n} \end{pmatrix} = \begin{pmatrix} b_0 & b_1 & \cdots & b_{m-1} & b_m & 0 & \cdots & 0 \\ 0 & b_0 & \cdots & \ddots & b_{m-1} & b_m & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & b_0 & b_1 & \cdots & b_m \end{pmatrix} \begin{pmatrix} X_0 \\ X_{-1} \\ \vdots \\ X_{-(n+m)} \end{pmatrix}. \tag{1.14}$$

From this we see that the action of an LTI filter is equivalent to multiplying by a Toeplitz matrix.

Sometimes it is beneficial to model a digital signal as a discrete time random process, in particular when the signal is subject to random noise. In this situation symmetric (and Hermitian) Toeplitz matrices occur naturally when constructing a *Wiener optimal filter* for a stationary random process. The difference between a signal and a random process is the assignment of a joint probability density for each finite subset of $\mathbf{X}$. For any such finite subset $\{X_{k_1}, X_{k_2}, \ldots, X_{k_r}\}$, we can define the covariances $\mathbb{E}(\overline{(X_i - \mu_i)}(X_j - \mu_j))$ where $\mathbb{E}$ is expectation with respect to the given probability density on $\{X_{k_1}, X_{k_2}, \ldots, X_{k_r}\}$. The covariance matrix, $C$, whose entries are $c_{i,j} = \mathbb{E}(\overline{(X_{k_i} - \mu_{k_i})}(X_{k_j} - \mu_{k_j}))$ is always Hermitian positive semidefinite, but we also have the following. The process is said to be time invariant or stationary if the joint probability density of $X_{k_1}, X_{k_2}, \ldots X_{k_r}$ is equal to the joint probability density of the shifted sequence $X_{k_1+s}, X_{k_2+s}, \ldots, X_{k_r+s}$ for all integer shifts $s$. For such processes the covariance matrix $C$ is a Toeplitz matrix [LSL99].

### Laurent operators

Probably the most theoretically pleasing type of Toeplitz matrix is the Laurent operator in equation (1.12), whose domain we assume here is $\ell^2(\mathbb{Z})$. The action of $L(f)$ for a symbol $f$ is merely multiplication by $f$ on $L^2(\mathbb{T})$ in disguise. Indeed, letting $\{e_k : k \in \mathbb{Z}\}$ be the standard orthonormal basis for $\ell^2(\mathbb{Z})$, we can define the unitary operator which satisfies

$$\mathcal{F} : L^2(\mathbb{T}) \to \ell^2, \qquad \mathcal{F}[z^k] = e_k \text{ for all } k \in \mathbb{Z}. \tag{1.15}$$

Then it is straightforward to check that for any functions $f, g \in L^2(\mathbb{T})$,

$$\mathcal{F}^* L(f) \mathcal{F}[g](z) = f(z)g(z). \tag{1.16}$$

Therefore all properties of Laurent operators can be derived from the properties of the multiplication operator by its symbol. For example, the spectrum of $L(f)$ is equal to the image $f(\mathbb{T})$, and $\|L(f)\|_2 = \sup_{z\in\mathbb{T}} |f(z)|$.

The unitary transform $\mathcal{F}$ is none other than the Fourier transform (in engineering circles this form of the Fourier transform is sometimes called the z-transform [Pro96]). The reduction to a multiplication operator is an operator-theoretic restatement of the convolution theorem, that $\mathcal{F}[f] * \mathcal{F}[g] = \mathcal{F}(fg)$, where $*$ denotes convolution of $\ell^2(\mathbb{Z})$ sequences.

The relationship between a Laurent operator and its symbol is in fact homomorphic: for functions $f, g \in L^2(\mathbb{T})$, $L(f)L(g) = L(g)L(f) = L(fg)$.

### Circulant matrices

A circulant matrix is a Toeplitz matrix such that $t_k = t_{k \bmod n}$. For example,

$$C_5(f) = \begin{pmatrix} t_0 & t_1 & t_2 & t_3 & t_4 \\ t_4 & t_0 & t_1 & t_2 & t_3 \\ t_3 & t_4 & t_0 & t_1 & t_2 \\ t_2 & t_3 & t_4 & t_0 & t_1 \\ t_1 & t_2 & t_3 & t_4 & t_0 \end{pmatrix}. \tag{1.17}$$

These matrices are to the discrete Fourier transform as Laurent operators are to the semi-discrete Fourier transform discussed above. Considering indices modulo $n$, the symbol of the above Toeplitz matrix is the polynomial

$$f(z) = t_0 + t_1 z + t_2 z^2 + t_3 z^3 + t_4 z^4, \tag{1.18}$$

but also functions such as $f(z) = t_3 z^{-2} + t_4 z^{-1} + t_0 + t_1 z + t_2 z^2$ where the indices have changed while preserving their residue modulo $n$. We can take the same approach as with Laurent operators, but we must instead work on the discrete function space $L^2(\mathbb{T}_n)$, where $\mathbb{T}_n$ is the set of $n$th roots of unity, and $\mathbb{C}^n$ with the standard orthonormal basis $\{e_0, e_1, \ldots, e_{n-1}\}$. Note that the ambiguity of the symbol due to modulo arithmetic disappears which we evaluate them only on $\mathbb{T}_n$. Define the unitary operator which satisfies

$$\mathcal{F}_n : L^2(\mathbb{T}_n) \to \mathbb{C}^n, \qquad \mathcal{F}_n[z^k] = e_k \text{ for } k = 0, 1, 2, \ldots, n-1. \tag{1.19}$$

Then for any $g \in L^2(\mathbb{T}_n)$, $\mathcal{F}_n^* C_n(f)\mathcal{F}[g](z) = f(z)g(z)$. Further, just as for Laurent operators, we have $C_n(f)C_n(g) = C_n(g)C_n(f) = C_n(fg)$ for all $f, g \in L^2(\mathbb{T}_n)$, but we must be careful to take the product of the symbols in $L^2(\mathbb{T}_n)$ where the indices of the coefficients can be considered modulo $n$. Hence $n \times n$ circulant matrices form a commutative algebra of matrices isomorphic to the algebra of polynomials in $L^2(\mathbb{T}_n)$.

The diagonalisation can be written more concretely. Let $F_n \in \mathbb{C}^{n\times n}$ be the Discrete Fourier Transform (DFT) matrix,

$$F_n = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & \cdots & \omega^{(n-1)(n-1)} \end{pmatrix}, \qquad \omega = \exp\left(\frac{2\pi i}{n}\right). \tag{1.20}$$

Then $F_n$ is unitary and simultaneously diagonalises all circulant matrices:

$$C_n(f) = F_n^H \begin{pmatrix} f(1) & & & \\ & f(\omega) & & \\ & & \ddots & \\ & & & f(\omega^{n-1}) \end{pmatrix} F_n. \tag{1.21}$$

Similarly to Laurent operators, the spectrum of $C_n(f)$ is the image $f(\mathbb{T}_n)$.

One of the most important algorithms in numerical analysis is the Fast Fourier Transform (FFT), which allows the computation of $F_n$ applied to a vector in $\mathcal{O}(n \log n)$ operations as opposed to the usual $\mathcal{O}(n^2)$. The theoretical foundations are due to Gauss as early as 1805 (which notably predates Fourier's own work the Fourier transform) [HJB85], but the algorithm was made widespread by Cooley and Tukey [CT65]. The basic idea comes from the multiplicative structure of the roots of unity in $F_n$ allowing a divide-and-conquer approach.

Using the FFT, one can apply a circulant matrix to a vector in $\mathcal{O}(n \log n)$ operators, using the formula $C_n(f)v = F_n^H \mathrm{diag}(F_n(t_0, t_1, \ldots, t_{n-1}))F_n v$. We can in fact use the FFT to apply any $n \times n$ Toeplitz matrix to a vector in $\mathcal{O}(n \log n)$ operations, by embedding the $n \times n$ Toeplitz matrix into a $2n-1 \times 2n-1$ circulant matrix, applying the fast algorithm and then dropping the added $n-1$ entries.

### Toeplitz and Hankel operators

We view the Toeplitz operator given in equation (1.11) as a submatrix of a Laurent operator. The properties of Toeplitz operators can then be derived from those of Laurent operators.

Related to Toeplitz operators are Hankel operators, which given a symbol $f(z) = \sum_{k=-\infty}^{\infty} h_k z^k$ is defined as[1]

$$H(f) = \begin{pmatrix} h_{-1} & h_{-2} & h_{-3} & h_{-4} & \cdots \\ h_{-2} & h_{-3} & h_{-4} & h_{-5} & \ddots \\ h_{-3} & h_{-4} & h_{-5} & h_{-6} & \ddots \\ h_{-4} & h_{-5} & h_{-6} & h_{-7} & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{1.22}$$

Defining the exchange operator $E : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$, $E(a_k)_{k\in\mathbb{Z}} = (a_{-k})_{k\in\mathbb{Z}}$, which can also with a mild abuse of notation act upon $\ell^2(\mathbb{N}_0)$, $\ell^2(\mathbb{Z}_-)$ and $L^2(\mathbb{T})$ in the obvious ways, we can write Laurent operators in the form [BS13, Sec. 2.13],

$$L(f) = \begin{pmatrix} ET(Ef)E & EH(Ef) \\ H(f)E & T(f) \end{pmatrix}. \tag{1.23}$$

The relation $L(f)L(g) = L(fg)$ for $f, g \in L^2(\mathbb{T})$ induces the multiplication relations

$$T(fg) = T(f)T(g) + H(f)H(Eg) \tag{1.24}$$

$$H(fg) = T(f)H(g) + H(f)T(Eg). \tag{1.25}$$

The multiplication relations for Toeplitz and Hankel operators are not as simple as those for Laurent operators, and so results about these operators are not as simple either. For example, the spectrum of a Toeplitz operator $T(f)$ is not $f(\mathbb{T})$ as in the Laurent case, but the union,

$$\sigma(T(f)) = f(\mathbb{T}) \cup \{\lambda \in f(\mathbb{D}) : \operatorname{wind}(f(\mathbb{T}), \lambda) \neq 0\}, \tag{1.26}$$

where $\operatorname{wind}(f(\mathbb{T}), \lambda)$ denotes the winding number of the curve $f(\mathbb{T})$ about the point $\lambda$ [RT92].

[1] Conventions on how the symbol defines a Hankel operator can vary from source to source.

### Finite Toeplitz matrices

The symbol analysis is also applicable to finite dimensional Toeplitz matrices, but we must be careful. For an $n \times n$ Toeplitz matrix, we should consider our symbols to lie in the space $L^2(\mathbb{T}_{2n-1})$ (yes $\mathbb{T}_{2n-1}$ and not $\mathbb{T}_n$). In other words, the symbol $f(z) = \sum_{k=1-n}^{n-1} t_k z^k$ is to be interpreted as equivalent to $f(z) = \sum_{k=0}^{n-1} t_k z^k + \sum_{k=n}^{2n-2} t_{k-2n+1} z^k$. All of these equivalent symbols will give the same $n \times n$ Toeplitz matrix.

The Hankel matrix defined by the symbol $f(z) = \sum_{k=1-n}^{n-1} h_k z^k$ in $L^2(\mathbb{T}_{2n-1})$ is

$$H_n(f) = \begin{pmatrix} h_{-1} & h_{-2} & \cdots & h_{1-n} & h_{n-1} \\ h_{-2} & h_{-3} & \cdots & h_{n-1} & h_{n-2} \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ h_{1-n} & h_{n-1} & \cdots & h_2 & h_1 \\ h_{n-1} & h_{n-2} & \cdots & h_1 & h_0 \end{pmatrix}. \tag{1.27}$$

Note that the indices decrease by 1 modulo $2n-1$ with each diagonal, so this definition is exactly the same as that used for Hankel operators, except that the indices are taken modulo $2n-1$.

The reason for this apparently gratuitous use of modulo arithmetic is that, just as a Toeplitz operator can be viewed as a submatrix of a Laurent operator, so too can an $n \times n$ Toeplitz matrix be thought of as a submatrix of a $2n-1 \times 2n-1$ circulant matrix, whose symbol is an element of $L^2(\mathbb{T}_{2n-1})$.

We must also define the exchange matrix $E \in \mathbb{C}^{n \times n}$,

$$E = E_{n\times n} = \begin{pmatrix} & & & & 1 \\ & & & 1 & \\ & & \iddots & & \\ & 1 & & & \\ 1 & & & & \end{pmatrix}.$$

The rectangular version $E_{m\times n}$ is defined by extending by zero in the direction of greater length.

Now, if we have a symbol of the form $f(z) = \sum_{k=1-n}^{n-1} t_k z^k$, then we may write the $2n-1 \times 2n-1$ circulant with symbol $f$ as

$$C_{2n-1}(f) = \begin{pmatrix} E_{n-1\times n} T_n(Ef) E_{n\times n-1} & E_{n-1\times n} H_n(Ef) \\ H_n(f) E_{n\times n-1} & T_n(f) \end{pmatrix}. \tag{1.28}$$

Compare this to equations (1.24) and (1.25). From the fact that $C_{2n-1}(f)C_{2n-1}(g) = C_{2n-1}(fg)$ for symbols $f, g \in L^2(\mathbb{T}_{2n-1})$, we find the following multiplication relations for Toeplitz and Hankel matrices,

$$T_n(fg) = T_n(f)T_n(g) + H_n(f)P_{n-1}H_n(Eg) \tag{1.29}$$

$$H_n(fg) = T_n(f)H_n(g) + H_n(f)P_{n-1}T_n(Eg), \tag{1.30}$$

where $P_{n-1} \in \mathbb{C}^{n\times n}$ is the projection $P_{n-1}(a_1, a_2, \ldots, a_n)^T = (a_1, a_2, \ldots, a_{n-1}, 0)^T$.

Despite these properties and the many more that follow on from these, there is no known simple characterisation for the eigenvalues of an arbitrary finite dimensional Toeplitz matrix [RT92], [BS13]. There are some well known results about the asymptotics of the eigenvalues as $n \to \infty$ [BS13], but research about the eigenvalues of $n \times n$ Toeplitz matrices continues [MMP99],[Hei01].

### 1.1.3 Orthogonal polynomials

Arguably the most important application of orthogonal polynomials is numerical integration a.k.a. quadrature [Gau04]. Given a probability measure $\mu$ on $\mathbb{R}$, how should one approximate the integral

$$I[f] = \int_{\mathbb{R}} f(s)\,\mathrm{d}\mu(s)? \tag{1.31}$$

Suppose we are restricted to approximations of the form

$$Q_n[f] = \sum_{k=1}^{n} w_k f(\lambda_k), \tag{1.32}$$

where $n$ is fixed and $w_1, \ldots, w_n, \lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Then the Gauss quadrature rule takes $\lambda_1, \ldots, \lambda_n$ to be the roots of $P_n$, the $n$th orthogonal polynomial for $\mu$, and the weights $w_1, \ldots, w_n$ equal to

$$w_k = \int_{\mathbb{R}} \frac{P_n(s)}{(s-\lambda_k)P_n'(\lambda_k)}\,\mathrm{d}\mu(s). \tag{1.33}$$

This choice of quadrature rule is optimal in the sense that it gives the *exact* result when $f$ is a polynomial of degree $2n-1$. Whether this type of optimality is best is up for debate [Tre08], [HO09], but nonetheless Gauss quadrature is known as a "jewel of numerical analysis" [Tre08].

By definition, the orthogonal polynomials for a measure $\mu$ on the real line provide a suitable basis for best approximation in the Hilbert space $L^2_\mu(\mathbb{R})$, but some families of orthogonal polynomials have extremely good properties for uniform approximation of functions. For example, the Chebyshev polynomials $T_n(s) = \cos(n \cos^{-1}(s))$, which are orthogonal with respect to the measure $\mathrm{d}\mu(s) = (1-s^2)^{-1/2}|_{s\in(-1,1)}\,\mathrm{d}s$ can be used to produce near-best uniform polynomial approximation for functions $f \in C([-1,1])$ either by projection or interpolation [Tre13]. This is the basis of the Matlab package Chebfun, which uses state-of-the-art algorithms for approximating and manipulating functions using Chebyshev expansions [DHT14]. The excellent approximation properties of Chebyshev polynomials translates into competitive methods for the solution of differential equations (which are used in Chebfun), Chebyshev spectral methods [Tre00].

The link between orthogonal polynomials and numerical linear algebra becomes clearer here with spectral methods. The solution of an ODE using an orthogonal polynomial basis requires the construction of matrices which represent the action of linear operators such as derivatives, integrals and pointwise multiplication by functions on the coefficients of an expansion in the polynomial basis. This is even true for nonlinear differential equations. Recently, spectral methods utilising other orthogonal bases such as ultraspherical polynomials [OT13] and Jacobi polynomials [VBL+16] have been derived and shown to have some superior properties, such as producing highly structured matrices.

Orthogonal polynomials (on the real line) satisfy a *three term recurrence*. That is, for any given probability measure $\mu$ there exists a real sequence $\alpha_0, \alpha_1, \alpha_2, \ldots$ and a sequence of positive real numbers $\beta_0, \beta_1, \beta_2, \ldots$ such that the orthonormal polynomials $P_0, P_1, \ldots$ for $\mu$ satisfy

$$sP_k(s) = \beta_{k-1}P_{k-1}(s) + \alpha_k P_k(s) + \beta_k P_{k+1}(s), \tag{1.34}$$

$$P_{-1}(s) = 0, \quad P_0(s) = 1. \tag{1.35}$$

One reason this is so interesting is that it leads to a practical recursive algorithm for evaluating finite expansions in orthogonal polynomials called Clenshaw's algorithm [Cle55], [Gau04]. To evaluate a function $f = \sum_{k=0}^n a_k P_k$ at the point $s$, compute the iteration

$$y_n = a_n, \quad y_{n+1} = 0 \tag{1.36}$$

$$y_k = a_k + \beta_k^{-1}(s - \alpha_k)y_{k+1} - \beta_{k+1}^{-1}\beta_k y_{k+2} \text{ for } k = n-1, n-2, \ldots, 1, 0 \tag{1.37}$$

$$f(s) = y_0. \tag{1.38}$$

Note that there is no requirement to know anything about the orthogonal polynomials $P_k$ besides their recurrence coefficients.

The three term recurrence for the orthonormal polynomials is also used to define the *Jacobi operator*. A Jacobi operator is given by the infinite-dimensional matrix [Dei00], [Tes00]

$$J = \begin{pmatrix} \alpha_0 & \beta_0 & & \\ \beta_0 & \alpha_1 & \beta_1 & \\ & \beta_1 & \alpha_2 & \ddots \\ & & \ddots & \ddots \end{pmatrix}.$$

In the case that the measure of orthonormality $\mu$ has compact support, $J$ defines a bounded self-adjoint operator $J : \ell^2 \to \ell^2$ with respect to the standard orthonormal basis $\{e_0, e_1, e_2, \ldots\}$. In fact, there is a one-to-one correspondence between probability measures on $\mathbb{R}$ with compact support, their orthonormal polynomials, and bounded Jacobi operators [AK65]. This measure is actually the *spectral measure* from the Spectral Theorem for self-adjoint operators on Hilbert space [Dei00]. That is, there is a unitary operator $U : \ell^2 \to L^2_\mu(\mathbb{R})$ (such that $Ue_k = P_k$) such that

$$UJU^*[f](s) = sf(s), \tag{1.39}$$

for all $f \in L^2(\mu)$.

The support of the measure $\mu$ is actually the spectrum of $J$. Also, if one takes a finite section of $J$,

$$J_n = \begin{pmatrix} \alpha_0 & \beta_0 & & & \\ \beta_0 & \alpha_1 & \beta_1 & & \\ & \beta_1 & \alpha_2 & \ddots & \\ & & \ddots & \ddots & \beta_{n-2} \\ & & & \beta_{n-2} & \alpha_{n-1} \end{pmatrix}, \tag{1.40}$$

then the eigenvalues of $J$ are the roots of $P_n$, which are also the nodes for the Gauss quadrature rules discussed above. The weight $w_k$ in the Gauss quadrature rule is equal to the square of the first entry of the $\lambda_k$-eigenvector.

### 1.1.4 Isospectral flows

Consider the following differential equation on the space of $n \times n$ matrices $\mathbb{C}^{n\times n}$:

$$\dot{X} = [A(X), X], \quad X(0) = X_0 \in \mathbb{C}^{n\times n}, \tag{1.41}$$

where $A : \mathbb{R} \times \mathbb{C}^{n\times n} \to \mathfrak{sl}(n)$[2] is a matrix-valued function that is locally Lipschitz in $X$ and continuous in $t$, and $[\cdot, \cdot]$ is the matrix commutator (also known as the Lie bracket). Then the eigenvalues of $X(t)$ are the same for all $t$, a surprising result considering the simplicity and generality of the expression. Indeed, denoting the eigenvalues of $X$ by $\lambda_1, \ldots, \lambda_n$, we have for any positive integer $j$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{n}\lambda_i^j &= \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{tr}(X^j) \\
&= \mathrm{tr}\left(\sum_{i=1}^{j} X^{i-1}\,[A(X), X]\,X^{j-i}\right) \\
&= \mathrm{tr}\left(\left[A(X), X^j\right]\right) \quad \text{(telescoping argument)} \\
&= 0 \quad \text{(all commutators are trace-free).}
\end{aligned}$$

Flows of the form in equation (1.41) are called *isospectral flows*. They are of particular interest to numerical analysts because of their connection to algorithms where the eigenvalues of a certain matrix remain fixed throughout the computation, such as the QR algorithm. In fact, there exists a choice of $A$ such that the resulting isospectral flow evaluated at integer values of $t$ gives the iterates of the QR algorithm (see Subsection 2.2.1).

In the theory of nonlinear ordinary and partial differential equations, isospectral flows are a particular breed of *integrable system*, which means that certain functionals of the solution called integrals are conserved. We just showed that for isospectral flows, the values of $\mathrm{tr}(X^j)$ for $j = 1, \ldots, n-1$ remain fixed for all $t$ (the cases $j \geq n$ follow from the Cayley-Hamilton Theorem).

Conversely, certain integrable systems which on the surface do not look like isospectral flows can be recast into an equivalent form which is an isospectral flow, called a *Lax formulation*[3]. We illustrate with two examples.

[2] we may subtract multiples of the identity to make $A$ lie in $\mathfrak{sl}(n)$

[3] In fact, any completely integrable Hamiltonian system can be written in Lax form [BV90]

The Toda lattice is a one-dimensional model for a crystal in solid state physics due to Morikazu Toda [Tod67]. The model has a chain of $n$ particles whose positions evolve according to Hamilton's equations of motion with Hamiltonian

$$H(q,p) = \frac{1}{2}\sum_{j=1}^{n} p_j^2 + \sum_{k=1}^{n-1} \exp(q_k - q_{k+1}), \tag{1.42}$$

where $q_k$ is displacement of the $k$th particle from equilibrium and $p_k$ is its corresponding momentum. In 1974 Flaschka showed that under the change of variables,

$$a_j = \frac{1}{2} p_j, \text{ for } j = 1, 2, \ldots, n, \tag{1.43}$$

$$b_k = \frac{1}{2}\exp((q_k - q_{k+1})/2), \text{ for } k = 1, 2, \ldots, n-1, \tag{1.44}$$

the Toda lattice equations are equivalent to the isospectral flow $\dot{Y} = [B, Y]$ with

$$Y(t) = \begin{pmatrix} a_1(t) & b_1(t) & 0 & \cdots & 0 \\ b_1(t) & a_2(t) & b_2(t) & \cdots & 0 \\ 0 & b_2(t) & a_3(t) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & b_{n-1}(t) \\ 0 & 0 & \cdots & b_{n-1}(t) & a_n(t) \end{pmatrix}, \tag{1.45}$$

and $B(Y) = Y_U - Y_L$ (the difference between the upper and lower triangular parts of $Y$) [Fla74]. Conservation of the eigenvalues of $Y$ is equivalent to the conservation laws of the original ODE [Tes01]. For example, conservation of momentum is equivalent to conservation of $\mathrm{tr}(Y)$ and conservation of energy is equivalent to conservation of $\mathrm{tr}(Y^2)$.

The Korteweg–de Vries equation (KdV equation) is a model for waves on shallow water surfaces [KV95]. It is a nonlinear dispersive Partial Differential Equation for $u : [0, \infty) \times \mathbb{R} \to \mathbb{R}$ described by

$$\frac{\partial u}{\partial t} - 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0, \quad u(0, x) = u_0(x), \tag{1.46}$$

where $u_0$ is often taken to lie in the Schwartz space $\mathcal{S}(\mathbb{R})$ of smooth functions whose derivatives of all orders tend to zero faster than any rational function as $|x| \to \infty$. In

1968, Peter Lax showed that for the linear operators

$$L_u(t) = -\frac{\partial^2}{\partial x^2} + M_{u(t,\cdot)}, \tag{1.47}$$

and

$$A_u(t) = -4\frac{\partial^3}{\partial x^3} + 3\left(M_{u(t,\cdot)}\frac{\partial}{\partial x} + \frac{\partial}{\partial x}M_{u(t,\cdot)}\right), \tag{1.48}$$

on $\mathcal{S}(\mathbb{R})$ where $M_v : f \mapsto vf$ is the multiplication operator, the KdV equation is equivalent to the isospectral flow

$$\frac{\partial L_u}{\partial t} = [A_u, L_u]\,. \tag{1.49}$$

Specifically, the reader can verify directly that the Lax pair $(L_u, A_u)$ satisfies

$$\frac{\partial L_u}{\partial t} + [L_u, A_u] = M_{u_t - 6uu_x + u_{xxx}}. \tag{1.50}$$

As with the Toda lattice, the eigenvalues of $L_u$ are integrals of the KdV equation [Lax68].

### 1.1.5 Inverse eigenvalue problems

The problem of finding a matrix with constrained structure and spectrum is an *inverse eigenvalue problem* [Chu98], [CG02],[CG05], [FNO87]. Inverse eigenvalue problems in general is an important and well developed area of numerical analysis, with a prominent book on the topic by Chu and Golub [CG05]. In physical applications the problem usually corresponds to reconstructing the parameters of a system from knowledge of its dynamical behaviour, in particular its natural frequencies or normal modes [Chu98].

To most mathematicians, the concept of an inverse eigenvalue problem is unfamiliar. However, many problems which are commonly used in applications, but not at first seen as inverse eigenvalue problems *are*, with the right perspective. For example, low rank matrix approximation problems are partially prescribed inverse eigenvalue problems. Additionally, preconditioning, another commonplace technique in numerical analysis, is a kind of inverse eigenvalue problem, in which the eigenvalues of a system are to be made more clustered in order to accelerate the convergence of iterative solvers (see Subsection 1.1.1).

Inverse eigenvalue problems are not just for the finite dimensional. One of the most famous inverse eigenvalue problems was posed by Kac in 1966 [Kac66]: Can one hear

the shape of a drum? More specifically, given a sequence of real numbers $\lambda_1, \lambda_2, \ldots$, can one find a unique Lipschitz domain $\Omega \subseteq \mathbb{R}^2$ such that the eigenvalues of the Laplacian on $\Omega$ (with Dirichlet boundary conditions) are $\lambda_1, \lambda_2, \ldots$. Spoiler alert — this was answered in the negative 26 years later with a demonstration of two domains whose Laplacians have equal spectra [GWW92].

In this thesis we will only concern ourselves with the fully prescribed inverse eigenvalue problem for finite dimensional matrices. Such problems are posed with a vector $\boldsymbol{\lambda} \in \mathbb{C}^n$ and a set $\mathcal{S} \subset \mathbb{C}^{n\times n}$, with the computational task,

$$\text{Find } X \in \mathcal{S} \text{ such that } \sigma(X) = \boldsymbol{\lambda}. \tag{1.51}$$

The following is a list of example structured inverse eigenvalue problems (see [CG02] and [CG05]):

- Jacobi matrices — tridiagonal with positive off-diagonal entries.
- Nonnegative matrices — all entries are positive real numbers [BP94].
- Stochastic matrices — nonnegative matrices whose rows sum to 1. These represent, for example, the transition probabilities of a Markov chain.
- Toeplitz matrices — matrices which are constant along each diagonal. See Section 1.2.
- Matrices with some prescribed entries.

These inverse eigenvalue problems can be viewed as any one of three different types of (constrained) optimisation problem:

(1) Find a matrix $X$ from the set $\mathbb{C}^{n\times n}$ which minimises some notion of distance to the set $\mathcal{S} \cap \{X \in \mathbb{C}^{n\times n} : \sigma(X) = \boldsymbol{\lambda}\}$.

(2) Find a matrix $X$ from the set $\mathcal{S}$ which minimises the error in the eigenvalues.

(3) Find a matrix $X$ in the set $\{X \in \mathbb{C}^{n\times n} : \sigma(X) = \boldsymbol{\lambda}\}$ such that some notion of distance to the set $X \in \mathcal{S}$ is minimised.

Which of these three points of view one takes affects how one might approach the computational problem. In the first regime one might consider an alternating direction method such as the Newton–projection algorithm in [CG02, §5.3]. In the second regime

one can parametrise the set $\mathcal{S}$ and optimise the parameters minimise the eigenvalue error, as in for example, [FNO87]. However, the eigenvalue error (in the Hausdorff distance for example) as a function of the entries of $X$, while continuous, is highly nonlinear, and is certainly not convex.

The third regime interests us most in this thesis. With this point of view we parametrise the set $\{X \in \mathbb{C}^{n\times n} : \sigma(X) = \boldsymbol{\lambda}\}$ and optimise those parameters to minimise the distance to the set $\mathcal{S}$. The author believes this to have some advantages. First, in many cases the set $\mathcal{S}$ is a linear, or at least convex space, and a projection onto $\mathcal{S}$ is simple to compute. Second, the complexity in set $\{X \in \mathbb{C}^{n\times n} : \sigma(X) = \boldsymbol{\lambda}\}$ discussed in the previous paragraph can be dealt with by noting that this set is in fact a manifold, and is acted upon by a Lie group (it is therefore a so-called homogeneous manifold; see Appendix B). The geometry of a manifold as opposed to unstructured set has several advantages. One can define distances and derivatives on the manifold, allowing a continuation approach, such as the isospectral flows developed in Chapter 2.

The idea for isospectral flows for the solution of inverse eigenvalues is quite simple. Given eigenvalues $\boldsymbol{\lambda}$, begin by computing a matrix $X_0$ with those desired eigenvalues (such as a diagonal matrix), and follow an isospectral flow $\dot{X} = [A(X), X]$, $X(0) = X_0$, as $t \to \infty$. The ideal situation is that the function $A$ has been designed so that the flow converges to matrices in the set $\mathcal{S}$.

Something that should be explicitly addressed is that these isospectral flows must at some point be solved numerically on a computer. It is not quite so simple though, because standard methods of solving ODE initial value problems such as linear multistep or Runge–Kutta methods fail to preserve the spectrum, defeating the point of the computation altogether [CIZ97]. Fortunately, methods of Geometric Numerical Integration allow the flow to be discretised while preserving the eigenvalues, the details of which we will not pursue in this thesis [IQ16], [IMKNZ00].

### 1.1.6 Infinite dimensional numerical linear algebra

The bread and butter of Numerical Linear Algebra is the following two problems [TBI97]. Let $A \in \mathbb{C}^{n\times n}$.

(i) Linear system problem: Given $b \in \mathbb{C}^n$, compute $x \in \mathbb{C}^n$ such that $Ax = b$

(ii) Eigenvalue problem: Compute the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $A$, along with eigenvectors.

Many other concepts in the field such as factorisations and iterations come from the desire to solve one or both of these two problems more effectively.

While these two problems are rich, fascinating and useful when taken at face value, it is worth noting that they often come from *discretising* and *truncating* an *infinite dimensional* linear operator. For a closed linear operator $L$ on the Hilbert space $\mathcal{H}$, the associated linear algebra problems are

(iii) Linear operator problem: Given $f \in \mathcal{H}$, compute $u \in \mathcal{H}$ such that $Lu = f$

(iv) Spectral problem: Compute the spectrum $\sigma(L)$, along with spectral measure and eigenvectors if appropriate.

For problem $(i)$ the operator $A$ is often a block operator involving a differential operator and boundary conditions [OT13].

These problems can be tackled with standard numerical linear algebra techniques by what is known as the *finite section method* (sometimes called the Galerkin method). Take the operators $P_1, P_2, \ldots$, in which $P_n : \mathcal{H} \to \mathbb{C}^n$, and define for an integer $n$,

$$A = P_n L P_n^+, \quad b = P_n f, \tag{1.52}$$

where $P_n^+ : \mathbb{C}^n \to \mathcal{H}$ is the Moore-Penrose pseudoinverse of $P_n$. The idea is then that if $n$ is taken sufficiently large, the solutions to these finite dimensional problems approximates the solution to the original problem to a desired accuracy. Indeed the finite section solutions converge to the proper solution as long as $L$ is invertible (with bounded inverse) and the projections $I_n = P_n^+ P_n$ satisfy $I_n \to I$ strongly in $\mathcal{H}$ as $n \to \infty$, by the following simple argument. Let $X = A^{-1}b$ and $u = L^{-1}f$. Then

$$\begin{aligned}
\|P_n^+ x - u\| &\leq \|P_n^+ x - I_n u\| + \|I_n u - u\| \\
&= \|P_n^+ A^{-1} b - I_n L^{-1} f\| + \|I_n u - u\| \\
&= \|I_n L^{-1}(I_n f - f)\| + \|I_n u - u\| \\
&\leq \|L^{-1}\| \|I_n f - f\| + \|I_n u - u\|
\end{aligned}$$

The strong convergence of $I_n$ to $I$ implies that this quantity converges to zero as $n \to \infty$.

In stark contrast, the finite section method for computing spectra can fail dramatically in some very simple cases [Arv94a], [Arv94b], [Han10], [LS04], [DP04]. For

example, in [Han10] Hansen considers the shift operator, which has entries

$$S = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & 1 & \\ & & & \ddots & \ddots \end{pmatrix}. \tag{1.53}$$

The spectrum of all finite sections of this operator is $\{0\}$, but the spectrum of the operator is the closed unit disc [BS13] (see also Section 1.1.2). So the finite section method fails dramatically. It fails even more dramatically for the Laurent operator version, where the spectrum of the finite sections are again, $\{0\}$, but the spectrum of the full operator is the unit *circle* — the computed eigenvalues and the actual spectrum have empty intersection. This phenomenon is called *spectral pollution*, where the eigenvalues of a finite section can appear anywhere in the convex hull of the essential spectrum and persist there for all $n \to \infty$ [LS04], [DP04].

While the failure of the finite section method is most spectacular for non-selfadjoint operators such as $S$ above, it also fails for self-adjoint operators, such as Toeplitz operators with discontinuous symbols [LS04]. For example, consider the selfadjoint Laurent operator with matrix entries

$$a_{i,j} = \begin{cases} \frac{2\sin\frac{i-j}{2}}{i-j} & \text{for } i,j \in \mathbb{Z}, i \neq j \\ 1 & \text{for } i,j \in \mathbb{Z}, i = j. \end{cases} \tag{1.54}$$

The symbol of this operator is discontinuous. It is the function that is 1 if $\mathrm{Re}\,(z) \geq 0$ and 0 otherwise. This operator has applications in the theory of Prolate Spheroidal Wave functions [Sle78] and Fourier extensions [MH16]. The spectrum of the operator is the image of the symbol on the unit circle (see Section 1.1.2 and [BS13]), which is the set $\{0, 1\}$. However, the eigenvalues of the finite sections fill the interval $[0, 1]$ as $n \to \infty$ [MH16].

There are theorems guaranteeing convergence of the finite section method for computing the spectrum in special cases or weakened notions of convergence. Good resources for these results are [Arv94a], [Arv94b], [Han08], [Han10], [Han11].

One example pertinent to this thesis is, let $J_n$ be the $n \times n$ principal submatrix of a Jacobi operator $J$ whose spectral measure is $\mu$ (see Subsection 1.1.3), and define

$$\mu_n = \sum_{k=1}^{n} w_k \delta_{\lambda_k},$$

where $\lambda_1, \ldots, \lambda_n$ are the $n$ eigenvalues of $J_n$, $w_k > 0$ is the square of the first entry of the $\lambda_k$-eigenvector, and $\delta_\lambda$ is the Dirac delta measure centered at $\lambda$. Then $\mu_n(f) \to \mu(f)$ for every compactly supported continuous function $f$ on the real line. The spectral measures $\mu_n$ of $J_n$ converge weakly to the spectral measure $\mu$ of $J$.

Even in the cases where the finite section method approximates the spectrum well, it is in some ways unsatisfactory. For example, take the *free Jacobi operator*,

$$\Delta = \begin{pmatrix} 0 & \frac{1}{2} & & \\ \frac{1}{2} & 0 & \frac{1}{2} & \\ & \frac{1}{2} & 0 & \ddots \\ & & \ddots & \ddots \end{pmatrix}. \tag{1.55}$$

The spectral measure is the semi-circle, $\mu_\Delta(s) = \frac{2}{\pi}\sqrt{1-s^2}$ supported on $[-1, 1]$ (see Subsection 1.1.3). This measure is absolutely continuous with respect to Lebesgue measure, but the approximations discussed above will all be discrete. If the spectral measure has a mixture of absolutely continuous parts and discrete parts, then it is not possible to distinguish them by looking at the approximations to the measure alone.

Another issue with a finite section approach is simply that the size of the truncation required is not known *a priori*, and so in practice many different sizes of truncation are used until the approximation appears correct. Even then there is no guarantee that the method has converged (if it does converge).

Using finite section methods along with proofs that the computed sequence of solutions will converge is known by some authors as *Infinite Dimensional Numerical Linear Algebra* [Arv94a], [Arv94b], [Han10]. However, there has been recent interest in going further than this by designing algorithms which operate on the infinite-dimensional objects directly. The distinction is subtle.

For the linear operator problem, one approach is to view a banded infinite dimensional linear system as an infinite linear recurrence. One of the earliest examples of this approach is Olver's algorithm [Olv67], which adaptively applies Gaussian elimination to the infinite dimensional system until the testable condition that a back substitution

can be performed is met. A recent descendent of Olver's algorithm is the adaptive QR method introduced by Olver (a recent descendent of the aforementioned Olver) and Townsend, which performs infinite dimensional numerical linear algebra for a spectral method in the solution of differential equations [OT13], [OT14].

One aspect of such an approach is that the infinite dimensional objects must be encodable in a form that can be stored on a computer. This encoding is related to the computer science principle of *lazy evaluation*, where the data itself is not stored, but a way to compute the data should the computer require that data. For example, the Laurent operator in equation (1.54) can be stored in a tailor-made data structure. When the compute requests the $(2, 3)$ element, it can compute that entry using the given formula and return it as if it were actually storing the data.

The ability to use such data structures to perform infinite dimensional numerical linear algebra requires more than just the ability to store the operators and vectors. They must be highly structured enough so that useful *operations* can applied to them (such as a QR factorisation) and the output of those operations is also a predictably, highly structured object which can then be stored too. Recent research on spectral methods produced operators with banded-plus-finite-rank matrix structure, with simple asymptotics of the entries, and led to the development of a practical framework for solving infinite dimensional linear systems on a computer [OT13], , [SO17] [Olvb]. The key to producing these highly structured matrices is an appropriate choice of basis, so it perhaps it is not unreasonable to suggest that many operators in applications can be represented by highly structured matrices with the right choice of bases.

The ApproxFun project in Julia [Olvb], pioneered in the main by Sheehan Olver and influenced by the Chebfun project in Matlab [DHT14], has implemented many of these ideas for the manipulation of operators and functions. The software is open source and available at https://github.com/JuliaApproximation/ApproxFun.jl.

Numerical methods for the spectral problem which do not resort to the finite section method are surprisingly thin on the ground. Deift, Li and Tomei studied the Toda flow with infinitely many variables [DLT85] (see Subsection 1.1.4). Hansen recently investigated the infinite dimensional QR algorithm [Han09], and found that for banded operators, any single entry of a single iterate of the QR algorithm for the full infinite dimensional matrix is exactly equal to that entry of that iterate computed by the QR algorithm applied on a sufficiently large finite section of the operator. Hence technically this approach does resort to finite sections, but so that the computation is equivalent to computing on the infinite dimensional object.

Infinite dimensional linear algebra techniques for the spectral problem on some highly structured operators are given in Chapters 4 and 5. Some of the ideas in these chapters have been implemented by the author and Sheehan Olver (University of Sydney) in the open source Julia package *SpectralMeasures*. Some of the code is included in Appendix A. It is freely available online at https://github.com/JuliaApproximation/SpectralMeasures.jl and uses the ApproxFun features extensively.

## 1.2 Outline and contributions of the thesis

This section does not attempt to provide a full historical and bibliographical context. That is done at the beginning of each chapter, and with fuller explanation of the contributions themselves.

Throughout the thesis, wherever a Theorem, Lemma, Proposition etc. is stated and it is due to another author, a citation will be included in brackets. If there is no citation in brackets then the result is original and due to this author.

### 1.2.1 Isospectral flows

Chapter 2 is about isospectral flows. Most of the material covered in the chapter is already in the literature such as [HM94] and known before the year 2000, but there are some minor novel results.

In Section 2.1 we give the basic elementary properties of isospectral flows. In Section 2.2 we discuss the basics of relationships between the QR algorithm, the Toda flow and the double bracket flow. These types of isospectral flow have gradient structure and, for appropriate initial data, converge as $t \to \infty$.

In Section 2.3 we discuss the Bloch-Iserles flow, which in contrast to the flows related to the QR algorithm are completely integrable Hamiltonian systems, so are oscillatory with dynamics that are diffeomorphic to inertial motion on a torus. We briefly discuss the new observation that the Lax pair for the KdV equation is a modified infinite dimensional Bloch-Iserles system. This new result is only briefly explored and so its consequences for the KdV equation and the Bloch-Iserles equation is not clear at present.

In Section 2.4 we derive gradient flows on isospectral manifolds with an arbitrary metric and discuss convergence of these flows to stationary points. Isospectral gradient flows for arbitrary potential functions $\Psi : \mathbb{C}^{n\times n} \to \mathbb{R}$ with respect to the so-called

normal metric were given by Brockett [Bro93] and gradients flows for which $\Psi$ is quadratic (such as the Toda flow) were derived in [CD90], [Bro93, Rem. 2], and [BG98, Prop. 2.3]. A small contribution of this chapter is to take these results and put them together into one uniform language to specify the isospectral gradient flows for an arbitrary metric and arbitrary potential function (Theorem 2.4.6). We also describe the stationary points of these flows, showing that they are independent of the choice of metric (Theorem 2.4.7), and also that the stability of the stationary point also does not depend on the metric chosen.

The results of Section 2.4 on isospectral gradient flows is applied to the QR algorithm in Section 2.5. This is used to put double bracket flows, Toda flows, and QR flows into a single framework of gradient flows with different metrics on the isospectral manifold, which has never been done before. The main contribution of this chapter is the following. We define a new isospectral flow called the *gradient QR flow* (see Theorem 2.5.8), which can be written

$$\dot{X} = \underbrace{[\pi_2(f(X)), X]}_{\text{QR flow}} + \underbrace{\frac{1}{2}\left[X, f(X)^H\right]}_{=0 \text{ if } X \text{ is normal}}. \tag{1.56}$$

This isospectral flow is the gradient flow for the function $\Psi(X) = \frac{1}{2}\|f(X) - D\|_F^2$ where $D = \operatorname{diag}(n, n-1, \ldots, 2, 1)$ with a certain metric (given in Definition 2.5.6). If the initial matrix $X(0)$ is normal, then this gradient flow coincides with the QR flow, but in general the term $\frac{1}{2}\left[X, f(X)^H\right]$ is nonzero. This gives some insight into the nonconvergence of the QR algorithm for nonnormal matrices. Note that there are explicit examples of nonconvergence for the nonshifted QR algorithm [Bat90],[Day96].

### 1.2.2 The symmetric Toeplitz inverse eigenvalue problem

Chapter 3 is about computing an $n \times n$ real symmetric Toeplitz matrix with prescribed spectrum $\boldsymbol{\lambda} \in \mathbb{R}^n$ using isospectral flows. The results of the chapter give incremental contributions and new insight to the current state of the art which was pioneered mainly by Moody Chu [CD89], [CD90], [Chu93], [Chu98], [CG02], [CG05].

The basic idea of taking a initial matrix $Y_0 \in \mathbb{R}^{n\times n}_{\text{sym}}$ with the prescribed spectrum (such as a diagonal matrix) and numerically simulating an isospectral flow that has been designed to converge to matrices with a certain structure in order to solve an inverse eigenvalue problem appears to have been first put forward by Chu and Driessel

[CD89], [CD90]. In [CD90] Chu and Driessel give the gradient descent flow,

$$\dot{Y} = [[P_T(Y), Y], Y], \tag{1.57}$$

where $P_T(Y)$ is the orthogonal projection of $\mathbb{R}^{n\times n}_{\text{sym}}$ onto the subspace of Toeplitz matrices. Chu and Driessel also introduced the following flow in [CD89], which we call Chu's flow.

$$\dot{Y} = [B(Y), Y], \tag{1.58}$$

where $B$ is the *Toeplitz annihilator*,

$$B(Y)_{i,j} = \begin{cases} y_{i,j-1} - y_{i+1,j} & \text{if } i < j, \\ 0 & \text{if } i = j, \\ y_{i,j+1} - y_{i-1,j} & \text{if } i > j. \end{cases} \tag{1.59}$$

To understand the motivation behind this flow, consider the case $n = 4$:

$$B(Y) = \begin{pmatrix} 0 & y_{1,1} - y_{2,2} & y_{1,2} - y_{2,3} & y_{1,3} - y_{2,4} \\ y_{2,2} - y_{1,1} & 0 & y_{2,2} - y_{3,3} & y_{2,3} - y_{3,4} \\ y_{3,2} - y_{2,1} & y_{3,3} - y_{2,2} & 0 & y_{3,3} - y_{4,4} \\ y_{4,2} - y_{3,1} & y_{4,3} - y_{3,2} & y_{4,4} - y_{3,3} & 0 \end{pmatrix}. \tag{1.60}$$

Clearly, $B(Y) = 0$ if and only if $Y$ is Toeplitz. Chu also showed that if $Y$ has distinct eigenvalues then $[B(Y), Y] = 0$ if and only if $Y$ is Toeplitz (see [Chu93] and Proposition 3.1.2).

In Section 3.1 we conduct a basic study these two isospectral flows (before returning to them both later in the chapter). This contributes to the previous numerical studies in [Chu93],[DS99],[Zan98]. In [CG05], Chu and Golub report numerical evidence of some stable stationary points of the gradient descent flow which are not Toeplitz matrices. This can be slightly problematic for the algorithm, which must be restarted from a different initial datum if such a stationary point is encountered.

Symmetric Toeplitz matrices are bisymmetric, which means they are symmetric along both the top-left-to-bottom-right and top-right-to-bottom-left diagonals. Let $X \in \text{Bisym}(n)$, the space of bisymmetric $n \times n$ matrices, and suppose $X$ has $n$ distinct eigenvalues. Then the bisymmetric isospectral manifold is

$$\mathcal{BI} = \{Y \in \text{Bisym}(n) : \text{eigs}(Y) = \text{eigs}(X)\}. \tag{1.61}$$

One the main contributions of this chapter is a full formal description of how the restriction to *bisymmetric* matrices affects the isospectral manifold. In Section 3.2, we prove that this manifold has $\binom{n}{p}$ connected components, where $p = \lceil \frac{n}{2} \rceil$. Each component has dimension $\frac{1}{2}p(p-1)+\frac{1}{2}q(q-1)$ where $q = n-p$, and may be parametrised by a connected Lie group of centrosymmetric orthogonal matrices. Aspects of this fact appear to be known in the literature, as Chu discusses the different connected components of the manifold for the $3 \times 3$ case but does not go into detail [Chu93].

The fact that each connected component of the bisymmetric isospectral manifold is acted upon by a Lie group with dimension $\frac{1}{2}p(p-1) + \frac{1}{2}q(q-1)$ is important. This allows us to parametrise the manifold by the associated Lie algebra, which has the small dimension. Using this, we can reduce the $3 \times 3$ bisymmetric isospectral flow to a one dimensional flow, the $4 \times 4$ bisymmetric isospectral flow to a two dimensional flow, and the $5 \times 5$ bisymmetric isospectral flow to a four dimensional flow and so on. In Subsection 3.2.3 we derive an analytical solution for the trajectories of $3 \times 3$ bisymmetric isospectral flows. Then in Section 3.3, conduct a numerical study of the $4 \times 4$ gradient flow and Chu flow for the inverse Toeplitz eigenvalue problem and gain some new insights that were not feasible if you only consider the 5 dimensional phase space rather than this reduced 2 dimensional one.

At the end of the chapter we briefly discuss an extremely impractical, brute force approach to the computation which produces isospectral iterates $Y_0, Y_1, \ldots$ which converge to a symmetric Toeplitz matrix. The reason is purely theoretical, to show that the Solvability Complexity Index of the problem is 1 (see [BAHNS15a] and Section 4.5).

### 1.2.3 Computing spectra of Jacobi operators

In Chapter 4 we show that the computation and theoretical study of the spectra and spectral measure of a Jacobi operator $J$ which is a structured perturbation of another Jacobi operator $D$ whose spectral theory is known, can be conducted using the *connection coefficient matrix* between $J$ and $D$. Almost all of this chapter consists of original results.

Suppose that $D$ is the second Jacobi operator, and let $Q_k(s)$ denote its orthonormal polynomials. The connection coefficient matrix between $J$ and $D$, denoted $C = C_{J \to D} = (c_{ij})_{i,j=0}^{\infty}$ is defined to be the upper triangular matrix representing the change

of basis between $(P_k)_{k=0}^{\infty}$ and $(Q_k)_{k=0}^{\infty}$ in the following manner:

$$P_k(s) = c_{0k}Q_0(s) + c_{1k}Q_1(s) + \cdots + c_{kk}Q_k(s). \tag{1.62}$$

Alternatively this can be written,

$$\begin{pmatrix} P_0(s) \\ P_1(s) \\ P_2(s) \\ \vdots \end{pmatrix} = C^T \begin{pmatrix} Q_0(s) \\ Q_1(s) \\ Q_2(s) \\ \vdots \end{pmatrix} \quad \text{for all } s \in \mathbb{C}. \tag{1.63}$$

Connection coefficient matrices have been well-studied [Ask75, GM09], but it does not appear to have been noted that the connection coefficients are relevant and useful in the spectral theory of Jacobi operators. When viewed as acting on finite vectors, $J$, $D$ and $C$ are related by

$$J = C^{-1}DC. \tag{1.64}$$

Consequently, when $C$ is a bounded and invertible operator on $\ell^2$, we have $\sigma(J) = \sigma(D)$. More significantly, we further show that when $C$ is neither bounded nor invertible, the matrix entries are still informative about the spectra of $J$ and $D$. For example, if we let $\nu$ denote the spectral measure for $D$, the connection coefficients matrix $C = C_{J\to D}$ determines the existence and certain properties of the Radon–Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ (see Appendix D.1). In Section 4.2 we derive new results regarding these relationships, including the formula

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = \sum_{k=0}^{\infty} c_{0,k}P_k, \tag{1.65}$$

whenever the series converges at least in the probabilists' weak sense.

In Sections 4.3, 4.4 and 4.5, we attention to the case where $D = \Delta$, the free Jacobi operator (see equation (1.55)). We assume that $J$ is a Jacobi operator of the form $J = \Delta + K$, where $K$ is compact. Jacobi operators of this form have been studied extensively because of their links to Schrödinger operators and to classical orthogonal polynomials [DS06a, DS06b, DN86, DE15, GNR16, GC80, KS03, NVA92, VAG89, VA90, VA94, VA91].

We prove the following more specific theorems about the spectra of this class of Jacobi operators $J$, and by an appropriate scaling and shifting by the identity, that of all Jacobi operators which are Toeplitz-plus-compact.

If $J$ is a *finite rank perturbation* of $\Delta$, i.e. there exists $n$ such that

$$\alpha_k = 0, \quad \beta_{k-1} = \frac{1}{2} \text{ for all } k \geq n, \tag{1.66}$$

- Theorem 4.3.8: The connection coefficient matrix $C_{J\to\Delta}$ can be decomposed into $C_{Toe} + C_{fin}$ where $C_{Toe}$ is Toeplitz, upper triangular and has bandwidth $2n-1$, and the entries of $C_{fin}$ are zero outside the $n-1 \times 2n-1$ principal submatrix.

- Theorem 4.3.21: let $c$ be the Toeplitz symbol of $C_{Toe}$. It is a degree $2n-1$ polynomial with $r \leq n$ roots inside the complex unit disc, all of which are simple. The spectrum of $J$ is

$$\sigma(J) = [-1,1] \cup \left\{ \lambda_k := \frac{1}{2}(z_k + z_k^{-1}) : c(z_k) = 0, |z_k| < 1 \right\}, \tag{1.67}$$

  and the spectral measure is given by the formula

$$\mu(s) = \frac{1}{p_C(s)}\mu_\Delta(s) + \sum_{k=1}^{r} \frac{(z_k - z_k^{-1})^2}{z_k c'(z_k) c(z_k^{-1})} \delta_{\lambda_k}(s), \tag{1.68}$$

  where $p_C(s) = \sum_{k=0}^{2n-1} c_{0,k} P_k(s) = \sum_{k=0}^{2n-1} \langle e_k, CC^T e_0 \rangle U_k(s)$.

We extend these results to the case where $J = \Delta + K$ where $K$ is a trace class operator. In that case, $C = C_{Toe} + C_K$ where $C_{Toe}$ is upper triangular Toeplitz and $C_K$ is compact as an operator in an appropriate topology we make clear in Section 4.4. Furthermore, let $c$ be the Toeplitz symbol of $C_{Toe}$. It is analytic in the unit disc with real inside the complex unit disc. The discrete eigenvalues, as in the Toeplitz-plus-finite-rank case are of the form $\frac{1}{2}(z_k + z_k^{-1})$ where $z_k$ are the roots of $c$ in the open unit disc.

Following the pioneering work of Ben-Artzi–Hansen–Nevanlinna–Seidel on the Solvability Complexity Index [BAHNS15a, BAHNS15b, Han11], we prove the following theorems about computability. We assume real number arithmetic, and the results do not necessarily apply to algorithms using floating point arithmetic.

- Theorem 4.5.7 : If $J$ is a Toeplitz-plus-finite-rank Jacobi operator, then in a finite number of operations, the absolutely continuous part of the spectral measure is computable exactly, and the locations and weights of the discrete part of the spectral measure are computable to any desired accuracy. If the rank is known $a$

*priori* then the algorithm can be designed to terminate with guaranteed error control.

- Theorem 4.5.9 : If $J = \Delta + K$ is a Toeplitz-plus-compact Jacobi operator, then in a finite number of operations, the spectrum of $J$ is computable to any desired accuracy in the Hausdorff metric on subsets of $\mathbb{R}$. If the quantity $\sup_{k\geq m} |\alpha_k| + \sup_{k\geq m} |\beta_k - \frac{1}{2}|$ can be estimated for all $m$, then the algorithm can be designed to terminate with guaranteed error control.

The significance of these computability theorems is that they extend the known class of operators whose spectra can be computed with error control.

### 1.2.4 Infinite dimensional QL algorithm

In Chapter 5 we discuss the infinite dimensional QL algorithm. Almost all of this chapter consists of original results.

The QL algorithm is the same as the QR algorithm, except that instead of computing QR factorisations we compute QL factorisations where $L$ is lower triangular. In finite dimensions the QR and QL algorithms are equivalent in the following sense. Applying some iterations of the QR algorithm to a matrix has the same effect as rotating the entries $\pi$ radians, applying the same number of iterations of the QL algorithm and then rotating back. Hence in finite dimensions there is no real difference between the QR and QL algorithms in terms of convergence [Par80].

For the QR algorithm on tridiagonal matrices, Wilkinson shifts can be applied to the algorithm give global convergence of the bottom–right entry to an eigenvalue of the input matrix [Wat07], [Par80]. The QR algorithm has been generalised to the infinite dimensional case of bounded operators on $\ell^2$ [Han08], [Han09], and so too has the related Toda flow [DLT85]. However, there is an issue with the use of shifts to accelerate convergence: there is no bottom–right entry! In finite dimensions, since the QL algorithm is equivalent to the QR algorithm after rotating the entries $\pi$ radians, we can force the top–left entry to converge rapidly to an eigenvalue using Wilkinson shifts, but this logic does not follow for the infinite dimensional case as there is no infinite dimensional analogue of such a rotation of entries.

Olver and Townsend proposed the following idea a footnote of [OT14]. In principle, if one could perform the QL algorithm to an infinite dimensional matrix, it could be possible to utilise shifts to yield rapid convergence of the top–left entry to an eigenvalue

(if the matrix has any point spectrum). However, there were no known methods to compute the QL factorisation of a (non-compact) infinite dimensional matrix.

One of the main contributions of this chapter is Theorem 5.1.17, which goes partway to solving the problem posed by Olver and Townsend. Here is the gist. For some bounded operators $A$ on $\ell^2$ we can find an analytical solution to $A = QL$. For example, if $A = \Delta - \frac{5}{4}I$ where $\Delta$ is the free Jacobi operator, then the QL factorisation of $A$ is

$$\Delta - \frac{5}{4}I = \begin{pmatrix} -\frac{\sqrt{3}}{2} & \frac{1}{2} & & & & \\ -\frac{\sqrt{3}}{4} & -\frac{3}{4} & \frac{1}{2} & & & \\ -\frac{\sqrt{3}}{8} & -\frac{3}{8} & -\frac{3}{4} & \frac{1}{2} & & \\ -\frac{\sqrt{3}}{16} & -\frac{3}{16} & -\frac{3}{8} & -\frac{3}{4} & \frac{1}{2} & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \frac{\sqrt{3}}{2} & & & & \\ -1 & 1 & & & \\ \frac{1}{4} & -1 & 1 & & \\ & -\frac{1}{4} & -1 & 1 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{1.69}$$

A proof of this new result (for a general Toeplitz-plus-finite-rank Jacobi operator) is given in Theorem 5.2.6. Such QL factorisations can be used to compute the QL factorisation of banded matrices which contain said matrix as the bottom-right submatrix. Specifically, suppose that $A$ is a banded, bounded operator on $\ell^2$ with bandwidth $b$ and block form

$$A = \begin{pmatrix} A_n & B \\ C & A_\infty \end{pmatrix}, \tag{1.70}$$

where $A_n \in \mathbb{R}^{n\times n}$ and $A_\infty$ has an *a priori* known QL factorisation $A_\infty = Q_\infty L_\infty$. Then the way to compute the QL factorisation of $A$ is seen by noting that

$$\begin{pmatrix} I_n & \\ & Q_\infty^T \end{pmatrix} \begin{pmatrix} A_n & B \\ C & A_\infty \end{pmatrix} = \begin{pmatrix} A_n & B \\ Q_\infty^T & L_\infty \end{pmatrix}. \tag{1.71}$$

The right hand side has finitely many nonzero entries above the diagonal (bandedness implies $B$ has finitely many nonzero entries). Therefore the standard approach of introducing zeros from the rightmost column applies, so this can be used to complete the QL factorisation in finitely many operations. Full details are given in Theorem 5.1.17.

One of the most surprising results to come out of this work is that the existence of a QL factorisation is not always guaranteed, unlike the case for the QR factorisation [Han08], [Han09]. We prove that a Jacobi operator has a QL factorisation if and only if the essential spectrum does not contain zero (see Theorem 5.2.2). Notably, the free

Jacobi operator $\Delta$ does not have a QL factorisation. In Theorem 5.1.3, Theorem 5.1.14 and Corollary 5.1.15 we prove generalisations of this for banded, selfadjoint operators, but their exact statements have some technical points we will explain there and not here.

In Section 5.1 we prove existence and nonexistence results for QL factorisations of bounded selfadjoint case and briefly indicate if there is an easy generalisation of a result to the non-selfadjoint case. In Section 5.2 we restrict these results to Jacobi operators and find that the statement of the results is simpler. Then in Section 5.2.2 we make practical considerations for running the QL algorithm for Jacobi operators on a computer, and derive a method to compute the QL factorisation of a Toeplitz-plus-finite-rank Jacobi operators, using only a finite amount of memory.

In Section 5.3 we consider the infinite dimensional QL algorithm which utilises these infinite dimensional QL factorisations. We prove that for a bounded Jacobi operator $J$ such that there is an eigenvalue $\lambda_0$ satisfying

$$0 < |\lambda_0| < \eta := \min_{\lambda \in \sigma(J) \setminus \lambda_0} |\lambda|, \tag{1.72}$$

the unshifted QL algorithm converges in the sense that the $(0,1)$ entry is $\mathcal{O}\left(\left|\frac{\lambda_0}{\eta}\right|^k\right)$. This implies (as is done in the finite dimensional case [Par80]) that if a shift is chosen sufficiently close to an isolated eigenvalue, then there will be rapid convergence of the top-left entry to that eigenvalue.

### 1.2.5 Computing functions of operators

To conclude the thesis we combine ideas from Chapters 4 and 5 to compute an invertible operator $U$ such that for a Jacobi operator $J$ that is a finite rank perturbation of $\Delta$, we have

$$UJU^{-1} = \begin{pmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \lambda_r & \\ & & & \Delta \end{pmatrix}, \tag{1.73}$$

where $\lambda_1, \ldots, \lambda_r$ are the discrete eigenvalues of $J$ (if there are any). Appropriate scaling and shifting by the identity gives a new canonical form for Toeplitz-plus-finite-rank

Jacobi operators. For functions $g : \sigma(J) \to \mathbb{R}$, we have

$$g(J) = \begin{pmatrix} g(\lambda_1) & & & & \\ & \ddots & & & \\ & & g(\lambda_r) & & \\ & & & g(\Delta) \end{pmatrix}. \tag{1.74}$$

In Proposition 5.4.1 we prove the apparently new result that if $g(s) = \sum_{k=0}^{m} a_k T_k(s)$, then

$$g(\Delta) = \frac{1}{2}\begin{pmatrix} 2a_0 & a_1 & a_2 & a_3 & \cdots \\ a_1 & 2a_0 & a_1 & a_2 & \ddots \\ a_2 & a_1 & 2a_0 & a_1 & \ddots \\ a_3 & a_2 & a_1 & 2a_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} - \frac{1}{2}\begin{pmatrix} a_2 & a_3 & a_4 & a_5 & \cdots \\ a_3 & a_4 & a_5 & a_6 & \cdots \\ a_4 & a_5 & a_6 & a_7 & \cdots \\ a_5 & a_6 & a_7 & a_8 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{1.75}$$

If a given function $g$ is Lipschitz in $[-1, 1]$, $g(s)$ can be approximated by a Chebyshev series (for example using Chebfun or ApproxFun) [Tre13]. Hence functions of $\Delta$ can be computed with relative ease, and applied to finite support vectors quickly using the FFT (because $g(\Delta)$ is the sum of a Toeplitz and a Hankel matrix).

The final part of the thesis utilises this for the solution of a discrete Schrödinger equation with double potential wells (demonstrating discrete quantum tunnelling), and some diffusion equations, including fractional order ones. A brief comparison to the traditional finite section method approach is made (see Subsection 1.1.6).

瞎子摸象，不识大体

(xiā zi mō xiàng, bù shí dà tǐ)

Lit. Blind men groping an elephant don't know the whole body
Fig. Not seeing the whole story, mistaking a part for the whole

# Chapter 2

# Isospectral flows

To analyse isospectral algorithms, we advocate an understanding of *isospectral flows*. These are the continuous versions of isospectral algorithms, functions $X : \mathbb{R} \to \mathbb{C}^{n\times n}$ which satisfy

$$X(t) = P(t)X(0)P(t)^{-1}, \qquad \text{for } t \in \mathbb{R},$$

where $P : \mathbb{R} \to \mathbb{C}^{n\times n}$ is a $C^1$ function such that $P(t)$ is invertible for all $t$. All such flows satisfy the differential equation $\dot{X}(t) = [A(t), X(t)]$ for a continuous function $A : \mathbb{R} \to \mathbb{C}^{n\times n}$, and methods of calculus and differential geometry can be used to elucidate their convergence properties [Wat84], [HM94], [Bro93], [Chu08]. In the theory of integrable systems, these flows are known as *Lax Pairs* [Lax68], [BV90], [Tes01], [Dei00].

This chapter in the main covers known results about isospectral flows, but has several new results. We begin in Section 2.1 with the elementary properties, most of which can be found in textbooks such as [HM94] and [CG05]. In Section 2.2 we discuss the (already known) connections between the Toda flow, double bracket flow, the QR algorithm and the QR flow:

- Toda flow: $\dot{Y} = [Y_U - Y_L, Y]$, where $Y \in \mathbb{R}^{n\times n}_{\text{sym}}$
- Double bracket flow: $\dot{Y} = [[S, Y], Y]$, where $Y, S \in \mathbb{R}^{n\times n}_{\text{sym}}$
- QR flow: $\dot{X} = [X, \pi_1(f(X))]$, where $f$ is a function which is analytic on the eigenvalues of $X \in \mathbb{C}^{n\times n}$,

where $\pi_1$ is the projection

$$\pi_1(Y) = Y_L - Y_L^H + i\text{Im}(Y_D). \tag{2.1}$$

It was shown by Symes that the QR algorithm which produces a sequence of isospectral iterates $Y^{(0)}$,$Y^{(1)}$, $Y^{(2)}$ is interpolated by a function of the QR flow, in that $\exp(f(Y(k))) = Y^{(k)}$ [Sym82] (see Theorem 2.2.11). Setting $f = \log$ gives the direct correspondence.

It is clear that the case $f(z) = z$ and $X \in \mathbb{R}^{n\times n}_{\text{sym}}$ the QR flow is equal to the Toda flow. Bloch observed that for tridiagonal matrices the Toda flow and the double bracket flow with $S = \text{diag}(n, n-1, \ldots, 1)$ are equal [Blo90] (see Lemma 2.2.9). Later, Chu showed that both the Toda flow and the double bracket flow in which $S$ is a diagonal matrix can both be written in the form

$$\dot{Y} = [A \circ Y, Y]\,, \tag{2.2}$$

where $A$ is a skew-symmetric matrix and $\circ$ denotes the Hadamard product (element-wise product), extending the relationship between the Toda flow and double bracket flow to all symmetric matrices [Chu95] (see Section 2.5.1).

In Section 2.3 we briefly discuss Bloch–Iserles flows, which are flows on symmetric matrices of the form

$$\dot{Y} = \left[N, Y^2\right] = [NY + YN, Y]\,, \qquad Y(0) = Y_0 \in \mathbb{R}^{n\times n}_{\text{sym}}, \tag{2.3}$$

for a fixed skew-symmetric matrix $N$. There are two reasons for including this short section in the thesis. First, is simply to give an example of an isospectral flow with different behaviour to other flows considered in the thesis. The second reason is to state a new result about infinite-dimensional Bloch–Iserles systems, which is that the Lax pair for the KdV equation can be parametrised in terms of the Bloch-Iserles system with $N = \partial_x$, the differential operator on $L^2(\mathbb{R})$. A full exploration of this fact is beyond the scope of this thesis, but it is worth the brief discussion.

In Section 2.4 we discuss gradient flows which evolve on an isospectral manifold. Gradient flows for quadratic potential functions on isospectral manifolds with respect to the normal metric are given in [CG05]. They state that a function of the form $\Psi : \mathbb{R}^{n\times n}_{\text{sym}} \to \mathbb{R}$, where

$$\Psi(Y) = \|Y - P(Y)\|_F^2, \tag{2.4}$$

where $P$ projects onto an affine subspace of $\mathbb{R}^{n\times n}_{\text{sym}}$, the isospectral gradient flow with respect to the normal metric is

$$\dot{Y} = [[P(Y), Y], Y], \tag{2.5}$$

which generalises the double bracket flow in which $P(Y) = S$.

There exists research in the literature on generalising the double bracket flow to be an isospectral gradient flow in a modified metric [BG98, Prop. 2.3], [Bro93, Rem. 2]. A general form of isospectral gradient flow for an arbitrary potential function was given in [Bro93], but this is only with respect to the normal metric. The three papers mentioned just now focus on an abstract framework involving compact semi-simple Lie algebras. There doesn't appear to exist in the literature, firstly, a translation of these results into more "applied" language, nor secondly, the straightforward generalisation of these results to an arbitrary potential and an arbitrary metric. In Theorem 2.4.6 we provide this: For a $C^1$ function $\Psi : \mathbb{C}^{n\times n} \to \mathbb{R}$ be a $C^1$ function. The isospectral gradient flow for $\Psi$ with respect to the metric $\langle [A, X], [B, X] \rangle_g = \langle A, L_X B \rangle_F$ as defined in (2.58) is of the form,

$$\dot{X}(t) = -\left[ L^{-1}_{X(t)} \pi_{\mathfrak{g}} \left[ \nabla\Psi(X(t)), X(t)^H \right], X(t) \right], \tag{2.6}$$

where $\pi_{\mathfrak{g}}$ denotes the orthogonal projection onto a Lie algebra which defined which isospectral deformations are allowed ($\mathfrak{g}$ is usually $\mathfrak{so}(n)$).

In Section 2.5 we define of a new isospectral flow we call a *gradient QR flow*. For an function $f$ analytic on the eigenvalues of a matrix $X_0 \in \mathbb{C}^{n\times n}$, the gradient $f$–QR flow is

$$\dot{X} = [X, \pi_1(\text{Herm}(f(X)))], \tag{2.7}$$

where $\text{Herm}(Y) = \frac{1}{2}(Y + Y^H)$. This flow has the property that the function $\Psi(X) = \frac{1}{2}\|f(X) - D\|_F^2$ where $D = \text{diag}(n, n-1, \ldots, 2, 1)$, is monotonically decreasing along the trajectories, and

$$\text{dist}(\text{Herm}(f(X(t))), \text{Diag}(n)) \to 0 \text{ as } \to \infty. \tag{2.8}$$

See Theorem 2.5.11 and Theorem 2.5.15. What is the relationship between the QR flow and the gradient QR flow? The gradient QR flow can be rewritten as $[X, \pi_1(\text{Herm}(f(X)))] = [X, \text{Herm}(f(X))] - [X, \pi_2(\text{Herm}(f(X)))] = \frac{1}{2}\left[X, f(X)^H\right] -$

$[X, \pi_2(f(X))]$. Hence the gradient QR flow is also equal to

$$\dot{X} = \underbrace{[\pi_2(f(X)), X]}_{\text{QR flow}} + \underbrace{\frac{1}{2}\left[X, f(X)^H\right]}_{=0 \text{ if } X \text{ is normal}}. \tag{2.9}$$

Therefore when $X_0$ is a normal matrix the QR flow and the gradient QR flow coincide. The significance of this fact is that it adds some insight into why theoretical guarantees for the convergence of the QR algorithm with normal initial matrices such as those in [EH75] and [Bat94] have been easily found by researchers, but convergence results for nonnormal matrices have been shown to be impossible in general [Bat90],[Day96]. One difference it is that for normal matrices, the QR algorithm interpolates a certain gradient flow, but for nonnormal matrices this specific gradient structure (if any) does not exist.

## 2.1 Elementary properties

An isospectral flow actually has a slightly stronger property than preservation of the eigenvalues.

**Proposition 2.1.1** (See [HM94]). *Let $X : \mathbb{R} \to \mathbb{C}^{n\times n}$ be a $C^1$ function. Then the following are equivalent.*

*(i) There exists a $C^1$ function $P : \mathbb{R} \to \mathrm{SL}(n)$ such that*

$$X(t) = P(t)X_0P(t)^{-1}. \tag{2.10}$$

*(ii) There exists a continuous function $A : \mathbb{R} \to \mathfrak{sl}(n)$ such that*

$$\dot{X}(t) = [A(t), X(t)]. \tag{2.11}$$

*The two auxiliary functions are related by*

$$\dot{P}(t) = A(t)P(t) \tag{2.12}$$

*Proof.* $(i) \implies (ii)$: Define $A(t) = \dot{P}(t)P(t)^{-1}$. Then by Jacobi's formula (Lemma C.1.2), $\mathrm{tr}(A(t)) = (\det(P(t)))^{-1}\frac{\mathrm{d}}{\mathrm{d}t}\det(P(t)) = 0$. We also see using the formula for the

derivative of the inverse (Lemma C.1.1), that $A$ gives the right time derivative for $X$,

$$\dot{X} = \dot{P}X(0)P^{-1} - PX(0)P^{-1}\dot{P}P^{-1} = [A, X]\,.$$

$(ii) \implies (i)$: Define $P(t)$ as the solution to the initial value problem

$$\dot{P}(t) = A(t)P(t), \quad P(0) = I.$$

Then $P(0)X_0P(0)^{-1} = IX_0I = X(0)$, and for all $t \in \mathbb{R}$, using the formula for the derivative of the inverse (Lemma C.1.1),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}P(t)X_0P(t)^{-1} &= \frac{\mathrm{d}P(t)}{\mathrm{d}t}X_0P(t)^{-1} + P(t)X_0\frac{\mathrm{d}}{\mathrm{d}t}\left(P(t)^{-1}\right)\\ &= A(t)P(t)X_0P(t)^{-1} - P(t)X_0P(t)^{-1}A(t)P(t)P(t)^{-1}\\ &= \left[A(t), P(t)X_0P(t)^{-1}\right].\end{aligned}$$

By uniqueness of the solution to the initial value problem (2.11), $X(t) = P(t)X_0P(t)^{-1}$. □

**Definition 2.1.2** (Isospectrality)**.** A $C^1$ flow $X : \mathbb{R} \to \mathbb{C}^{n\times n}$ is *isospectral* if there exists a $C^1$ function $P : \mathbb{R} \to \mathrm{SL}(n)$ which parametrises a similarity transformation for $X$,

$$X(t) = P(t)X(0)P(t)^{-1} \text{ for all } t \in \mathbb{R}.$$

*Remark* 2.1.3. Note that the existence of this similarity transformation is not implied just by the eigenvalues remaining fixed. Consider the $C^1$ function

$$X(t) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}.$$

The eigenvalues are 1 and 1 for all $t$, but there is no similarity transformation describing $X$ from $X(0)$ (because $X(0)$ is the identity). The extra property of isospectral flows we have imposed is that the algebraic *and* geometric multiplicities of the eigenvalues must be preserved. Equivalently, the (unique) Jordan decomposition of the matrix is preserved by the flow.

*Remark* 2.1.4. Isospectral flows are sometimes written in the form

$$\dot{Z}(t) = [Z(t), C(t)]\,. \tag{2.13}$$

These are equivalent to the above mentioned form, but the similarity structure of these flows is slightly different:

$$Z(t) = R(t)^{-1}Z(0)R(t), \quad C(t) = R(t)^{-1}\dot{R}(t). \tag{2.14}$$

We use the following definition for functions of matrices, which is consistent with that in [Hig08].

**Definition 2.1.5.** We say that a function $f : G \subseteq \mathbb{C} \to \mathbb{C}$ is a *matrix function* for the matrix $A \in \mathbb{C}^{n\times n}$ if the values

$$f^{(j)}(\lambda_i), \quad i = 1, 2, \ldots, s, \quad j = 0, 1, \ldots, n_i - 1, \tag{2.15}$$

exist. Here $\lambda_1, \ldots, \lambda_s$ are the eigenvalues of $A$, each with index $n_i$ in their Jordan normal form (see [Hig08, pp. 2–3]).

*Remark* 2.1.6. Note that polynomials are matrix functions for all matrices. Conversely a matrix function applied to a specific matrix is equal to some polynomial applied to the matrix [Hig08].

**Proposition 2.1.7** (See [HM94])**.** *Let $f$ be a matrix function for $X_0 \in \mathbb{C}^{n\times n}$ (see Definition 2.1.5). Suppose that $X : \mathbb{R} \to \mathbb{C}^{n\times n}$ satisfies the isospectral flow $\dot{X}(t) = [A(t), X(t)]$, $X(0) = X_0$. Then $f(X)$ satisfies the isospectral flow*

$$\frac{\mathrm{d}}{\mathrm{d}t} f(X(t)) = [A(t), f(X(t))], \quad f(X(0)) = f(X_0). \tag{2.16}$$

*Proof.* Let $P(t)$ be the auxiliary flow as defined in Proposition 2.1.1, so that $X(t) = P(t)X_0P(t)^{-1}$ and $\dot{P}(t) = A(t)P(t)$. Then by one of the main properties of matrix functions, we have that $f(X(t)) = P(t)f(X_0)P(t)^{-1}$ [Hig08], so using the formula for the derivative of the inverse (Lemma C.1.1), we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} f(X(t)) &= \dot{P}(t)f(X_0)P(t)^{-1} + P(t)f(X_0)\frac{\mathrm{d}}{\mathrm{d}t}P(t)^{-1} \\
&= A(t)P(t)f(X_0)P(t)^{-1} - P(t)f(X_0)P(t)^{-1}A(t)P(t)P(t)^{-1} \\
&= [A(t), f(X(t))].
\end{aligned}$$

Noting $f(X(0)) = f(X_0)$ completes the proof. □

### 2.1.1 Symmetric isospectral flows

Here we discuss real, symmetric isospectral flows since they are an oft-considered special case [HM94], [DNT83], [Bro91], [BI06].

**Proposition 2.1.8** (See [HM94]). *Let $B : \mathbb{R} \to \mathfrak{so}(n)$ be continuous and $Y_0 \in \mathbb{R}^{n\times n}_{\text{sym}}$. Then the unique solution to*

$$\dot{Y}(t) = [B(t), Y(t)], \quad Y(0) = Y_0, \tag{2.17}$$

*lies in $\mathbb{R}^{n\times n}_{\text{sym}}$ for all $t$. Furthermore, the auxiliary flow in Proposition 2.1.1,*

$$\dot{Q}(t) = B(t)Q(t), \quad Q(0) = I, \tag{2.18}$$

*evolves in $\mathrm{SO}(n)$.*

*Proof.* By Proposition 2.1.1, we may write $Y(t) = Q(t)Y(0)Q(t)^{-1}$, where $\dot{Q}(t) = B(t)Q(t)$, $Q(0) = I$. Note that $Q(0) = I$ is orthogonal, and for all $t$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(QQ^T) &= \dot{Q}Q^T + Q\dot{Q}^T \\ &= BQQ^T + QQ^TB^T \\ &= B + B^T \\ &= 0.\end{aligned}$$

Hence $Q(t)$ is orthogonal and $Y(t) = Q(t)Y(0)Q(t)^T \in \mathbb{R}^{n\times n}_{\text{sym}}$ for all $t$. The determinant of an orthogonal matrix is $\pm 1$, so since $\det(Q(0)) = 1$ and the flow is continuous, $Q$ must lie in $\mathrm{SO}(n)$ for all $t$. □

One reason to study symmetric isospectral flows is that their fixed points are easy to characterise, at least under the assumption that $Y$ has distinct eigenvalues. This facilitates the design of the flow to meet our computational desires.

**Lemma 2.1.9.** *Let $D \in \mathrm{Diag}(n)$ with distinct diagonal entries. Then for any $N \in \mathbb{C}^{n\times n}$,*

$$[N, D] = 0 \iff N \in \mathrm{Diag}(n).$$

*Proof.* This follows from the explicit calculation: $[N, D]_{ij} = n_{ij}(d_{jj} - d_{ii})$. □

**Proposition 2.1.10.** *Let $Y \in \mathbb{R}^{n\times n}_{\text{sym}}$ have distinct eigenvalues and $B \in \mathfrak{so}(n)$. Then*

$$[B, Y] = 0 \iff B = 0$$

*Proof.* Since $Y$ is symmetric, we have the factorisation $Y = QDQ^T$, where $D \in \text{Diag}(n)$ $Q \in \text{SO}(n)$. Since $B$ is skew-symmetric, we can factor it into $B = QAQ^T$, where $A = Q^T BQ$ is a skew-symmetric matrix. Then we have the following:

$$[B, Y] = 0 \iff [A, D] = 0 \iff A \in \text{Diag}(n) \iff A = 0 \iff B = 0,$$

by Lemma 2.1.9. □

By the above proposition, we see that if $Y_0$ has distinct eigenvalues, the fixed points of an autonomous isospectral flow are precisely the matrices similar to $Y_0$ such that $B(Y) = 0$. This is a useful fact for the design of algorithms involving isospectral flows, since we know that the flow converges if and only if the condition $B(Y) = 0$ is satisfied in the limit. Despite this, we have no reason in general to expect an isospectral flow to converge to a fixed point $Y^*$ simply because $B(Y^*) = 0$, even if it is the unique matrix similar to $Y_0$ with this property. It is some special property of $B$ which guarantees convergence over an alternative scenario such as a periodic orbit.

### 2.1.2 Normal isospectral flows

A matrix $A \in \mathbb{C}^{n\times n}$ is said to be *normal* if $AA^H = A^H A$. Here we discuss isospectral flows of normal matrices because we will show later that normal matrices have special behaviour in the QR algorithm.

**Proposition 2.1.11.** *Let $A : \mathbb{R} \to \mathfrak{su}(n)$ be continuous and $X_0 \in \mathbb{C}^{n\times n}$ be normal. Then the solution to*

$$\dot{X} = [A, X], \quad X(0) = X_0 \tag{2.19}$$

*is normal for all $t$. Further, the auxiliary flow $\dot{Q} = AQ$ evolves in the Lie group $\text{SU}(n)$.*

*Proof.* The argument that $Q$ is unitary is identical to that used in Proposition 2.1.8. The determinant can take any value on the unit circle here though, so we must show that the determinant is constant. Using the Jacobi formula (Lemma C.1.2),

$$\frac{\mathrm{d}}{\mathrm{d}t}\det(Q) = (\det(Q))\text{tr}(\dot{Q}Q^H) = (\det(Q))\text{tr}(A) = 0.$$

Hence $\det(Q(t)) = \det(Q(0)) = 1$ for all $t$, so $Q$ evolves in $\mathrm{SU}(n)$. It is now simple to show that $X$ is normal for all $t$:

$$\begin{aligned}\left[X, X^H\right] &= \left[QX_0Q^H, QX_0^HQ^H\right] \\ &= Q\left[X_0, X_0^H\right]Q^H \\ &= 0.\end{aligned}$$

This completes the proof. □

*Remark* 2.1.12. Note that the proof of Proposition 2.1.11 also shows that for nonnormal matrices $X_0$, and continuous flows $A$ evolving in $\mathfrak{su}(n)$, the nonnormality of $X$ as measured by a unitary-invariant norm $\|\cdot\|$ such as the 2-norm or Frobenius norm is preserved:

$$\left\|\left[X, X^H\right]\right\|_F = \left\|Q\left[X_0, X_0^H\right]Q^H\right\|_F = \left\|\left[X_0, X_0^H\right]\right\|_F.$$

## 2.2 The QR algorithm and isospectral flows

Isospectral flows came to the attention of numerical analysts in the 1980's after Symes showed that the famous QR algorithm was interpolated by an isospectral flow at integer values of $t$ [Sym82], [DNT83], [Wat84]. In this section we discuss the QR algorithm and isospectral flows that are associated to it: Toda flows, double bracket flows, and their generalisations.

### 2.2.1 The QR algorithm

Let $X \in \mathbb{C}^{n\times n}$. Then there exists $Q \in \mathrm{SU}(n)$ and an upper triangular matrix $R$ such that

$$X = QR. \tag{2.20}$$

This factorisation is known as a *QR factorisation*. The factorisation is usually made unique by requiring the diagonal entries of $R$ to be positive, or else the entire row is zero. This is what we will assume throughout.

The QR factorisation is a fundamental concept in numerical linear algebra [TBI97], [Par80], one reason for which is its use in the QR algorithm. The QR algorithm is the go-to algorithm for computing the eigenvalues of a general matrix [Cip00], [Wat08], despite being invented over half a century ago by Francis and Kublanovskaya (independently) [Fra61], [Kub62]. The basic form of the algorithm is as follows.

**Definition 2.2.1** (Basic QR algorithm)**.** The basic QR algorithm starting from $X_0 \in \mathbb{C}^{n\times n}$ generates the following sequences of matrices:

1. $X^{(0)} \leftarrow X_0$
2. **for** $k = 1, 2, \ldots$ **do**
3. Compute the QR factorisation: $Q^{(k)}R^{(k)} = X^{(k-1)}$
4. $X^{(k)} \leftarrow R^{(k)}Q^{(k)}$

We simply compute the QR factorisation, multiply $R$ by $Q$, and repeat. The first thing to note is that the iterates are all similar:

$$X^{(k)} = R^{(k)}Q^{(k)} = Q^{(k)H}Q^{(k)}R^{(k)}Q^{(k)} = Q^{(k)H}X^{(k-1)}Q^{(k)}, \tag{2.21}$$

so the QR algorithm is a discrete isospectral flow.

How does such a simple algorithm compute eigenvalues? It is easy to see by the uniqueness of the QR factorisation that upper triangular matrices are fixed points of the iteration. If, for example, the eigenvalues of a Hermitian matrix $X^{(0)} \in \mathbb{C}^{n\times n}$ have distinct absolute values and we write

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|,$$

then the elements of $X^{(k)}$ generated by the QR algorithm satisfy $x_{ij}^{(k)} = \mathcal{O}(|\lambda_j/\lambda_i|^k)$ [Wil65], [Wat07]. Hence not only are upper triangular matrices fixed points of the iteration, but $X^{(k)}$ *converges* to an upper triangular matrix as $k \to \infty$. More generally, when $X^{(0)}$ is normal, $R^{(k)}$ converges to a diagonal matrix as $k \to \infty$ [EH75]. There is no "magic" global convergence result for all matrices, particularly when $X^{(0)}$ is nonnormal. We will discuss this later in the chapter.

The QR algorithm is rarely applied to a full matrix. In practice it is first reduced to upper Hessenberg form by Householder's algorithm in $\mathcal{O}(n^2)$ operations. This approach takes advantage of the fact that the QR algorithm preserves the lower bandwidth, which follows from the fact that the similarity transformations can be viewed as coming from triangular matrices as well as unitary matrices,

$$X^{(k)} = R^{(k)}Q^{(k)} = R^{(k)}Q^{(k)}R^{(k)}(R^{(k)})^{-1} = R^{(k)}X^{(k)}(R^{(k)})^{-1}. \tag{2.22}$$

Reduction to Hessenberg form also reduces the number of operations and storage required to compute the QR factorisations. We will not dwell upon the efficient, optimised implementation of algorithms in this thesis.

Table 2.1 Commonly used shifting strategies for the QR algorithm.

| $\sigma(z)$ | Name | Notes |
|---|---|---|
| $z$ | Non shift | The basic QR algorithm |
| $z-\eta$ | Linear shift | Most common type of shift |
| $(z-\eta)(z-\overline{\eta})$ | Francis double shift | Two linear shifts done in succession |
| $\exp(z)$ | Toda shift | Iterates interpolate the Toda flow |

In practice the QR algorithm is also modified by including *shifts*. We take shifts to be matrix functions for $X_0$ as in Definition 2.1.5. This departs from what is usually defined in the literature, as will be explained below in Remark 2.2.3.

**Definition 2.2.2** (Shifted QR algorithm)**.** The shifted QR algorithm starting from $X_0 \in \mathbb{C}^{n\times n}$ with shifts $\sigma_1, \sigma_2, \ldots$ generates the following sequences of matrices.

1. $X^{(0)} \leftarrow X_0$
2. **for** $k = 1, 2, \ldots$ **do**
3. Compute the QR factorisation: $Q^{(k)}R^{(k)} = \sigma_k(X^{(k-1)})$
4. $X^{(k)} \leftarrow (Q^{(k)})^H X^{(k-1)} Q^{(k)}$

*Remark* 2.2.3. Shifts are usually taken to be a complex number. The approach taken here in which a shift is a matrix function generalises the notion of a shift and allows us to include things like double shifts and Toda shifts given in Table 2.1.

The shifted QR algorithm is clearly isospectral, and just as in the basic case, the lower bandwidth is preserved because the similarity transformation is also performed by an upper triangular matrix:

$$\begin{aligned} X^{(k)} &= (Q^{(k)})^H X^{(k-1)} Q^{(k)} \\ &= R^{(k)}\sigma_k(X^{(k-1)})^{-1} X^{(k-1)} \sigma_k(X^{(k-1)})(R^{(k)})^{-1} \\ &= R^{(k)} X^{(k-1)} \left(R^{(k)}\right)^{-1}. \end{aligned} \tag{2.23}$$

The similarity transformation by upper triangular matrices shown in equations (2.22) and (2.23) will only work if $\sigma_k(X^{k-1})$ is nonsingular because we took an inverse. Hence we make the following definition.

**Definition 2.2.4** (Regular shifts)**.** For a matrix $X_0 \in \mathbb{C}^{n\times n}$, a shift $\sigma$ is *regular* if $\sigma(X)$ is nonsingular.

What happens if we use an *irregular* shift?

**Proposition 2.2.5** (Perfect shifting)**.** *Let $X_0 \in \mathbb{C}^{n\times n}$, and let $\sigma$ be a shift for $X_0$ such that $\sigma(X_0)$ is nonsingular with kernel of dimension $k$. Then one iteration of the shifted QR algorithm with shift $\sigma$ makes the last $k$ rows of $\sigma(X^{(1)})$ zero.*

*Proof.* Since the dimension of the kernel is invariant under the Hermitian transpose, $\sigma(X_0)^H$ also has a $k$-dimensional kernel. Let $Q_1 \in \mathbb{C}^{n\times n}$ be a unitary matrix whose first $k$ columns form an orthonormal basis for the kernel of $\sigma(X_0)^H$, and such that the last $n-k$ columns form an orthonormal basis for the orthogonal complement. Then $Q_1^H\sigma(X_0)$ has its final $k$ rows zero. Hence we can partition $Q_1^H\sigma(X_0)$ into

$$Q_1^H\sigma(X_0) = \begin{pmatrix} A_{11} & A_{12} \\ 0_{k\times n-k} & 0_{k\times k} \end{pmatrix}$$

If $A_{11} = Q_2R_2$ is the unique QR factorisation then the orthogonal matrix

$$Q = Q_1 \begin{pmatrix} Q_2 & 0_{n-k\times k} \\ 0_{k\times n-k} & I_{k\times k} \end{pmatrix}$$

is such that the last $k$ rows of $Q^H\sigma(X_0)$ are all zero. Hence the last $k$ rows of $\sigma(X^{(1)}) = Q^H\sigma(X_0)Q$ are also all zero. □

**Corollary 2.2.6.** *Let $X_0 \in \mathbb{C}^{n\times n}$ have an eigenvalue $\lambda \in \mathbb{C}$ of geometric multiplicity $k$. Then one iteration of the shifted QR algorithm with shift $\sigma(z) = z - \lambda$ makes the last $k$ rows of $X^{(1)}$ zero, except the diagonal entries, all of which are $\lambda$.*

*Proof.* Since $\lambda$ is an eigenvalue of $X_0$ of geometric multiplicity $k$, $\sigma(X)$ is nonsingular with a $k$ dimensional kernel. Hence by Proposition 2.2.5 the last $k$ rows of $\sigma(X^{(1)})$ are zero. Adding $\lambda I$ to this gives the required result. □

*Remark* 2.2.7. Corollary 2.2.6 shows that if we use a shift $\sigma(z) = z - \lambda$ where $\lambda$ is an eigenvalue, then the eigenvalue and its multiplicity become explicit in the matrix after one step of the shifted QR algorithm. This may seem useless at first, because we would need to know the eigenvalue with which to shift *before* using the shift to make the eigenvalue explicit. However, in practice the result also holds in an approximate sense. By this we mean that if we use a shift which is sufficiently close to an eigenvalue of $X_0$, then the entries which in Corollary 2.2.6 are zero after one step, in this case converge to zero rapidly.

### 2.2.2 Toda flow

The Toda lattice was discussed in Subsection 1.1.4. It was in 1974 that Flaschka demonstrated the Toda flow [Fla74] and in the 1980s it was later found that the Toda flow is connected to the QR algorithm [Mos75], [Sym82], [DNT83],[Wat84], [Kos79]. Symes proved the following theorem.

**Theorem 2.2.8** (Symes [Sym82]). *Let $Y(t)$ be the solution to the Toda lattice isospectral flow with initial condition $Y(0) = Y_0$, and let $Y^{(0)}, Y^{(1)}, Y^{(2)}, \ldots$ be the iterates of the basic QR algorithm with initial condition $\exp(Y_0)$. Then $Y(k) = Y^{(k)}$ for $k = 0, 1, 2, \ldots$.*

*Proof.* We will prove this in more generality in subsection 2.2.4. □

We will explore Toda flows further in Section 2.4.

### 2.2.3 Double bracket flow

The following isospectral flow is Brockett's double bracket flow [Bro91].

$$\dot{Y} = [[S, Y], Y], \quad Y_0 \in \mathbb{R}^{n\times n}_{\text{sym}}, \quad S \in \mathbb{R}^{n\times n}_{\text{sym}}. \tag{2.24}$$

Note that since $\|Y\|_F^2 = \sum_i \lambda_i^2 = \|Y_0\|_F^2$ for all $t$, we have:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|Y - S\|_F^2 &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\|Y\|_F^2 + \frac{1}{2}\|S\|^2 - \langle Y, S\rangle_F\right) \\
&= -\left\langle \dot{Y}, S\right\rangle_F \\
&= -\langle [[S, Y], Y], S\rangle_F \\
&= -\langle [S, Y], [S, Y]\rangle_F \quad \text{(Lemma C.2.1)} \\
&= -\| [S, Y] \|_F^2 \\
&\leq 0.
\end{aligned}$$

The Frobenius norm of the difference between $Y$ and $S$ is non-increasing, and since $\frac{1}{2}\|Y - S\|_F^2 \geq 0$ for all $t$, we conclude by the Monotone Convergence Theorem that $\frac{1}{2}\|Y - S\|_F^2$ converges as $t \to \infty$. Therefore $\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|Y - S\|_F^2 \to 0$ as $t \to \infty$, and so we have the following condition at infinity:

$$\lim_{t\to\infty} [S, Y(t)] = 0. \tag{2.25}$$

Now, if $S$ is a diagonal matrix with distinct eigenvalues, then by Lemma 2.1.9, $Y(t)$ converges to a diagonal matrix as $t \to \infty$. In fact, if the diagonal entries of $S$ are in increasing order then the only *stable* stationary point also has its diagonal entries in increasing order [Bro91]. This implies that the double-bracket flow is a continuous sorting process. It also possible to choose $S$ and $Y_0$ so that the system solves a linear programming problem on a convex polytope, but this is more subtle [Bro91].

If $S$ is not a diagonal matrix, then since it is symmetric we can diagonalise it orthogonally into $S = Q^T \tilde{S} Q$. Then the matrix $\tilde{Y} = Q^T Y Q$ satisfies:

$$\dot{\tilde{Y}} = [[\tilde{S}, \tilde{Y}], \tilde{Y}], \quad \tilde{Y}(0) = Q^T Y_0 Q, \tag{2.26}$$

so the above analysis applies. As long as $S$ or $Y_0$ has distinct eigenvalues, the double-bracket flow will converge to the best isospectral approximation to $S$ in Frobenius norm.

In the tridiagonal case, the Toda flow and the double bracket flow are identical [Blo90], [BG98].

**Lemma 2.2.9** (Bloch [Blo90]). *Let $S = \operatorname{diag}(n, n-1, \ldots, 1)$ and let $Y_0 \in \mathbb{R}^{n\times n}_{\text{sym}}$ be symmetric tridiagonal. Then the double bracket flow (2.24) is equal to the Toda flow.*

*Proof.* Since $S$ is diagonal, we have

$$[S, Y]_{ij} = (s_i - s_j) y_{ij} = (j - i) y_{ij}.$$

Hence if $Y$ is tridiagonal, $[S, Y] = Y_U - Y_L$. Since the double bracket flow preserves the tridiagonal structure the proof is complete [ADH97]. □

The connection between the Toda flow and the double bracket flow shows that the double bracket flow is also related to the QR algorithm. We will see more on this later.

### 2.2.4 QR flow

Here we follow [Wat84]. The QR factorisation for matrices in the general linear Lie group $\mathrm{GL}(n, \mathbb{C})$ is linked to the following decomposition of the general linear Lie algebra.

$$\mathfrak{gl}(n, \mathbb{C}) = \mathfrak{su}(n) \oplus \mathfrak{supp}(n, \mathbb{C}). \tag{2.27}$$

Here $\mathfrak{supp}(n, \mathbb{C})$ is the Lie algebra of upper triangular matrices with real diagonal entries. For a matrix $Y \in \mathbb{C}^{n\times n}$, let $Y_L, Y_U, Y_D$ denote the strictly lower triangular,

strictly upper triangular, and diagonal parts of $Y$ respectively. Define the linear projections

$$\pi_1(Y) = Y_L - Y_L^H + i\mathrm{Im}(Y_D) \tag{2.28}$$
$$\pi_2(Y) = Y_U + Y_L^H + \mathrm{Re}(Y_D), \tag{2.29}$$

which project onto $\mathfrak{su}(n)$ and $\mathfrak{supp}(n, \mathbb{C})$ respectively.

**Definition 2.2.10** ($f$-QR flow)**.** Let $X_0 \in \mathbb{C}^{n\times n}$ and $f$ be a matrix function for $X_0$ (see Definition 2.1.5). The $f$-QR flow is

$$\dot{X} = [X, \pi_1(f(X))] = [\pi_2(f(X)), X], \quad X(0) = X_0, \tag{2.30}$$

where $\pi_1$ and $\pi_2$ are as defined above.

We may write $X(t) = Q(t)^H X_0 Q(t) = R(t) X_0 R(t)^{-1}$ where

$$\dot{Q} = Q\pi_1(f(X)), \quad Q(0) = I \tag{2.31}$$
$$\dot{R} = \pi_2(f(X))R, \quad R(0) = I. \tag{2.32}$$

Since $\pi_1(X)$ always lies in the Lie algebra of skew-Hermitian matrices $\mathfrak{su}(n)$ and $\pi_2(X)$ always lie in the Lie algebra of upper-triangular matrices with real diagonal entries $\mathfrak{supp}(n, \mathbb{C})$, $Q$ and $R$ lie in the Lie groups $\mathrm{SU}(n)$ and $\mathrm{SUpp}(n)$ of unitary and upper triangular matrices with positive diagonal respectively.

**Theorem 2.2.11** (Symes [Sym82]. See also [Wat84])**.** *The $f$-QR flow satisfies the following*

$$\exp(tf(X_0)) = Q(t)R(t) \tag{2.33}$$
$$\exp(tf(X(t))) = R(t)Q(t). \tag{2.34}$$

*Hence* $\exp(f(X(k))) = X_k$ *where* $X_k$ *is generated by the QR algorithm starting from* $X_0$ *with shifts* $\sigma_k(z) = \exp(f(z))$.

*Proof.* Let us deal with the first assertion:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}QR &= Q\pi_1(f(X))R + Q\pi_2(f(X))R \\ &= Qf(X)R\end{aligned}$$

$$= f(X_0)QR.$$

By uniqueness of solutions to analytic initial value problems, we have that $\exp(tf(X_0)) = Q(t)R(t)$. The second assertion follows because for $t > 0$, the map $z \mapsto z^t$ is analytic on the eigenvalues of $X_0$. Thus,

$$X(t)^t = Q(t)^T X_0^t Q(t) = R(t)Q(t).$$

Now, consider the first iteration, $k = 0$. We have $X(0) = Q(1)R(1)$ and $X(1) = R(1)Q(1)$. Since the flow is autonomous, the flow starting from $X(1)$ is exactly the same as the original flow for $t \geq 1$. Therefore the QR flow gives the iterates of the QR algorithm, when sampled at unit intervals. □

We will explore this isospectral flow in much more detail in Section 2.4.

## 2.3 Bloch–Iserles flow

### 2.3.1 Bloch–Iserles flow

Bloch and Iserles introduced an isospectral flow that has a Hamiltonian structure and a Lie–Poisson structure [BI06]. Let $Y_0 \in \mathbb{R}^{n\times n}_{\text{sym}}$ and $N \in \mathfrak{so}(n)$. The Bloch–Iserles flow is as follows.

$$\dot{Y} = \left[N, Y^2\right] = \left[NY + YN, Y\right], \quad Y(0) = Y_0. \tag{2.35}$$

If we write $H_3(Y) = \operatorname{tr}(Y^3)$ and $\mathcal{J}_3 = \operatorname{ad}_N$, then $\mathcal{J}_3$ induces a Poisson bracket for functions $f, g : \mathbb{R}^{n\times n}_{\text{sym}} \to \mathbb{R}$ given by

$$\{f, g\}_{\mathcal{J}_3} = \langle \nabla f, \mathcal{J}_3 \nabla g \rangle_F, \tag{2.36}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. Using this, the Bloch–Iserles equation can be viewed as a Hamiltonian system $\dot{Y} = \mathcal{J}_3 \nabla H_3(Y)$.

Also, writing $H_2(Y) = \operatorname{tr}(Y^2)$ and $\mathcal{J}_2 = \operatorname{ad}_{YN+NY}$, another Poisson bracket is induced in the same way and we can write the system as $\dot{Y} = \mathcal{J}_2 \nabla H_2(Y)$. Since the operator $\mathcal{J}_2$ depends homogeneously on $Y$, writing the system in this forms shows that it has Lie–Poisson structure [BI06].

As such, this system has very different behaviour to the double bracket system, despite the vector fields appearing very similar. The Bloch-Iserles system is a completely integrable Hamiltonian system ([BBI+09]) with bounded trajectory (since $\|Y\|_F^2$ is an integral of motion), hence by the Arnold–Liouville Theorem its dynamics are diffeomorphic to inertial motion on a torus. On the other hand the double bracket flow is a gradient flow, which converges to a fixed point as $t \to \infty$. This is demonstrated in Figure 2.1. We compute numerical solutions to the Bloch-Iserles flow and the double bracket flow with

$$N = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \tag{2.37}$$

and initial datum

$$Y_0 = \begin{pmatrix} -5 & 4 & 1 & -1 \\ 4 & -1 & 0 & -2 \\ 1 & 0 & 2 & 3 \\ -1 & -2 & 3 & 0 \end{pmatrix}. \tag{2.38}$$

The numerical method used was a Runge–Kutta–Munthe-Kaas forward Euler method on the isospectral manifold using the Cayley map (see Appendix B for the Cayley map) [Zan98], [CIZ97], [IMKNZ00]. The specific scheme is

$$Y_{k+1} = (I + \frac{1}{2}B(Y_k))^{-1}(I - \frac{1}{2}B(Y_k))Y_k(I - \frac{1}{2}B(Y_k))^{-1}(I + \frac{1}{2}B(Y_k)), \tag{2.39}$$

where $B(Y) = NY + YN$ and $B(Y) = SY - YS$ for the Bloch-Iserles and double bracket flows respectively. Despite the two vector fields being both of the form $\dot{Y} = [VY - YV^T, Y]$ for a matrix $V$, their flows have very different behaviours: the Bloch-Iserles flow is oscillatory whereas the double bracket flow is convergent.

### 2.3.2 KdV is a modified Bloch-Iserles system

Consider the Bloch-Iserles system [BI06] with an added linear term:

$$\dot{X} = [XN + NX + M, X], \quad X(0) = X_0. \tag{2.40}$$

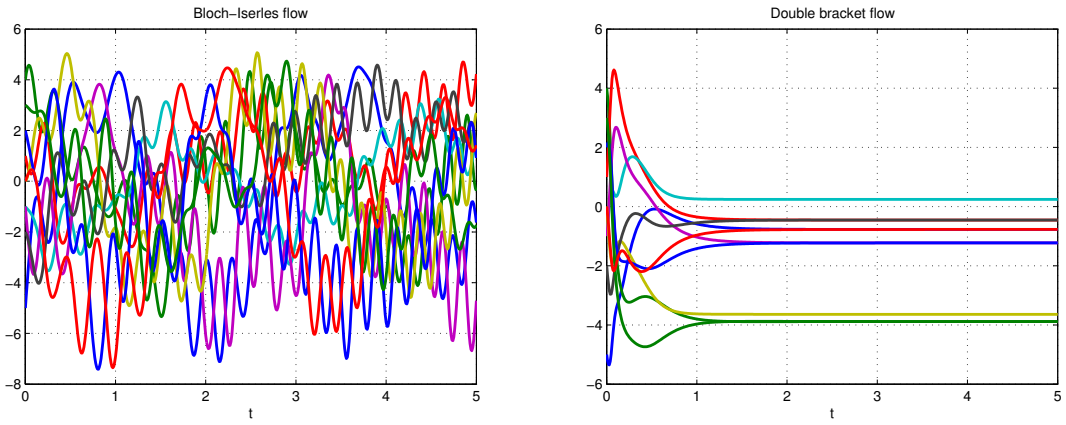



Fig. 2.1 The left image is a trajectory of the Bloch-Iserles flow $\dot{Y} = [NY + YN, Y]$, $Y(0) = Y_0$ and the right image is a trajectory of a double bracket flow $\dot{Y} = [SY - YS, Y]$, $Y(0) = Y_0$, where $N$ and $S$ are given in equation (2.37) and $Y_0$ is given in equation (2.38). Each coloured line represents the trajectory of a single entry of of the matrix. Despite the two vector fields being both of the form $\dot{Y} = [VY - YV^T, Y]$ for a matrix $V$, their flows have very different behaviours: the Bloch-Iserles flow is oscillatory whereas the double bracket flow is convergent.

Here $N$ and $M$ are skew-selfadjoint linear operators and $X_0$ is selfadjoint. Let us call this a *modified Bloch-Iserles system*. If we perform the transformation $Y = e^{-tM} X e^{tM}$, then

$$\begin{aligned}\dot{Y} &= -[M, Y] + e^{-tM}[XN + NX + M, X]e^{tM} \\ &= -[M, Y] + [Ye^{-tM}Ne^{tM} + e^{-tM}Ne^{tM}Y + M, Y] \\ &= [Ye^{-tM}Ne^{tM} + e^{-tM}Ne^{tM}Y, Y].\end{aligned}$$

So $Y$ solves a kind of time-dependent Bloch-Iserles system. Note that if $[M, N] = 0$ then the solution to (2.40) is $X = e^{tM} Y e^{-tM}$ where $Y$ solves the actual Bloch-Iserles system $\dot{Y} = [YN + NY, Y]$.

The KdV equation described in Subsection 1.1.4 can be written using the operators

$$L(t) = -\partial_x^2 + M_{u(t,\cdot)},$$

$$A(t) = -4\partial_x^3 + 3\left(M_{u(t,\cdot)}\partial_x + \partial_x M_{u(t,\cdot)}\right),$$

in the form of an isospectral flow [Lax68],

$$L_t - [A, L] = M_{u_t - 6uu_x + u_{xxx}} = 0.$$

**Lemma 2.3.1.** *The operator $A$ can be written as a function of $L$ and $\partial_x$,*

$$A = 3\left(\partial_x L + L\partial_x\right) + 2\partial_x^3. \tag{2.41}$$

*Proof.* We simply expand the right hand side using $L = -\partial_x^2 + M_{u(\cdot,t)}$ like so.

$$\begin{aligned} 3\left(\partial_x L + L\partial_x\right) + 2\partial_x^3 &= -3\partial_x^3 + 3\partial_x M_u - 3\partial_x^3 + 3M_u\partial_x \\ &\quad + 2\partial_x^3 \\ &= -4\partial_x^3 + 3\left(M_{u(t,\cdot)}\partial_x + \partial_x M_{u(t,\cdot)}\right), \end{aligned}$$

as required. □

The above Lemma shows that the isospectral flow for the KdV is an infinite dimensional modified Bloch-Iserles system with $N = 3\partial_x$ and $M = 2\partial_x^3$.

$$L_t = \left[3(\partial_x L + L\partial_x) + 2\partial_x^3, L\right]. \tag{2.42}$$

Since $\partial_x$ and $\partial_x^3$ commute, the solution can be written in terms of the solution to a Bloch-Iserles system:

$$L(t) = e^{2t\partial_x^3} K(3t) e^{-2t\partial_x^3}, \tag{2.43}$$

$$\dot{K} = [\partial_x K + K\partial_x, K], \quad K(0) = L(0). \tag{2.44}$$

This observation appears to be new. What are the implications? A full answer is beyond the scope of this thesis and requires further work, but here are some ideas. It may be possible to derive new results about KdV from those of the Bloch-Iserles system or vice versa. It is also worth exploring how the KdV hierarchy might correspond to a Bloch-Iserles hierarchy. This could also lead to a new numerical method for solvign the KdV equation as follows.

It was shown in [Kau16] that the solution to a Bloch-Iserles system $\dot{Y} = [NY + YN, Y]$, $Y(0) = Y_0 \in \mathbb{R}^{n\times n}_{\text{sym}}$ can be expressed using a Magnus series expression of the form

$$Y(t) = \exp(\Omega(t)) \tag{2.45}$$

$$\Omega(t) = t\{Y_0\} + \frac{1}{2}t^2\{[\{Y_0\}, Y_0]\} + t^3\left(\frac{1}{6}\{[\{[\{Y_0\}, Y_0]\}, Y_0]\}\right. \tag{2.46}$$

$$\left. + \frac{1}{6}\{[\{Y_0\}, [\{Y_0\}, Y_0]]\} + \frac{1}{12}[\{[\{Y_0\}, Y_0]\}, \{Y_0\}]\right) + \cdots, \tag{2.47}$$

where for any $Y \in \mathbb{R}^{n\times n}_{\text{sym}}$, $\{Y\} = NY + YN$. An interesting point about such an expansion is that the solution is generated purely using the initial datum $Y_0$, $N$, linear combinations, the Lie bracket $[\cdot,\cdot]$ and the bracket $\{\cdot\}$. It is said that the expansion is written using only two "letters", $Y_0$ and $N$. A similar expression was found by Iserles for the double bracket equation and some other Lie-algebraic equations that can be appropriately expressed in terms of a finite "alphabet" [Ise02].

From this it follows that the solution to the KdV equation $L_t = [A, L]$ can be written using an expression which uses only $L(0)$ and $\partial_x$, namely

$$L(t) = \exp(2t\partial_x^3)\exp(\Omega(3t))\exp(-2t\partial_x^3) \tag{2.48}$$

$$\Omega(t) = t\{L(0)\} + \frac{1}{2}t^2\{[\{L(0)\}, L(0)]\} + t^3\left(\frac{1}{6}\{[\{[\{L(0)\}, L(0)]\}, L(0)]\}\right. \tag{2.49}$$

$$\left. +\frac{1}{6}\{[\{L(0)\}, [\{L(0)\}, L(0)]]\} + \frac{1}{12}[\{[\{L(0)\}, L(0)]\}, \{L(0)\}]\right) + \cdots, \tag{2.50}$$

where for a self-adjoint operator $L$ on $L^2(\mathbb{R})$, $\{L\} = \partial_x L + L\partial_x$. Note that convergence issues have not been discussed, so the series can only be said to be formal at present.

## 2.4 Isospectral gradient flows

In this section we show how to design an isospectral flow which computes local minima of an objective function as $t \to \infty$. From here on we will use terminology from Riemannian geometry and Lie theory; a basic introduction and setting of notation is given in Appendix B.

### 2.4.1 The isospectral manifold, or adjoint orbit

Let $\mathcal{G}$ be a Lie subgroup of $\mathrm{GL}(n, \mathbb{C})$ with associated Lie algebra $\mathfrak{g}$. Then for any element $X_0 \in \mathbb{C}^{n\times n}$, the associated *isospectral manifold* is the set

$$\mathcal{I}_{\mathcal{G}}(X_0) = \left\{PX_0P^{-1} : P \in \mathcal{G}\right\}. \tag{2.51}$$

This manifold is sometimes called the *adjoint orbit* for $X_0$ by $\mathcal{G}$, because it is precisely the orbit of $X_0$ under the Adjoint group action $\mathrm{Ad} : \mathcal{G} \times \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$.

**Proposition 2.4.1** (See [HM94]). *The tangent space $T_X\mathcal{I}_\mathcal{G}(X_0)$ at any $X \in \mathcal{I}_\mathcal{G}(X_0)$ is the space*

$$T_X\mathcal{I}_\mathcal{G}(X_0) = \left\{[A, X] : A \in \mathrm{C}_\mathfrak{g}(X)^\perp\right\}, \tag{2.52}$$

*where $\mathrm{C}_\mathfrak{g}(X) = \{B \in \mathfrak{g} : [B, X] = 0\}$ is the centraliser of $X$ in $\mathfrak{g}$ and $\perp$ denotes the orthogonal complement in $\mathfrak{g}$ with respect to the Frobenius inner product.*

*Furthermore, there is a linear bijection,*

$$\mathrm{ad}_X : \mathrm{C}_\mathfrak{g}(X)^\perp \to T_X\mathcal{I}_\mathcal{G}(X_0). \tag{2.53}$$

*Proof.* Let $A \in \mathrm{C}_\mathfrak{g}(X)^\perp$, and define the path $\mu(s) = \exp(sA)X\exp(-sA)$, which lies in $\mathcal{I}_\mathcal{G}(X_0)$ since the exponential takes $\mathfrak{g}$ to $\mathcal{G}$. Then $\mu(0) = X$ and $\dot\mu(0) = [A, X]$. Hence by definition of tangent spaces $[A, X] \in T_X\mathcal{I}_\mathcal{G}(X_0)$.

Now suppose that $\mu$ is a smooth path in $\mathcal{I}_\mathcal{G}(X_0)$ such that $\mu(0) = X$. Then

$$\mu(s) = P(s)XP(s)^{-1}, \quad P(0) = I,$$

for some path $P$ in $\mathcal{G}$. Then we have

$$\begin{aligned}\dot\mu(s)\Big|_{s=0} &= \dot P(s)XP(s)^{-1} - P(s)XP(s)^{-1}\dot P(s)P(s)^{-1}\Big|_{s=0}\\ &= \left[\dot P(s)P(s)^{-1}, P(s)XP(s)^{-1}\right]\Big|_{s=0}\\ &= \left[\dot P(0), X\right].\end{aligned}$$

By Definition B.2.3, $\dot P(0) \in \mathfrak{g}$, so $\dot\mu(0) = [A, X]$ for some $A \in \mathfrak{g}$. We may take $A \in \mathrm{C}_\mathfrak{g}(X)^\perp$ since $[A, X]$ is unaffected by addition of elements of $\mathrm{C}_\mathfrak{g}(X)$ to $A$.

To complete the proof and show that $\mathrm{ad}_X$ is a bijection, note that it is onto and its kernel in $\mathrm{C}_\mathfrak{g}(X)^\perp$ is zero by definition. □

There are a couple of points to take from Proposition 2.4.1. First is that the results of Propositions 2.1.1, 2.1.8 and 2.1.11 are special cases in which $(\mathcal{G}, \mathfrak{g}) = (\mathrm{SL}(n), \mathfrak{sl}(n))$, $(\mathcal{G}, \mathfrak{g}) = \mathrm{SO}(n), \mathfrak{so}(n))$ and $(\mathcal{G}, \mathfrak{g}) = \mathrm{SU}(n), \mathfrak{su}(n))$ respectively. Any isospectral flow of the form

$$\dot X = [A, X], \quad A \in \mathfrak{g},$$

can be written as

$$X = PX(0)P^{-1}, \quad P \in \mathcal{G},$$

where $\mathfrak{g}$ is the Lie algebra associated to $\mathcal{G}$ as in Definition B.2.3, and vice versa. The second point is that the map $\mathrm{ad}_X$ is invertible on the tangent space $T_X\mathcal{I}_\mathcal{G}(X_0)$, so we may unambiguously use the operator

$$\mathrm{ad}_X^{-1} : T_X\mathcal{I}_\mathcal{G}(X_0) \to \mathrm{C}_\mathfrak{g}(X)^\perp. \tag{2.54}$$

### 2.4.2 Metrics and gradient flows

Suppose we are given a continuously differentiable function $\Psi$ from a manifold $\mathcal{M}$ into $\mathbb{R}$, and we want to find a local minimum of it. One approach is to set up a flow on $\mathcal{M}$ which always points in the direction of steepest descent of $\Psi$. This motivates what is known as the gradient descent flow.

**Definition 2.4.2** (Gradient flows)**.** Let $\Psi$ be a $C^1$ function from the Riemannian manifold $(\mathcal{M}, g)$ into $\mathbb{R}$. The gradient (descent) flow of $\Psi$ in $(\mathcal{M}, g)$ is the flow whose trajectories satisfy

$$\dot{X}(t) = -\nabla_g \Psi(X(t)). \tag{2.55}$$

*Remark* 2.4.3. Note that the gradient flow is dependent on the metric chosen for the tangent spaces of $\mathcal{M}$.

The justification for the gradient flow is in the following calculation:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Psi(X(t)) &= \left\langle \nabla_g \Psi(X(t)), \dot{X}(t) \right\rangle_g \\ &= -\left\|\nabla_g \Psi(X(t))\right\|_g^2 = -\|\dot{X}(t)\|_g^2.\end{aligned}$$

We see that $\Psi$ decreases along the trajectory $X$ and that it does so in an optimal way, because the Cauchy-Schwarz inequality implies that given the values of $X(t)$ and $\|\dot{X}(t)\|_g$, the value of $\left\langle \nabla_g \Psi(X(t)), \dot{X}(t) \right\rangle_g$ is minimised if $\dot{X}$ solves (2.55).

**Theorem 2.4.4** (See [HM94])**.** *Let $\Psi : (\mathcal{M}, g) \to \mathbb{R}$ be a $C^1$ function such that the sublevel sets $S_\alpha = \{X \in \mathcal{M} : \Psi(X) \leq \alpha\}$ are compact. Then distance of the trajectories of the gradient flow for $\Psi$ in $(\mathcal{M}, g)$ to the set of critical points of $\Psi$ converges to 0 as $t \to \infty$.*

*Proof.* Let $X(t)$ be the trajectory of the gradient flow for some initial datum. Then $\frac{\mathrm{d}}{\mathrm{d}t}\Psi(X(t)) \leq 0$ for all $t$. Suppose for a contradiction that there exists an $\varepsilon < 0$ such that $\frac{\mathrm{d}}{\mathrm{d}t}\Psi(X(t)) < \varepsilon$ for all $t$. Then $\Psi(X(t)) \to -\infty$ as $t \to \infty$, which contradicts the

fact that $\Psi$ is bounded below (which is an immediate consequence of the compactness of the sublevel sets and continuity of $\Psi$). Hence $\frac{\mathrm{d}}{\mathrm{d}t}\Psi(X(t)) \to 0$ as $t \to \infty$. Since $\frac{\mathrm{d}}{\mathrm{d}t}\Psi(X(t)) = -\left\|\nabla_g \Psi(X(t))\right\|_g^2$, we have that $\nabla_g \Psi(X(t)) \to 0$.

Now, for a contradiction let $t_1, t_2, t_3, \ldots$ be a strictly increasing divergent sequence such that $X(t_n)$ is always at least distance $\varepsilon > 0$ from the set of critical points[1]. Then since $\Psi$ has compact sublevel sets and the flow must remain on a sublevel set for all time, there exists a subsequence $t_{n_1}, t_{n_2}, \ldots$ such that $X(t_{n_k}) \to \tilde{X}$ as $k \to \infty$. Since $\Psi$ is $C^1$ and $\nabla_g \Psi(X(t)) \to 0$ as $t \to \infty$, we must have that $\nabla_g \Psi(\tilde{X}) = 0$. Hence $(X(t_{n_k}))_{k=1}^{\infty}$ converges to the set of critical points of $\Psi$. This is a contradiction. Hence $X(t)$ converges to the set of critical points of $\Psi$ as $t \to \infty$. □

*Remark* 2.4.5. Note that Theorem 2.4.4 does not imply that every solution converges to an equilibrium point, merely that its distance to a possibly continuous set of critical points of $\Psi$ converges to zero. However, if the critical points of $\Psi$ are all isolated points then the flow converges to an equilibrium point [HM94, App. C].

What do gradient flows look like on an isospectral manifold? We must first discuss metrics.

The *Frobenius metric* on the isospectral manifold is the metric induced from the embedding $\mathcal{I}_{\mathcal{G}}(X_0) \subseteq \mathbb{C}^{n^2}$,

$$\langle [A,X],[B,X] \rangle_F = \sum_{i,j} \overline{[A,X]}_{i,j} [B,X]_{i,j}. \tag{2.56}$$

The *normal metric* for the Frobenius metric on the isospectral manifold is the metric inherited from $\mathcal{G}$ as follows. For $A, B \in C_{\mathfrak{g}}(X)^{\perp}$,

$$\langle [A,X],[B,X] \rangle_N = \langle A, B \rangle_F . \tag{2.57}$$

The normal metric is a standard construction for homogeneous manifolds as in Definition B.3.4 [Bro93].

Alternatively, noting as we did in the previous section that the adjoint map $\mathrm{ad}_X$ is invertible on the tangent space $T_X \mathcal{I}_{\mathcal{G}}(X_0)$, we can write $\langle R, S \rangle_N = \left\langle \mathrm{ad}_X^{-1} R, \mathrm{ad}_X^{-1} S \right\rangle_F$.

Now let us consider a general metric on the isospectral manifold. For $A, B \in C_{\mathfrak{g}}(X)^{\perp}$, we can write a general metric $\langle \cdot, \cdot \rangle_g$ in terms of the Frobenius metric by

$$\langle [A,X],[B,X] \rangle_g = \langle A, L_X B \rangle_F , \tag{2.58}$$

[1] The distance referred to here is that of the geodesic distance induced by the metric $g$.

for some selfadjoint positive definite linear operator $L_X \in \mathrm{GL}(C_{\mathfrak{g}}(X)^{\perp})$ depending smoothly on $X$ (otherwise $g$ is not a metric). The normal metric mentioned above has $L_X A = A$ and the Frobenius metric has $L_X A = [[A, X], X^H]$.

We can characterise all the gradient flows in the metric $g$ in terms of the (easier to find) gradient on $\mathbb{C}^{n\times n}$ endowed with the Frobenius metric, denoted $\nabla\Psi$.

**Theorem 2.4.6.** *Let $\Psi : \mathbb{C}^{n\times n} \to \mathbb{R}$ be a $C^1$ function. The gradient of $\Psi$ on the isospectral manifold $\mathcal{I}_{\mathcal{G}}(X_0)$ with respect to the metric $\langle [A, X], [B, X]\rangle_g = \langle A, L_X B\rangle_F$ as defined in (2.58), is the vector field*

$$\nabla_g \Psi(X) = \left[L_X^{-1}\pi_{\mathfrak{g}}\left[\nabla\Psi(X), X^H\right], X\right], \tag{2.59}$$

*where $\nabla\Psi$ is the gradient of $\Psi$ in $\mathbb{C}^{n\times n}$ endowed with the Frobenius metric and $\pi_{\mathfrak{g}}$ is the orthogonal projection onto $\mathfrak{g}$.*

*Hence the trajectories of the isospectral gradient flow for $\Psi$ are described by*

$$\dot{X}(t) = -\left[L_{X(t)}^{-1}\pi_{\mathfrak{g}}\left[\nabla\Psi(X(t)), X(t)^H\right], X(t)\right]. \tag{2.60}$$

*Proof.* Let $X : \mathbb{R} \to \mathcal{I}_{\mathcal{G}}(X_0)$ be a $C^1$ path. By the isomorphism (2.53), $\nabla_g\Psi(X(t)), \dot{X}(t) \in T_{X(t)}\mathcal{I}_{\mathcal{G}}(X_0)$ implies that $\nabla_g\Psi(X(t)) = [A(t), X(t)]$ and $\dot{X}(t) = [B(t), X(t)]$ for some continuous paths $A(t), B(t)$ in $\mathrm{C}_{\mathfrak{g}}(X(t))^{\perp}$.

Now, for any $t$, first using the definition of the metric $g$ and then the bracketed comments, we have

$$\begin{aligned}
\langle A, B\rangle_F &= \left\langle \nabla_g\Psi(X), \left[L_X^{-1}B, X\right]\right\rangle_g \\
&= \left\langle \nabla\Psi(X), \left[L_X^{-1}B, X\right]\right\rangle_F \quad \text{(Definition B.1.6)} \\
&= \left\langle \left[\nabla\Psi(X), X^H\right], L_X^{-1}B\right\rangle_F \quad \text{(Lemma C.2.1)} \\
&= \left\langle \pi_{\mathfrak{g}}\left[\nabla\Psi(X), X^H\right], L_X^{-1}B\right\rangle_F \quad (L_X^{-1}B \in \mathfrak{g}) \\
&= \left\langle L_X^{-1}\pi_{\mathfrak{g}}\left[\nabla\Psi(X), X^H\right], B\right\rangle_F \quad (L_X \text{ is self-adjoint}).
\end{aligned}$$

This last line only makes sense if $\left[\nabla\Psi(X), X^H\right]$ is orthogonal to $C_{\mathfrak{g}}(X)$ because of the domain of definition of $L_X^{-1}$. This is indeed the case, since if $Z \in C_{\mathfrak{g}}(X)$ then by Lemma C.2.1,

$$\left\langle \left[\nabla\Psi(X), X^H\right], Z\right\rangle_F = \langle \nabla\Psi(X), [Z, X]\rangle_F = 0.$$

Finally, since $B \in \mathrm{C}_{\mathfrak{g}}(X)^{\perp}$ can be made arbitrary whilst keeping $A$ fixed, we must have $A = L_X^{-1}\pi_{\mathfrak{g}}\left[\nabla\Psi(X), X^H\right]$, which gives the desired result. □

In the literature, the most common metric used to define an isospectral gradient flow is the normal metric [HM94], [CG05]. Also, it is often real symmetric matrix flows that are used. In this case an isospectral gradient flow for the function $\Psi : \mathbb{R}^{n\times n}_{\text{sym}} \to \mathbb{R}$ is given by

$$\dot{Y}(t) = -\left[\left[\nabla\Psi(Y), Y\right], Y\right]. \tag{2.61}$$

It is so simple because in the normal metric $L_X$ is the identity operator, and $\nabla\Psi(Y), Y \in \mathbb{R}^{n\times n}_{\text{sym}}$, so that $[\nabla\Psi(Y), Y] \in \mathfrak{so}(n)$, which means that the projection $\pi_{\mathfrak{so}}$ is not necessary.

What types of matrices do isospectral gradient flows compute? By Theorem 2.4.4 they converge to the set of stationary points for $\Psi$. The following theorem characterises the stationary points of functions on an isospectral manifold.

**Theorem 2.4.7.** *Let $\Psi$ be defined on the isospectral manifold $\mathcal{I}_{\mathcal{G}}(X_0)$. Then no matter what the metric, the stationary points of the gradient flow in Theorem 2.4.6 satisfy*

$$\pi_{\mathfrak{g}}\left[\nabla\Psi(X), X^H\right] = 0. \tag{2.62}$$

*Proof.* The stationary points of the flow are those points $X$ such that the quantity $\left[L_X^{-1}\pi_{\mathfrak{g}}\left[\nabla\Psi(X), X^H\right], X\right]$ is zero. If $X$ satisfies this, then by definition of the metric $g$ in equation (2.58),

$$\begin{aligned}\left\|\pi_{\mathfrak{g}}\left[\nabla\Psi(X), X^H\right]\right\|_F^2 &= \left\langle\left[\pi_{\mathfrak{g}}\left[\nabla\Psi(X), X^H\right], X\right], \left[L_X^{-1}\pi_{\mathfrak{g}}\left[\nabla\Psi(X), X^H\right], X\right]\right\rangle_g \\ &= 0.\end{aligned}$$

This completes the proof. □

It is also important to characterise *stable* equilibria. This is done via the Hessian on the manifold.

**Theorem 2.4.8.** *Let $\Psi : \mathbb{C}^{n\times n} \to \mathbb{R}$ be a $C^2$ function. The Hessian of $\Psi$ on the isospectral manifold $\mathcal{I}_{\mathcal{G}}(X_0)$ at the point $X$ is the bilinear form on $C_{\mathfrak{g}}(X)^{\perp}$*

$$A \mapsto \left\langle\nabla^2\Psi(X)\left([A, X]\right) + \left[A^H, \nabla\Psi(X)\right], [A, X]\right\rangle_F, \tag{2.63}$$

*where $\nabla\Psi$ and $\nabla^2\Psi$ are the gradient and Hessian of $\Psi$ on $\mathbb{C}^{n\times n}$ in the Frobenius metric.*

*Proof.* By Hadamard's Lemma B.2.9, for any $\varepsilon$ sufficiently small and $A \in C_{\mathfrak{g}}(X)^{\perp}$, we have

$$e^{\varepsilon A} X e^{-\varepsilon A} = X + \varepsilon\,[A, X] + \frac{\varepsilon^2}{2}\,[A, [A, X]] + \mathcal{O}(\varepsilon^3).$$

Hence by Taylor's Theorem,

$$\Psi\left(e^{\varepsilon A}Xe^{-\varepsilon A}\right) = \Psi(X) + \left\langle \nabla\Psi(X), \varepsilon\left[A,X\right] + \frac{\varepsilon^2}{2}\left[A,\left[A,X\right]\right]\right\rangle_F + \frac{1}{2}\left\langle \nabla^2\Psi(X)\left(\varepsilon\left[A,X\right]\right), \varepsilon\left[A,X\right]\right\rangle_F + \mathcal{O}(\varepsilon^3).$$

Apply Lemma C.2.1 to the double commutator and take the quadratic terms in $\varepsilon$ to obtain the result. □

**Corollary 2.4.9.** *A stationary point $X$ of an isospectral gradient flow is stable if and only if*

$$\left\langle \nabla^2\Psi(X)\left([A,X]\right) + \left[A^H, \nabla\Psi(X)\right], [A,X]\right\rangle_F \geq 0 \text{ for all } A \in C_{\mathfrak{g}}(X)^{\perp}. \tag{2.64}$$

*Remark* 2.4.10. Note that Theorem 2.4.7 on stationary points and Corollary 2.4.9 on stable stationary points are completely independent of the metric chosen. This means that we cannot use the metric to change where the stationary points are or whether they are stable or not. The introduction of metrics here, which does not really appear in the literature unless we are in the setting of arbitrary semi-simple Lie algebras [Blo90],[Bro93],[BG98], is not as powerful as it may seem at first.

As an example to demonstrate Theorem 2.4.6, Theorem 2.4.7 and Corollary 2.4.9, consider the following.

**Jacobi eigenvalue flow** The following flow was introduced by Chu and Driessel in 1990 [CD90]. Consider the function $\Psi(X) = \frac{1}{2}\|X - \operatorname{diag}(X)\|_F^2$ for symmetric matrices. By Theorem 2.4.6, the gradient flow on $\mathcal{I}_{\mathrm{SL}(n)}$ is

$$\dot{X} = \left[\left[\operatorname{diag}(X), X\right], X\right]. \tag{2.65}$$

By Theorem 2.4.7, the stationary points of this flow are $X$ such that $[\operatorname{diag}(X), X] = 0$. By Corollary 2.4.9, the stable stationary points are $X$ such that

$$\left\langle [A,X] - \operatorname{diag}\left([A,X]\right) + \left[A^H, X - \operatorname{diag}(X)\right], [A,X]\right\rangle_F \geq 0 \text{ for all } A \in C_{\mathfrak{so}(n)}(X)^{\perp},$$

which is equivalent to the condition

$$\left\langle [A, \operatorname{diag}(X)] - \operatorname{diag}\left([A,X]\right), [A,X]\right\rangle_F \geq 0 \text{ for all } A \in C_{\mathfrak{so}(n)}(X)^{\perp}. \tag{2.66}$$

It is clear that if $X$ is a diagonal matrix then it is a stationary point because $[\operatorname{diag}(X), X] = 0$. Also, these stationary points are stable, because the stability condition becomes $\langle [A, X], [A, X] \rangle_F \geq 0$. Since this is zero if and only if $A = 0$, we have that the stationary point is isolated. Driessel showed that non-diagonal stationary points are unstable [Dri87], [CG05, Ch. 7]. Hence the flow is (almost everywhere) globally convergent to a diagonal matrix. Chu and Golub note the similarities between this flow and the Jacobi eigenvalue algorithm [CG05].

## 2.5 QR flows as gradient flows

In this section we will first show that for symmetric tridiagonal matrices, Toda, double bracket and f-QR flows (with $f(z) = z$) are all isospectral gradient flows for the function $\Psi(X) = \frac{1}{2}\|X - S\|_F^2$ for certain choices of $S$. Then we show how their generalisations are also related to these gradient flows, but with a specific choice of metric on the manifold.

**Proposition 2.5.1** (Double bracket flows)**.** *Let $S \in \mathbb{C}^{n\times n}$. The isospectral gradient flow for the function $\Psi(X) = \frac{1}{2}\|X - S\|_F^2$ on $\mathcal{I}_{\mathrm{GL}(n,\mathbb{C})}(X_0)$ with the normal metric is*

$$\dot{X} = \left[\left[S - X, X^H\right], X\right]. \tag{2.67}$$

*When $X(0)$ and $S$ are Hermitian, the flow is the double bracket flow*

$$\dot{X} = [[S, X], X]. \tag{2.68}$$

*When $X(0)$ and $S$ are skew-Hermitian, the flow is the double bracket flow*

$$\dot{X} = [[-S, X], X]. \tag{2.69}$$

*Proof.* The gradient of $\Psi$ in Euclidean space is $\nabla\Psi(X) = X - S$. The results follow from Theorem 2.4.6. □

**Corollary 2.5.2.** *For symmetric tridiagonal matrices, the Toda flow, double bracket flow with $S = \operatorname{diag}(n, n-1, \ldots, 1)$ and the $f$-QR flow with $f(z) = z$ are equal to the gradient flow for the function $\Psi(X) = \frac{1}{2}\|X - S\|_F^2$ in the normal metric.*

*Proof.* Combine Lemma 2.2.9 with Proposition 2.5.1. □

To better understand the relationships between Toda, double bracket and QR flows for more general matrices, we must make a few definitions and lemmata.

**Definition 2.5.3.** We denote the operator which performs the Hadamard product by a matrix $S \in \mathbb{C}^{n\times n}$ as $H_S$. Explicitly, the operator $H_S$ is defined for $X \in \mathbb{C}^{n\times n}$ as

$$(H_S X)_{ij} = s_{ij} x_{ij}. \tag{2.70}$$

**Lemma 2.5.4** (See [Blo90])**.** *Let $D \in \mathbb{C}^{n\times n}$ be a diagonal matrix and $X \in \mathrm{C}^{n\times n}$. Then*

$$[D, X] = H_{\hat{D}}(X), \tag{2.71}$$

*where $\hat{d}_{ij} = d_i - d_j$.*

**Lemma 2.5.5** ([Blo90])**.** *Let $D = \mathrm{diag}(n, n-1, \ldots, 2, 1)$ and define $S \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ to be the matrix with entries*

$$s_{ij} = |i - j|. \tag{2.72}$$

*Then for a Hermitian matrix $Y$, we have*

$$[D, Y] = -H_S \pi_1(Y), \tag{2.73}$$

*where $H_S$ is the Hadamard product operator in Definition 2.5.3, and $\pi_1$ is the linear operator in the QR flow (see Definition 2.2.10).*

*Proof.* Since $Y$ is Hermitian, we have

$$\pi_1(Y)_{ij} = \begin{cases} y_{ij} & \text{if } i > j \\ 0 & \text{if } i = j \\ -y_{ji} & \text{if } i < j \end{cases} \tag{2.74}$$

By Lemma 2.5.4, $[D, Y] = H_{\hat{D}} Y$, where $\hat{d}_{ij} = n - i - n + j = j - i$. This gives the desired result. □

Let us make the following new definition.

**Definition 2.5.6** (QR metric)**.** Let $S \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ be the matrix with entries $s_{ij} = |i - j|$. The *QR metric* on an isospectral manifold $\mathcal{I}_{\mathrm{SU}(n)}(X_0)$ for any $X_0 \in \mathbb{C}^{n\times n}$ at the point $X \in \mathcal{I}_{\mathrm{SU}(n)}(X_0)$, for any $A, B \in C_{\mathfrak{su}(n)}(X)^{\perp}$, is

$$\langle [A, X], [B, X] \rangle_{QR} = \langle A, H_S B \rangle_F , \tag{2.75}$$

where $H_S$ is the Hadamard operator in Definition 2.5.3 and $S$ has elements $s_{ij} = |i-j|$ for $i \neq j$ and the diagonal entries may be arbitrary positive real numbers.

*Remark* 2.5.7. We will see in the proof of Theorem 2.5.8 why we allowed the diagonal entries of $S$ to be arbitrary positive real numbers. This makes sure that the QR metric is indeed a Riemannian (positive definite) metric, but for the purposes of the gradient QR flows we define momentarily, the diagonal entries of $S$ have no effect.

**Theorem 2.5.8** (Gradient $f$-QR flow)**.** *Define the function* $\Psi(X) = \frac{1}{2}\|f(X) - D\|_F^2$, *where* $f$ *is a matrix function as in Definition 2.1.5 and* $D = \operatorname{diag}(n, n-1, \ldots, 2, 1)$. *The isospectral gradient flow for* $\Psi$ *in the* QR metric *is*

$$\dot{X} = [X, \pi_1(\operatorname{Herm}(f(X)))]\,, \tag{2.76}$$

*where* $\operatorname{Herm}(Y) = \frac{1}{2}(Y + Y^H)$.

*Proof.* The isospectral gradient flow must be of the form $\dot{X}(t) = [X(t), B(t)]$ for some $B(t) \in C_{\mathfrak{su}(n)}(X(t))^\perp$, the orthogonal complement of the centraliser of $X(t)$ in $\mathfrak{su}(n)$. We will show that we may take $B(t) = \pi_1(\operatorname{Herm}(f(X)))$. We compute

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\Psi(X(t)) &= \operatorname{Re}\left\langle \frac{\mathrm{d}}{\mathrm{d}t} f(X), f(X) - D \right\rangle_F \\
&= \operatorname{Re}\left\langle [f(X), B]\,, f(X) - D \right\rangle_F \quad \text{(Proposition 2.1.7)} \\
&= \operatorname{Re}\left\langle B, \left[f(X)^H, f(X) - D\right] \right\rangle_F \quad \text{(Lemma C.2.1)} \\
&= \operatorname{Re}\left\langle B, \pi_{\mathfrak{su}(n)}\left[f(X)^H, f(X) - D\right] \right\rangle_F \quad (B \in \mathfrak{su}(n)) \\
&= -\operatorname{Re}\left\langle B, [\operatorname{Herm}(f(X)), D] \right\rangle_F \quad (D \in \operatorname{Herm}(X)),
\end{aligned}$$

where $\operatorname{Herm}(Y) = \frac{1}{2}(Y + Y^H)$. By Lemma 2.5.5, $[\operatorname{Herm}(f(X)), D] = H_S \pi_1(\operatorname{Herm}(f(X)))$ where $S$ has entries $s_{ij} = |i-j|$. Here is where the arbitrariness of the diagonal entries of $S$ from Definition 2.5.6 come into play. Because $\operatorname{Herm}(f(X))$ is Hermitian, $\pi_1(\operatorname{Herm}(f(X)))$ is zero on the diagonal.

Inserting this into the equation, we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\Psi(X(t)) &= \operatorname{Re}\left\langle B, H_S \pi_1(\operatorname{Herm}(f(X))) \right\rangle_F \\
&= \operatorname{Re}\left\langle [X, B], [X, \pi_1(\operatorname{Herm}(f(X)))] \right\rangle_{QR}.
\end{aligned}$$

If $\frac{\mathrm{d}}{\mathrm{d}t}\Psi(X(t)) = -\|\dot{X}(t)\|_{QR}^2$ for all trajectories then the flow is a gradient flow in the QR metric. This happens if $B(X) = -\pi_1(\mathrm{Herm}(f(X)))$, in which case

$$\dot{X} = [X, \pi_1(\mathrm{Herm}(f(X)))]\,.$$

This completes the proof. □

**Corollary 2.5.9.** *The $f$-QR flow for a* normal *matrix $X_0 \in \mathbb{C}^{n\times n}$ is equal to the gradient $f$-QR flow in Theorem 2.5.8.*

*Proof.* The gradient flow in Theorem 2.5.8 can be rewritten as $[X, \pi_1(\mathrm{Herm}(f(X)))] = [X, \mathrm{Herm}(f(X))] - [X, \pi_2(\mathrm{Herm}(f(X)))] = \frac{1}{2}\left[X, f(X)^H\right] - [X, \pi_2(f(X))]$. Hence the gradient QR flow is

$$\dot{X} = \underbrace{[\pi_2(f(X)), X]}_{\text{QR flow}} + \underbrace{\frac{1}{2}\left[X, f(X)^H\right]}_{=0 \text{ if } X \text{ is normal}}\,. \tag{2.77}$$

If $X$ is normal at one value of $t$ it must be normal for all $t$ by Proposition 2.1.11. Now, by [Hig08], since $X$ is normal we have $\left[X, f(X)^H\right] = 0$. Therefore the gradient $f$-QR flow and the $f$-QR flow are equal. □

The significance of Theorem 2.5.8 is that it adds some insight into why theoretical guarantees for the convergence of the QR algorithm with normal initial matrices such as those in [EH75] and [Bat94] have been easily found by researchers, but convergence results for nonnormal matrices have been shown to be impossible in general [Bat90],[Day96]. One difference it is that for normal matrices, the QR algorithm interpolates a certain gradient flow, but for nonnormal matrices this specific gradient structure (if any) does not exist.

We are now able to reprove already established convergence results (see for example [EH75] and [Bat94]) for the QR algorithm on normal matrices, but here instead using Theorem 2.5.8.

**Lemma 2.5.10.** *Let $Y$ be Hermitian. Then $[Y, \pi_1(Y)] = 0$ implies that $Y$ is diagonal.*

*Proof.* Since the diagonal of $Y$ is real, we have $\pi_1(Y) = Y_L - Y_L^H$. Define the matrix $L = Y_L + \frac{1}{2}Y_D$. Then we have $\pi_1(Y) = L - L^H$ and $Y = L + L^H$. Hence,

$$\begin{aligned}[Y, \pi_1(Y)] &= \left[L + L^H, L - L^H\right] \\ &= 2\left[L^H, L\right].\end{aligned}$$

This implies that $L$ is a diagonal matrix, because diagonal entries of $L^H L$ are $|l_{11}|^2, |l_{22}|^2, \ldots, |l_{nn}|^2$ and diagonal entries of $LL^H$ are $|l_{11}|^2, |l_{22}|^2+|l_{21}|^2, \ldots, \sum_{k=1}^{n} |l_{nk}|^2$. Hence $L^H L = LL^H$ implies that the off-diagonal elements of $L$ are zero. This gives that $Y$ is diagonal. □

**Theorem 2.5.11.** *The stationary points of the gradient $f$-QR flow in Theorem 2.5.8 satisfy*

$$\operatorname{Herm}(f(X)) \in \operatorname{Diag}(n). \tag{2.78}$$

*Proof.* It is clear that points $X$ satisfying $\operatorname{Herm}(f(X)) \in \operatorname{Diag}(n)$ have $\pi(\operatorname{Herm}(f(X)) = 0$ and hence are stationary.

Now suppose that the flow is stationary. Then $[X, \pi_1(\operatorname{Herm}(f(X)))] = 0$. By Remark 2.1.6, $f(X)$ is equal to a polynomial in $X$. Hence $[f(X), \pi_1(\operatorname{Herm}(f(X)))] = 0$. We also have $\left[f(X)^H, \pi_1(\operatorname{Herm}(f(X)))\right] = [-\pi_1(\operatorname{Herm}(f(X))), f(X)]^H = 0$. Hence

$$[\operatorname{Herm}(f(X)), \pi_1(\operatorname{Herm}(f(X)))] = 0. \tag{2.79}$$

By Lemma 2.5.10, $\operatorname{Herm}(f(X)) \in \operatorname{Diag}(n)$. □

*Remark* 2.5.12. This theorem shows that stationary points of the gradient $f$-QR flow have $f(X) = \Omega + \Lambda$ where $\Omega \in \mathfrak{su}(n)$ and $\Lambda \in \operatorname{Diag}(n)$.

**Lemma 2.5.13.** *A stationary point $X \in \mathbb{C}^{n\times n}$ of the gradient $f$-QR flow is stable if and only if*

$$\operatorname{Re}\left\langle [A, D], [A, \operatorname{Herm}(f(X))] \right\rangle_F \leq 0 \text{ for all } A \in C_{\mathfrak{su}(n)}(X)^{\perp}, \tag{2.80}$$

*where $D = \operatorname{diag}(n, n-1, \ldots, 1)$.*

*Remark* 2.5.14. This stability criterion is not easy to check at all. The point we wish to make here is that the stability of the stationary points only depends on the ordered eigenvalues of $\operatorname{Herm}(f(X))$, and not on the skew-Hermitian part of $f(X)$.

*Proof.* Let $Z$ be a $C^1$ path on the isospectral manifold $\mathcal{I}_{\mathfrak{su}(n)}(X)$ such that $Z(0) = X$. Then there exists a path $A$ in $C_{\mathfrak{su}(n)}(Z)$ such that $\dot{Z}(t) = [A(t), Z(t)]$. In the proof of Theorem 2.5.8 we showed that for the function $\Psi(X) = \frac{1}{2}\|f(X) - D\|_F^2$, and a flow $Z$ satisfying $\dot{Z} = [A, Z]$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\Psi(Z(t)) = \operatorname{Re}\left\langle A(t), [D, \operatorname{Herm}(f(Z(t)))] \right\rangle_F.$$

Let us take a further derivative with respect to $t$. By Lemma 2.1.7,

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Psi(Z(t)) &= \operatorname{Re}\langle A(t), [D, \operatorname{Herm}([A(t), f(Z(t))])]\rangle_F \\ &\quad + \operatorname{Re}\left\langle \dot{A}(t), [D, \operatorname{Herm}(f(Z(t)))]\right\rangle_F \\ &= \operatorname{Re}\langle A(t), [D, [A(t), \operatorname{Herm}(f(Z(t)))]]\rangle_F \\ &\quad + \operatorname{Re}\left\langle \dot{A}(t), [D, \operatorname{Herm}(f(Z(t)))]\right\rangle_F\end{aligned}$$

By Theorem 2.5.11, $\operatorname{Herm}(f(X)) \in \operatorname{Diag}(n)$, so $[D, \operatorname{Herm}(f(X))] = 0$. Hence, taking $t = 0$ in this second derivative gives

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Psi(Z(t))\right|_{t=0} = \operatorname{Re}\langle A(0), [D, [A(0), \operatorname{Herm}(f(X))]]\rangle_F .$$

Since $A(0)$ can be arbitrary in $C_{\mathfrak{su}(n)}(X)$, we have the desired result. □

**Theorem 2.5.15** (Convergence of normal Toda flow and QR algorithm [EH75])**.** *Let $X_0 \in \mathbb{C}^{n\times n}$ be a normal matrix. Then the Toda flow $\dot{X} = [X, \pi_1(X)]$ converges to a set of matrices of the form*

$$X = P \begin{pmatrix} Z_1 & & & \\ & Z_2 & & \\ & & \ddots & \\ & & & Z_r \end{pmatrix} P^T, \tag{2.81}$$

*for some permutation matrix $P$. Here each block $Z_i$ is of the form $Z_i = \gamma_i I + \Omega_i$ where $\gamma_i \in \mathbb{R}$ and each $\Omega_i$ is skew-Hermitian.*

*If $X_0$ is also nonsingular, the (normal) unshifted QR algorithm converges to a set of matrices of the form*

$$X = P \begin{pmatrix} \gamma_1 Q_1 & & & \\ & \gamma_2 Q_2 & & \\ & & \ddots & \\ & & & \gamma_r Q_r \end{pmatrix} P^T, \tag{2.82}$$

*for some permutation matrix $P$ and unitary matrices $Q_1, \ldots, Q_r$.*

*Remark* 2.5.16. The singular case is treated by Proposition 2.2.5.

*Proof.* The Toda flow is the $f$-QR flow where $f(z) = z$. Corollary 2.5.9 implies that since $X_0$ is normal, this Toda flow is a gradient QR flow as in Theorem 2.5.8. Combining Theorem 2.5.11 on the stationary points and Theorem 2.4.4 on convergence to the set of stationary points shows that $\mathrm{Herm}(X(t)) \to \mathrm{Diag}(n)$ as $t \to \infty$. Hence $X = \Omega + \Lambda$ where $\Omega \in \mathfrak{su}(n)$ and $\Lambda \in \mathrm{Diag}(n)$. By normality we have

$$[\Omega, \Lambda] = [\mathrm{Skew}(X), \mathrm{Herm}(X)] = 0.$$

Hence $\Omega_{ij} = 0$ unless $i = j$ or $\Lambda_i = \Lambda_j$. This gives the form of the limit required.

The unshifted QR algorithm is the $f$-QR flow for the function $f(z) = \log(z)$. Note that the matrix logarithm of $X_0$ exists (though not necessarily unique) because it is nonsingular [Hig08, Thm 1.27], and the matrix logarithm of $X$ can be defined uniquely by similarity transform from $X_0$. We can use the above reasoning to deduce that $\log(X)$ is of the form in equation (2.81). Taking an exponential gives the desired form. □

*Remark* 2.5.17. This result not new, but we have found it in a completely new way via isospectral gradient flows. See for example [EH75].

### 2.5.1 Scaled Toda-like flows

Let us briefly mention how *scaled Toda-like flows* discussed in a paper of Chu fit into this section [Chu95]. Chu noted that the full Toda flow on symmetric matrices is an example of a system of the form

$$\dot{Y} = [H_N(Y), Y]\,,$$

where $N$ is a skew-symmetric matrix. $N$ has elements $n_{ij} = \mathrm{sign}(j - i)$ in the Toda case. He also noted Bloch's observation that $n_{ij} = j - i$ produces the double bracket flow (Lemma 2.5.4). The form of equation (2.5.1) is similar to the gradient flow for $\Psi(Y) = \frac{1}{2}\|Y - D\|_F^2$ where $D = \mathrm{diag}(n, n-1, \ldots, 1)$ on the symmetric isospectral manifold with metric

$$\langle [A, Y], [B, Y] \rangle_g = \langle A, H_S B \rangle_F, \tag{2.83}$$

for all $A, B \in C_{\mathfrak{so}(n)}(Y)$ and symmetric matrix $S$ with positive entries off the diagonal. In that case the flow is

$$\dot{Y} = [H_S\,[D, Y]\,, Y] \tag{2.84}$$

which by Lemma 2.5.4 is equal to the flow

$$\dot{Y} = [H_N(Y), Y]\,, \tag{2.85}$$

where $N$ has entries

$$n_{ij} = s_{ij}(j - i). \tag{2.86}$$

In order for $S$ to have positive entries off the diagonal, $N$ must have positive entries on the strictly upper triangular part. When $N$ does not satisfy this, then the flow corresponds to a gradient flow in a non-Riemannian metric.

闭门造车，出则合辙

(bì mén zào chē, chū zé hé zhé)

Lit. If you build a cart behind closed doors, then only when it comes out do you know if it fits the track

Fig. A critical reference to a person who works in solitude and disregards the outside world

# Chapter 3

# The symmetric Toeplitz inverse eigenvalue problem

In this chapter we discuss the symmetric Toeplitz inverse eigenvalue problem, that of finding an $n \times n$ real symmetric Toeplitz matrix with prescribed spectrum $\boldsymbol{\lambda} \in \mathbb{R}^n$.

### 3.0.1 Motivation

Motivation for studying this problem can be considered from the point of view of applications, pure mathematics, and also numerical algorithms.

From a pure mathematical perspective, the inverse eigenvalue problem for *real symmetric* Toeplitz matrices is interesting because of its surprising difficulty. As discussed in Subsection 1.1.2, different subclasses of Toeplitz matrix such as circulant matrices, Toeplitz operators and Laurent operators have very elegant properties which lead to simple specification of their spectra in terms of their symbol. Furthermore, there are also elegant results about the *asymptotics* of the eigenvalues of Toeplitz matrices as the dimension $n$ goes to infinity. Nonetheless, there is no simple characterisation of the eigenvalues of a general Toeplitz matrix, or even a real symmetric one [TE05]. The *existence* of an $n \times n$ real symmetric Toeplitz matrix with arbitrary spectrum was a wide open problem until 1994 when Henri Landau provided a nonconstructive topological proof [Lan94].

The nonconstructive nature of Landau's proof leads to the intrigue from the numerical algorithms perspective. There is still no known effective numerical algorithm proven converge to a real symmetric Toeplitz given an arbitrary vector of real eigenvalues. There are numerical algorithms proposed by authors Friedland et al. [FNO87], Chu–

Driessel [CD89], [CD90], Chu–Golub [CG02], Laurie [Lau01], and Trench [Tre97], which have been empirically observed to be effective algorithms with global convergence, but the convergence proofs are still lacking. In this chapter we make only partial headway in the analysis of some of these given algorithms, but we also give an *impractical* but convergent numerical algorithm (see Theorem 3.4.1), to at least settle some theoretical questions regarding computability.

From the applications point of view, a real symmetric Toeplitz matrix arises most naturally as the covariance matrix of a stationary discrete random process [Gra06], [Hay08], [Pro96]. A stationary process $\mathbf{X} = (X_k)_{k\in\mathbb{Z}}$ (as described in Subsection 1.1.2) has joint probability density functions satisfying

$$f(X_{k_1}, X_{k_2}, \ldots, X_{k_n}) = f(X_{k_{1+s}}, X_{k_{2+s}}, \ldots, X_{k_{n+s}}), \tag{3.1}$$

for all shifts $s \in \mathbb{Z}$ i.e. the density depends solely on the relative time differences between the samples. The covariance matrix, $C$, whose entries are defined to be $c_{i,j} = \mathbb{E}((X_{k_i} - \mu_{k_i})(X_{k_j} - \mu_{k_j}))$ where $\mathbb{E}$ is the expectation with respect to the joint probability and $\mu_k = \mathbb{E}(X_k)$, under this assumption is a real symmetric Toeplitz matrix. A consequence of the theory of Principal Component Analysis (see Subsection 1.1.1) is that the eigenvectors of a covariance matrix can be interpretted as an orthogonal splitting of the signal vector. Eigenvectors corresponding to large eigenvalues correspond the salient features of the signal and eigenvectors corresponding to small eigenvalues correspond to random noise. The inverse eigenvalue problem for real symmetric Toeplitz matrices is therefore relevant to areas such as signal processing theory, control theory and system identification [Gra06], [Hay08], [Pro96]. Applications also extend to the trigonometric moment problem and orthogonal polynomials on the unit circle [AK65].

### 3.0.2 Numerical algorithms for the inverse eigenvalue problem

Approaches to solving the symmetric Toeplitz inverse eigenvalue problem can be divided into two camps: Newton iterations [FNO87], [Lau88], [Lau91], [Tre97], [Lau01], [CG05], and isospectral flows [CG05, Chu93, DS99, Chu98, CG02]. In this thesis we are more interested in the latter, but the results of this chapter may have consequences for either approach.

The idea for an isospectral flow approach is inspired by the QR algorithm. The intent is to design a matrix function $B : \mathbb{R}^{n\times n}_{\text{sym}} \to \mathfrak{so}(n)$ such that the flow $\dot{Y} = [B(Y), Y]$ *converges* to a Toeplitz matrix from initial data $Y(0) = Y_0$. Here $Y_0$ is intended to be

a matrix we can easily prescribe the spectrum of (for example an diagonal matrix). Then we can use a geometric integrator to numerically solve the flow whilst preserving the eigenvalues [IMKNZ00], [Zan98].

In this chapter we numerically solve all of our isospectral flows using the most basic geometric integrator,

$$Y_{k+1} = (I - \frac{h}{2}B(Y_k))^{-1}(I + \frac{h}{2}B(Y_k))Y_k(I + \frac{h}{2}B(Y_k))^{-1}(I - \frac{h}{2}B(Y_k)), \tag{3.2}$$

which is a Runge–Kutta–Munthe-Kaas forward Euler method on the isospectral manifold using the Cayley map:

$$\text{Cay} : \mathfrak{so}(n) \to \text{SO}(n), \qquad \text{Cay}(\Omega) = \left(I - \frac{1}{2}\Omega\right)^{-1}\left(I + \frac{1}{2}\Omega\right). \tag{3.3}$$

See Appendix B for more information on the Cayley map. There are of course much better methods which could be used, but this is very simple to implement and suffices for our purposes [Zan98], [CIZ97], [IMKNZ00, App.].

In terms of whether in rigorous terms this defines an effective algorithm, we say the following. If the flow converges to a Toeplitz matrix, then there exists a final time $t$ such that that $Y(t)$ is "close" to a symmetric Toeplitz matrix in Frobenius distance (for example), and if the numerical method is a convergent one, then the step size $h$ in the discretisation can be taken small enough so that the final iterate $Y_N$ is "close" to the analytical solution $Y(t)$, and then by the triangle inequality $Y_N$ is "close" to a symmetric Toeplitz matrix. This shows that we have two requirements for this to constitute a bona fide algorithm: the flow converges to a Toeplitz matrix, and the numerical method used is a convergent one. Note that except for the small effect of rounding errors, the eigenvalues of $Y_N$ are *exactly* as required, because of the use of a geometric integrator. Therefore, the major challenge which must first be tackled is to understand the convergence (or lack of) of the continuous flows to Toeplitz matrices. This is the focus of the chapter. Error estimates for Lie group integrators are discussed in [IMKNZ00].

There are two principal candidates for the choice of matrix function $B : \mathbb{R}^{n\times n}_{\text{sym}} \to \mathfrak{so}(n)$ which are intended to solve the symmetric Toeplitz inverse eigenvalue problem. The first is a gradient flow for the function $\Psi(Y) = \frac{1}{2}\|Y - P_T(Y)\|_F^2$, where $P_T$ is the orthogonal projection from $\mathbb{R}^{n\times n}_{\text{sym}}$ onto the subspace of Toeplitz matrices, first published

in [CD90],

$$\dot{Y} = [[P_T(Y), Y], Y]. \tag{3.4}$$

For solutions of this flow, $\Psi(Y)$ is nonincreasing and flows converge to points such that $[P_T(Y), Y] = 0$ (first shown in [CD90] and follows from Theorem 2.4.4 and Theorem 2.4.7). Stable stationary points that are *not* Toeplitz matrices (which can be a problem for a numerical algorithm) were reported (numerically) in [CG05], and it is suggested there to investigate the stationary points as a possible avenue of research. In Section 3.1.1 we *prove* using a symbolic calculation that for the $3 \times 3$ case there are stable stationary points which are not Toeplitz and in Subsection 3.3 we give numerical examples of stable non-Toeplitz stationary points in the $4 \times 4$ case.

The second flow is Chu's flow, first studied in the unpublished manuscript [Chu93], but later studied in [Chu94], [Chu98], [DS99], [DS02], [CG02], [CG05]. Chu's flow is of the form

$$\dot{Y} = [B(Y), Y], \tag{3.5}$$

where $B$ is the *Toeplitz annihilator*,

$$B(Y)_{i,j} = \begin{cases} y_{i,j-1} - y_{i+1,j} & \text{if } i < j, \\ 0 & \text{if } i = j, \\ y_{i,j+1} - y_{i-1,j} & \text{if } i > j. \end{cases} \tag{3.6}$$

To understand the motivation behind this flow, consider the case $n = 4$:

$$B(Y) = \begin{pmatrix} 0 & y_{1,1} - y_{2,2} & y_{1,2} - y_{2,3} & y_{1,3} - y_{2,4} \\ y_{2,2} - y_{1,1} & 0 & y_{2,2} - y_{3,3} & y_{2,3} - y_{3,4} \\ y_{3,2} - y_{2,1} & y_{3,3} - y_{2,2} & 0 & y_{3,3} - y_{4,4} \\ y_{4,2} - y_{3,1} & y_{4,3} - y_{3,2} & y_{4,4} - y_{3,3} & 0 \end{pmatrix}. \tag{3.7}$$

Clearly, $B(Y) = 0$ if and only if $Y$ is Toeplitz. Chu also showed that if $Y$ has distinct eigenvalues then $[B(Y), Y] = 0$ if and only if $Y$ is Toeplitz (see Proposition 3.1.2). Therefore, since eigenvalues remain fixed throughout the flow, for initial data with distinct eigenvalues, if Chu's flow converges then it can only converge to a Toeplitz matrix.

### 3.0.3 Landau's Theorem and eigenvalue parity

Let us discuss some details of Landau's Theorem as it will allow the explanation of the contributions of this chapter to the analysis of the inverse eigenvalue problem.

**Theorem 3.0.1** (Landau [Lan94]). *Any vector* $\boldsymbol{\lambda} \in \mathbb{R}^n$ *is the spectrum of an* $n \times n$ *real symmetric Toeplitz matrix.*

An $n \times n$ symmetric Toeplitz matrix is defined by its first row $t_0, t_1, \ldots, t_{n-1}$. First note that it is sufficient to consider matrices which are trace free, because we can add a multiple of the identity to any Toeplitz matrix and it will remain Toeplitz. Landau assumes that $t_0 = 0$ and $t_1 = 1$ and considers the following map $\Lambda : \mathbb{R}^{n-2} \to \mathbb{R}^{n-2}$. $\Lambda$ takes $(t_2, t_3, \ldots, t_{n-1})^T \in \mathbb{R}^{n-2}$, and finds the eigenvalues $\lambda_1 \leq \ldots \leq \lambda_n$ of the symmetric Toeplitz matrix with first row $(0, 1, t_2, \ldots, t_{n-1})$, since the trace of this matrix is zero and it is not the zero matrix, $\lambda_1 < 0$, so the following vector in $\mathbb{R}^{n-2}$ is well-defined:

$$\Lambda(t_2, \ldots, t_{n-1}) = (y_2, y_3, \ldots, y_{n-1}), \text{ where } y_k = \lambda_k / |\lambda_1|. \tag{3.8}$$

The image of $\Lambda$ is clearly not the whole of $\mathbb{R}^{n-2}$, because the eigenvalues were sorted in the process. The image is a simplex defined as the intersection of $n-1$ linear inequalities:

$$L_n = \left\{ (y_2, \ldots, y_{n-1})^T \in \mathbb{R}^{n-2} : \begin{array}{c} -1 \leq y_2 \leq y_3 \leq \cdots \leq y_{n-1} \\ y_2 + \cdots + y_{n-2} + 2y_{n-1} \leq 1 \end{array} \right\}. \tag{3.9}$$

The last inequality comes from $\sum_{k=1}^{n} \lambda_k = 0$. This thus far is quite elementary; Landau's key insight is his identification of a candidate preimage $\Lambda^{-1}(L_n)$ — a set he terms *regular* Toeplitz matrices — and a proof that $\Lambda$ does indeed map this restricted set of Toeplitz matrices onto $L_n$, using the topological degree of $\Lambda$.

What is a regular Toeplitz matrix? Symmetric Toeplitz matrices are part of a larger space of matrices called centrosymmetric matrices, those matrices $X$ such that

$EXE = X$, where $E$ is the exchange matrix,

$$E = E_{n\times n} = \begin{pmatrix} & & & & 1 \\ & & & 1 & \\ & & \iddots & & \\ & 1 & & & \\ 1 & & & & \end{pmatrix}.$$

A vector $v$ is said to be *even* if $Ev = v$ and *odd* if $Ev = -v$. Let $X \in \mathbb{R}^{n\times n}_{\text{sym}}$ be a symmetric and centrosymmetric matrix, which is also called a bisymmetric matrix. Then there exists an orthonormal basis of eigenvectors $v_1, v_2, \ldots, v_p, w_1, w_2, \ldots, w_q$ such that the $v_i$'s are even and the $w_i$'s are odd [CB76]. Here $p = \lceil n/2 \rceil$ and $q = \lfloor n/2 \rfloor$.

A bisymmetric matrix has *alternating parity* if the greatest eigenvalue is even, and if the $k$th greatest eigenvalue is odd then the $k+1$st is even. A real symmetric Toeplitz matrix is *regular* if it has distinct eigenvalues, alternating parity and so do all of it's principal $k \times k$ minors for $k = 2, \ldots, n$.

The alternating parity assumption isn't necessary for existence of a real symmetric Toeplitz matrix. In Theorem 3.3.4 we prove the new result that there exists a $3 \times 3$ real symmetric Toeplitz matrix with even eigenvalues $\lambda_1, \lambda_2$ and odd eigenvalue $\mu_1$ if and only if

$$(\lambda_1 - \mu_1)(\mu_1 - \lambda_2) + 2(\lambda_1 - \lambda_2)^2 \geq 0.$$

### 3.0.4 Bisymmetric isospectral flows

Symmetric Toeplitz matrices are bisymmetric, which means they are symmetric along both the top-left-to-bottom-right and top-right-to-bottom-left diagonals. Let $X \in \text{Bisym}(n)$, the space of bisymmetric $n \times n$ matrices, and suppose $X$ has $n$ distinct eigenvalues. Then the bisymmetric isospectral manifold is

$$\mathcal{BI} = \{Y \in \text{Bisym}(n) : \text{eigs}(Y) = \text{eigs}(X)\}. \tag{3.10}$$

In Section 3.2, we prove that this manifold has $\binom{n}{p}$ connected components, where $p = \lceil \frac{n}{2} \rceil$. Each component has dimension $\frac{1}{2}p(p-1) + \frac{1}{2}q(q-1)$ where $q = n - p$, and may be parametrised by a representative matrix in the connected Lie group of centrosymmetric orthogonal matrices. This is not a completely new result, but is certainly a contribution as a fully formalised proof of the result, as Chu discusses the

different connected components of the manifold for the $3 \times 3$ case but does not go into detail [Chu93]. Each component corresponds to a choice of parity for these $n$ distinct eigenvalues. Hence, Landau's Theorem informs us of a connected component that is guaranteed to contain a symmetric Toeplitz matrix — the connected component containing matrices with alternating parity.

The fact that each connected component of the bisymmetric isospectral manifold is acted upon by a Lie group with dimension $\frac{1}{2}p(p-1) + \frac{1}{2}q(q-1)$ is important. This allows us to parametrise the manifold by the associated Lie algebra, which has the same small dimension. Using this, we can reduce the $3 \times 3$ bisymmetric isospectral flow to a one dimensional flow, the $4 \times 4$ bisymmetric isospectral flow to a two dimensional flow, and the $5 \times 5$ bisymmetric isospectral flow to a four dimensional flow and so on. In Subsection 3.2.3 we derive an analytical solution for the trajectories of $3 \times 3$ bisymmetric isospectral flows. Then in Section 3.3, conduct a numerical study of the $4 \times 4$ gradient flow and Chu flow for the inverse Toeplitz eigenvalue problem and gain some insights that have until now been occurring in a 5 or 6 dimensional space rather than this 2 dimensional one.

At the end of the chapter we briefly discuss an extremely impractical, brute force approach to the computation which produces isospectral iterates $Y_0, Y_1, \ldots$ which converge to a symmetric Toeplitz matrix. The reason is purely theoretical, to show that the Solvability Complexity Index of the problem is 1 (see [BAHNS15a] and Section 4.5).

## 3.1 Isospectral flows for Toeplitz inverse eigenvalue problems

In this section we give two types of isospectral flows that can be used to compute a symmetric Toeplitz matrix with prescribed spectrum.

### 3.1.1 Isospectral gradient flows

Isospectral gradient flows in an arbitrary metric were derived in Chapter 2. In this section we apply the results to the symmetric Toeplitz inverse eigenvalue problem.

**Proposition 3.1.1.** *The space of real symmetric matrices can be decomposed into orthogonal subspaces*

$$\mathbb{R}^{n\times n}_{\text{sym}} = \mathcal{T}_n \oplus \mathcal{Z}_n, \tag{3.11}$$

*where $\mathcal{T}_n$ is the space of real symmetric Toeplitz matrices and $\mathcal{Z}_n$ is the space of real symmetric matrices whose diagonals sum to zero:*

$$\mathcal{Z}_n = \left\{ Z \in \mathrm{S}_n : \sum_{i=1}^{n-j} Z_{i,i+j} = 0,\ \forall j \in \{0, 1, \ldots, n-1\} \right\}. \tag{3.12}$$

*Proof.* For any $T \in \mathcal{T}_n$ and $Z \in \mathcal{Z}_n$, $\langle T, Z \rangle_F = 0$. Furthermore, the dimension of $\mathcal{T}_n$ is $n$ and the dimension of $\mathcal{Z}_n$ is $\frac{1}{2}n(n-1)$, which add up to the dimension of $\mathbb{R}^{n\times n}_{\mathrm{sym}}$. □

We denote the projection operators onto $\mathcal{T}_n$ and $\mathcal{Z}_n$ by $\mathcal{P}_T$ and $\mathcal{P}_Z$ respectively. In coordinates, $\mathcal{P}_T$ simply averages each diagonal:

$$\mathcal{P}_T(Y)_{i,j} = \frac{1}{n - |j-i|} \sum_{l-k=j-i} y_{k,l} \tag{3.13}$$

and $\mathcal{P}_Z(Y) = Y - \mathcal{P}_T(Y)$.

Recall the setup of an isospectral gradient flow: We have a function $f : \mathbb{R}^{n\times n}_{\mathrm{sym}} \to \mathbb{R}$, an isospectral manifold $\mathcal{I}_{\mathrm{SO}(n)}(Y_0)$ for a starting matrix $Y_0$, and Riemannian metric

$$\langle [A, Y], [B, Y] \rangle_g = \langle A, L_Y B \rangle_F, \tag{3.14}$$

for $A, B \in C_{\mathfrak{so}(n)}(Y)$. By Theorem 2.4.6, the isospectral gradient flow in this setup is

$$\dot{Y} = -\left[ L_Y^{-1} \left[ \nabla\Psi(Y), Y \right], Y \right].$$

The most obvious function to optimise in order to obtain a symmetric Toeplitz matrix is

$$f(Y) = \|P_Z(Y)\|_F^2. \tag{3.15}$$

The isospectral gradient flow for this function with the normal metric is

$$\dot{Y} = [[P_T(Y), Y], Y]. \tag{3.16}$$

Chu and Golub study this flow in [CG02], [CG05]. By Theorem 2.4.4 a gradient flow with isolated stationary points always converges to a stationary point. Therefore, the study of the stationary points is a sufficient analysis for the convergence or nonconvergence question for Toeplitz matrices.

By Theorem 2.4.7, the stationary points of this flow are points such that $[P_T(Y), Y] = 0$, which is equivalent to

$$[P_T(Y), P_Z(Y)] = 0. \tag{3.17}$$

Since points $Y \in \mathcal{T}$ are global minimisers, these stationary points are all stable. By Corollary 2.4.9, the points $Y \notin \mathcal{T}$ are stable if and only if

$$\langle P_Z([A, Y]) - [A, P_Z(Y)], [A, Y]\rangle_F \geq 0 \text{ for all } A \in \mathfrak{so}(3),$$

because $\nabla f(Y) = P_Z(Y)$ and $\nabla^2 f(Y) = P_Z$ for $Y \notin \mathcal{T}$. Now, since $[A, P_Z(Y)] = P_Z([A, P_Z(Y)]) + P_T([A, P_Z(Y)])$, and using Lemma C.2.1, we have stability if and only if the operator

$$A \mapsto [Y, P_Z([A, P_T(Y)] - P_T([A, P_Z(Y)]] \tag{3.18}$$

is positive definite on $\mathfrak{so}(n)$. Note the this formula allows one to numerically check the stability of a stationary point by computing the eigenvalues of this operator when considered as a matrix mapping $\mathbb{R}^{\frac{1}{2}n(n-1)} \to \mathbb{R}^{\frac{1}{2}n(n-1)}$. Indeed we do this below symbolically for $3 \times 3$ case.

In the space of symmetric matrices with Frobenius norm equal to 1, $f(Y) = \|P_Z(Y)\|_F^2$ has a straightforward set of stationary points. The Toeplitz matrices $\mathcal{T}$ are the global minimisers with $f(Y) = 0$ and the complement $\mathcal{Z}$ are global maximisers. The stationary points on the isospectral manifold, which are characterised by $[P_T(Y), P_Z(Y)] = 0$, however, include other kind of matrices. Let us look at specific cases for $n$.

To simplify matters we will consider only trace free matrices. The convergence behaviour of the flow (3.16) is not affected by this restriction, since $[[P_T(Y + \alpha I), Y + \alpha I], Y + \alpha I] = [[P_T(Y), Y], Y]$.

In the $2 \times 2$ case we have

$$\mathbb{R}^{2\times 2}_{\text{sym}} \cap \mathfrak{sl}(2) = \text{Span}\left\{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\right\} = \{T_1, Z_1\}.$$

Since $[T_1, Z_1] = 0$, the only stationary points of the flow are multiples of $T_1$ and multiples of $Z_1$. Since $T_1$ is a minimiser and $Z_1$ is a maximiser, $T_1$ is stable and $Z_1$ is unstable.

In the $3\times 3$ case the stationary points are more involved. We have the decomposition.

$$\mathbb{R}^{3\times 3}_{\text{sym}} \cap \mathfrak{sl}(3) = \text{Span}\left\{ \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix} \right.$$
$$\left. \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -1 \end{pmatrix} \right\}$$
$$= \{T_1, T_2, T_3, Z_1, Z_2, Z_3\}\,.$$

This particular choice of basis is not orthogonal in the Frobenius inner product, but satisfies $[T_i, Z_i] = 0$ for $i = 1, 2, 3$, and commutation of any other pair of elements is nonzero. Hence by the stationarity condition $[P_Z(X), P_T(X)] = 0$, we have stationary points of the form,

$$Y = \begin{pmatrix} b & 0 & a \\ 0 & -2b & 0 \\ a & 0 & b \end{pmatrix}, \quad \begin{pmatrix} b & a-b & a \\ a-b & 0 & a+b \\ a & a+b & -b \end{pmatrix}, \quad \begin{pmatrix} b & a+b & -a \\ a-b & 0 & a-b \\ -a & a-b & -b \end{pmatrix}$$
$$= Y_1, Y_2, Y_3.$$

What about the stability of these points? Recall that a stationary point for this flow is stable if and only if the operator given in equation (3.18) is positive definite. The following snippet of Matlab code will compute a $3 \times 3$ matrix representation (with respect to a basis for $\mathfrak{so}(3)$) of that operator for arbitrary $a$ and $b$ in $Y_1$, and then compute the eigenvalues of that matrix.

```
syms a b 'real'
Z = [b 0 0; 0 -2*b 0; 0 0 b];
T = [0 0 a; 0 0 0; a 0 0];
Y = T+Z; % Stat. pt.
L = sym(zeros(3)); % Manifold Hessian around stat. pt.
for i = 1:2
 for j = i+1:3
  for k = 1:2
   for l = k+1:3
    Eij = zeros(3);Eij(i,j) = 1;Eij(j,i) = -1; % so(3) basis element
    Ekl = zeros(3);Ekl(k,l) = 1;Ekl(l,k) = -1; % so(3) basis element
    C = Y*Eij-Eij*Y; % [Y,Eij]
```

```
13      PTC = toeplitz([mean(diag(C)), mean([C(1,2),C(2,3)]), C(1,3)]);
14      PZC = C - PTC; % P_Z([Y,Eij])
15      L((i-1)+(j-1),(k-1)+(l-1)) = trace((Y*Ekl-Ekl*Y)*(PZC-(Z*Eij-Eij*
            Z)));
16    end
17   end
18  end
19 end
20 eigs = eig(L)
```

The matrix for stability of $Y_1$ (with respect to a certain basis for $\mathfrak{so}(3)$) computed by this Matlab code (denoted `L` in the code) is

$$L = \begin{pmatrix} a^2 - 6ab - 9b^2 & 0 & a^2 + 9b^2 \\ 0 & 8a^2 & 0 \\ a^2 + 9b^2 & 0 & a^2 - 6ab - 9b^2 \end{pmatrix}. \tag{3.19}$$

The eigenvalues of this matrix are $16a^2, 2a^2 - 6ba, -18b^2 - 6ab$. Therefore if we take for example $a = -4$ and $b = 1$, then the stationary point $Y_1$ is stable but not Toeplitz. In Section 3.3, we show how the parity of eigenvalues (discussed above) inform us how to avoid these stable stationary points.

We could consider the $4\times4$ case, but the dimension of the flow is then $\dim(\mathfrak{so}(4)) = 6$, which is quite difficult to deal with for many reasons. In Section 3.3 we restrict the flow to bisymmetric matrices, because then the flow can be reduced to 2 dimensional flow. This has never been done before in the literature, and it makes observing the behaviour of the flow tractable.

### 3.1.2 Chu's flow

The following flow was introduced by Chu [Chu93] as a method for solving the inverse eigenvalue problem for symmetric Toeplitz matrices.

$$\dot{Y} = [B(Y), Y], \quad Y_0 \in \mathbb{R}^{n\times n}, \sigma(Y_0) = (\lambda_1, \ldots, \lambda_n)^T, \tag{3.20}$$

where $B$ is the *Toeplitz annihilator*,

$$B(Y)_{i,j} = \begin{cases} y_{i,j-1} - y_{i+1,j} & \text{if } i < j, \\ 0 & \text{if } i = j, \\ y_{i,j+1} - y_{i-1,j} & \text{if } i > j. \end{cases} \tag{3.21}$$

To understand how this works, consider the case $n = 4$:

$$B(Y) = \begin{pmatrix} 0 & y_{1,1} - y_{2,2} & y_{1,2} - y_{2,3} & y_{1,3} - y_{2,4} \\ y_{2,2} - y_{1,1} & 0 & y_{2,2} - y_{3,3} & y_{2,3} - y_{3,4} \\ y_{3,2} - y_{2,1} & y_{3,3} - y_{2,2} & 0 & y_{3,3} - y_{4,4} \\ y_{4,2} - y_{3,1} & y_{4,3} - y_{3,2} & y_{4,4} - y_{3,3} & 0 \end{pmatrix}. \tag{3.22}$$

It is clear that if $Y$ is symmetric, then $B(Y)$ is skew-symmetric. Also, $B(Y) = 0$ if and only if $Y$ is a Toeplitz matrix. Less clear is that the flow in equation (3.20) is stationary if and only if $Y$ is Toeplitz.

**Proposition 3.1.2** ([Chu93]). *If $Y_0$ has distinct eigenvalues, then all fixed points of the Chu flow are Toeplitz matrices.*

*Proof.* This follows immediately from Proposition 2.1.10. □

This proves that for distinct eigenvalues Chu's flow cannot converge to a non-Toeplitz matrix. Note however, that if $Y_0$ does not have distinct eigenvalues there can be non-Toeplitz fixed points. It was found by Chu [Chu93] that

$$\begin{pmatrix} a & 0 & -3a \\ 0 & -2a & 0 \\ -3a & 0 & a \end{pmatrix}$$

is a stationary point of the flow. In Section 3.3.2 we show that this matrix can be avoided by using the parity of the eigenvalues.

In contrast to the gradient flow in Subsection 3.1.1, there is no reason to expect that Chu's flow will converge to a fixed point. Indeed there are periodic orbits. In Section 3.3.2 we prove that Chu's flow with initial data of the form

$$Y_0 = \begin{pmatrix} x & 0 & z \\ 0 & -2x & 0 \\ z & 0 & x \end{pmatrix} \tag{3.23}$$

gives a periodic orbit if and only if $zx + x^2 < 0$.

If there are non-Toeplitz fixed points, and periodic orbits in Chu's flow, then why is it worth studying for this problem?

Chu's flow has been studied in several papers [Chu93], [DS99], [DS02], [DS03]. All claim that Chu's flow *always* converges in practice, despite these theoretical results

showed the contrary. Figure 3.1 shows why this might not be a contradiction. In Figure 3.1, Chu's flow is solved numerically using a basic geometric integrator,

$$Y_{k+1} = (I - \frac{h}{2}B(Y_k))^{-1}(I + \frac{h}{2}B(Y_k))Y_k(I + \frac{h}{2}B(Y_k))^{-1}(I - \frac{h}{2}B(Y_k)), \tag{3.24}$$

which is a Runge–Kutta–Munthe-Kaas forward Euler method on the isospectral manifold using the Cayley map:

$$\text{Cay} : \mathfrak{so}(n) \to \text{SO}(n), \qquad \text{Cay}(\Omega) = \left(I - \frac{1}{2}\Omega\right)^{-1}\left(I + \frac{1}{2}\Omega\right). \tag{3.25}$$

See Appendix B for more information on the Cayley map [Zan98], [CIZ97], [IMKNZ00]. The initial condition used is

$$Y_0 = \begin{pmatrix} 1 & 1 & -3 \\ 1 & -2 & 1 \\ -3 & 1 & 1 \end{pmatrix}. \tag{3.26}$$

What we see is that the flow begins periodic with period approximately 1.047, but around time $t = 4.2$ the dynamics change and the flow converges to a Toeplitz matrix.

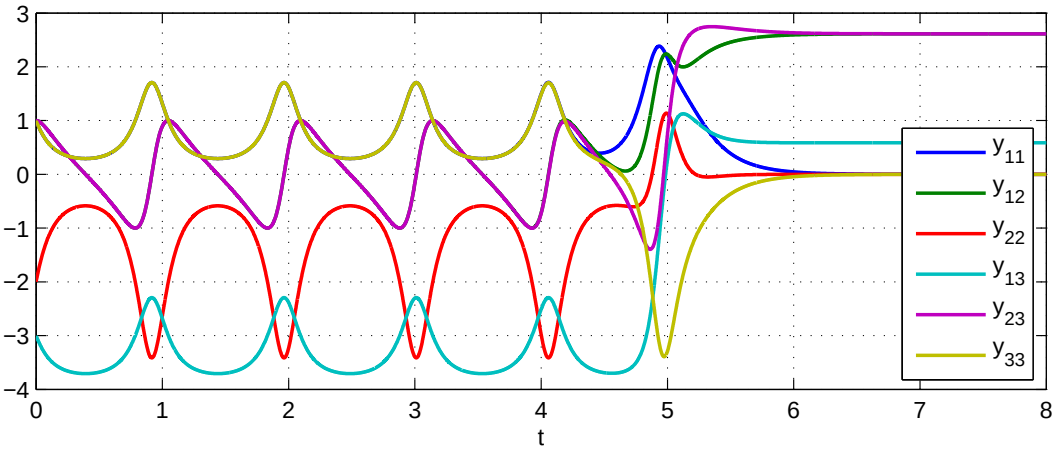



Fig. 3.1 This is a numerical solution to Chu's flow (equation (3.20)) solved using a Cayley Forward Euler method (equation (3.24)), a stepsize $h = 0.002$, and initial condition given in equation (3.26). Each line is an entry of the symmetric matrix as described in the legend. In exact arithmetic, the trajectory is periodic with period approximately 1.047 (the periodicity is proven in Subsection 3.2.3). However, the numerical solution quickly leaves the orbit due to discretisation and rounding errors. If the step size $h$ is changed then the time $t = hn$ at which the flow leaves the orbit changes.

In Subsection 3.3.2 we discuss Chu's flow restricted to bisymmetric matrices, but first we must derive some results about the manifold such a flow evolves on.

## 3.2 The bisymmetric isospectral manifold

In this section we describe fine properties of centrosymmetric matrices, (a superset of bisymmetric matrices), leading to discussion of the structure of isospectral manifolds of bisymmetric matrices.

### 3.2.1 Centrosymmetric matrices

Define the *exchange matrix*,

$$E = E_{n\times n} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{pmatrix}. \tag{3.27}$$

**Definition 3.2.1.** A matrix $X \in \mathbb{R}^{n\times n}$ is *centrosymmetric* if $EXE = X$ (equivalently, if $x_{n+j-1,n+i-1} = x_{ij}$ for all $i, j$). A matrix is *bisymmetric* if it is both symmetric and centrosymmetric (equivalently, if $x_{n+i-1,n+j-1} = x_{ij} = x_{ji} = x_{n+j-1,n+i-1}$ for all $i, j$). These two spaces we denote $\mathrm{Centro}(n)$ and $\mathrm{Bisym}(n)$ respectively.

Informally, a matrix is centrosymmetric if, when the entries are rotated $\pi$ radians, the matrix remains the unchanged. In the literature, matrices that are invariant under reflecting the entries across the antidiagonal are known as *persymmetric* [MMD03]; we are not interested in this particular property in this particular chapter.

Centrosymmetric matrices have a nice block structure, and can be block diagonalised by the involutory orthogonal matrix $K$, as seen in Table 3.1.

**Lemma 3.2.2** ([CB76])**.** $\mathrm{Centro}(n)$ *is an algebra isomorphic to* $\mathbb{R}^{r\times r} \times \mathbb{R}^{s\times s}$.

*Proof.* $\mathrm{Centro}(n)$ is clearly a linear space. To see it is an algebra, for $X, Y \in \mathrm{Centro}(n)$, $EXYE = EXEEYE = XY$ so $XY \in \mathrm{Centro}(n)$. The isomorphism is given by $(X_1, X_2) \mapsto K\mathrm{diag}(X_1, X_2)K$. □

Table 3.1 Structure of centrosymmetric matrices. This is a modified version of that in [CG05, p. 87], [CG02], [CB76], modified so that the matrix $K$ here is symmetric. This makes working with these block structures easier as we do not need to worry about the difference between $K$ and $K^T$.

| | Even $n$ | Odd $n$ |
|---|---|---|
| $X$ | $\begin{pmatrix} A & CE \\ EC & EAE \end{pmatrix}$ | $\begin{pmatrix} A & b & CE \\ d^T & e & d^TE \\ EC & Eb & EAE \end{pmatrix}$ |
| $\sqrt{2}K$ | $\begin{pmatrix} I & E \\ E & -I \end{pmatrix}$ | $\begin{pmatrix} I & 0 & E \\ 0 & \sqrt{2} & 0 \\ E & 0 & -I \end{pmatrix}$ |
| $KXK$ | $\begin{pmatrix} A+C & 0 \\ 0 & E(A-C)E \end{pmatrix}$ | $\begin{pmatrix} A+C & \sqrt{2}b & 0 \\ \sqrt{2}d^T & e & 0 \\ 0 & 0 & E(A-C)E \end{pmatrix}$ |

**Lemma 3.2.3** ([CB76])**.** *For $X \in \mathrm{Centro}(n)$, any simple eigenvector is either even or odd.*

*Proof.* If $Xv = \lambda v$ then $XEv = EXv = \lambda Ev$. By simplicity of $\lambda$, $Ev = \pm v$, so $v$ is either odd or even. □

**Lemma 3.2.4** ([CB76])**.** *Let $X \in \mathrm{Bisym}(n)$. Then there exists an orthonormal basis of eigenvectors $\{v_1, v_2, \ldots, v_p, w_1, w_2, \ldots, w_q\}$ for $X$ such that the $v_i$'s are even and the $w_i$'s are odd. Here $p = \lceil n/2 \rceil$ and $q = \lfloor n/2 \rfloor$.*

*Proof.* Using Table 3.1, $KXK = \mathrm{diag}(X_1, X_2)$ where $X_1 \in \mathrm{Sym}(p)$ and $X_2 \in \mathrm{Sym}(q)$. Let $Q_1 \in \mathrm{SO}(p)$, $Q_2 \in \mathrm{SO}(q)$ such that $Q_1^T X_1 Q_1$ and $Q_2^T X_2 Q_2$ are diagonal. Then the eigenvectors of $X$ are the orthonormal columns of $K\mathrm{diag}(Q_1, Q_2)K$, the first $p$ of which are even and the last $q$ odd. □

**Definition 3.2.5** (Eigenvalue parity)**.** Let $X \in \mathrm{Bisym}(n)$. By Lemma 3.2.4, we can say that $X$ has $p = \lceil n/2 \rceil$ even eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_p$ and $q = \lfloor n/2 \rfloor$ odd eigenvalues $\mu_1 \le \mu_2 \le \cdots \le \mu_q$.

**Definition 3.2.6.** Define the Lie groups $\mathrm{CO}(n)$ and $\mathrm{SCO}(n)$ by

$$\mathrm{CO}(n) = K\mathrm{diag}(\mathrm{O}(p), \mathrm{O}(q))K, \tag{3.28}$$

$$\mathrm{SCO}(n) = K\mathrm{diag}(\mathrm{SO}(p), \mathrm{SO}(q))K, \tag{3.29}$$

where $K$, $p$ and $q$ are as in Table 3.1 and Lemma 3.2.4.

**Lemma 3.2.7.** *The Lie algebra derived from both* $\mathrm{CO}(n)$ *and* $\mathrm{SCO}(n)$ *is*

$$\mathfrak{sco}(n) = K\mathrm{diag}(\mathfrak{so}(p), \mathfrak{so}(q))K. \tag{3.30}$$

*Proof.* Consider a smooth path $\rho : (-\epsilon, \epsilon) \to \mathrm{SCO}(n)$ such that $\rho(0) = I$. Write $\rho = K\mathrm{diag}(\rho_1, \rho_2)K$. Then

$$\rho'(0) = K\mathrm{diag}(\rho_1'(0), \rho_2'(0))K \in K\mathrm{diag}(\mathfrak{so}(p), \mathfrak{so}(q))K. \tag{3.31}$$

This completes the proof that $\mathfrak{sco}(n)$ is the derived Lie algebra for $\mathrm{SCO}(n)$. The proof for $\mathrm{CO}(n)$ is analogous. □

**Lemma 3.2.8.** *We have the following technical facts regarding* $\mathrm{CO}(n)$, $\mathrm{SCO}(n)$ *and* $\mathfrak{sco}(n)$.

1. $\mathrm{CO}(n) = \mathrm{O}(n) \cap \mathrm{Centro}(n)$
2. $\mathfrak{sco}(n) = \mathfrak{so}(n) \cap \mathrm{Centro}(n)$
3. $\mathrm{SCO}(n) = \exp(\mathfrak{sco}(n))$
4. $\mathrm{SCO}(n)$ *is a connected normal Lie subgroup of* $\mathrm{CO}(n)$ *with quotient group isomorphic to the Klein 4 group* $V_4$ *(i.e.* $\mathrm{CO}(n)$ *has 4 connected components)*
5. $\mathrm{CO}(n)$ *and* $\mathrm{SCO}(n)$ *are quadratic Lie groups as in Subsection B.3.1 and* $\mathrm{SCO} = \mathrm{Cay}(\mathfrak{sco}(n))$*, where* Cay *is the Cayley map* $\Omega \to (I + \frac{1}{2}\Omega)^{-1}(I - \frac{1}{2}\Omega)$

*Proof.* 1. Let $Q = K\mathrm{diag}(Q_1, Q_2)K \in \mathrm{CO}(n)$. Then $Q$ is orthogonal as it is the product of three orthogonal matrices, and it is centrosymmetric by Table 3.1. Hence $\mathrm{CO}(n) \subseteq \mathrm{O}(n) \cap \mathrm{Centro}(n)$. Now for the reverse inclusion. If a matrix $Q$ is centrosymmetric then it is of the form $Q = K\mathrm{diag}(Q_1, Q_2)K$ by Table 3.1. If $Q$ is also orthogonal, then $\mathrm{diag}(Q_1, Q_2) = KQK$ is orthogonal, so that $Q_1$ and $Q_2$ are orthogonal, meaning $Q$ is in fact a member of $\mathrm{CO}(n)$.

2. Essentially the same proof as for part 1.

3. Since exp is analytic, we can simply show $\exp(\mathfrak{sco}(n)) = \exp\left(K\operatorname{diag}(\mathfrak{so}(p), \mathfrak{so}(q))K\right) = K\exp(\operatorname{diag}(\mathfrak{so}(p), \mathfrak{so}(q)))K = K\operatorname{diag}(\mathrm{SO}(p), \mathrm{SO}(q))K = \mathrm{SCO}(n)$.

4. $\mathrm{SCO}(n)$ is a connected normal Lie subgroup of $\mathrm{CO}(n)$ because $\mathrm{SO}(p) \times \mathrm{SO}(q)$ is a connected normal Lie subgroup of $\mathrm{O}(p) \times \mathrm{O}(q)$. The quotient group is isomorphic to $\mathrm{O}(p)/\mathrm{SO}(p) \times \mathrm{O}(q)/\mathrm{SO}(q) \cong V_4$.

5. We have that $\mathrm{CO}(n) = \left\{Q \in \mathrm{GL}(n) : Q^TQ = I,\, Q^TEQ = E\right\}$, so it is quadratic as in Subsection B.3.1 and the Cayley map sends $\mathfrak{sco}(n) \to \mathrm{SCO}(n)$.

□

An alternative formulation of $\mathrm{CO}(n)$ is as the *indefinite orthogonal group* $\mathrm{O}_E(p, q)$ of orthogonal matrices invariant under the nondegenerate bilinear form $\langle v, w\rangle_E = v^TEw$. In this interpretation $\mathrm{SCO}(n) = \mathrm{SO}^+_E(p, q)$ the connected component of $\mathrm{O}_E(p, q)$ containing the identity. The most famous split orthogonal group is the Lorentz group $\mathrm{SO}^+(1, 3)$, important in electromagnetism and special relativity, whose nondegenerate bilinear form is $\langle v, w\rangle_{1,3} = -v_1w_1 + v_2w_2 + v_3w_3 + v_4w_4$.

A nice way to think about $\mathrm{SPO}(n)$ is that it is the Lie group of centrosymmetric orthogonal matrices connected to the identity. We can go further than this and describe a basis of *bisymmetry preserving Givens rotations*.

**Lemma 3.2.9** (Bisymmetry preserving Givens rotations (see for example [MMD03]))**.** *Any element $U \in \mathrm{SCO}(n)$ can be written as a product of elements of the lexicographically ordered set*

$$\begin{aligned}\mathcal{B} := &\left\{U_{ij}(\theta) : 1 \le i < j \le \frac{n}{2}\right\} \cup \left\{V_{i,j}(\theta) : 1 \le i < \frac{n+1}{2} = j\right\} \\ &\cup \left\{W_{ij}(\theta) : 1 \le i < \frac{n}{2} < j \le n - i\right\}.\end{aligned}$$

*Each element is the $n \times n$ identity matrix, with the exception that the $i$th, $j$th, $n+1-i$th and $n+1-j$th rows and columns of $U_{ij}(\theta)$, $V_{ij}(\theta)$ and $W_{ij}(\theta)$ are (respectively)*

$$\begin{pmatrix} c & s & 0 & 0 \\ -s & c & 0 & 0 \\ 0 & 0 & c & -s \\ 0 & 0 & s & c \end{pmatrix}, \begin{pmatrix} \frac{c+1}{2} & \frac{s}{\sqrt{2}} & \frac{c-1}{2} \\ -\frac{s}{\sqrt{2}} & c & -\frac{s}{\sqrt{2}} \\ \frac{c-1}{2} & \frac{s}{\sqrt{2}} & \frac{c+1}{2} \end{pmatrix}, \begin{pmatrix} c & 0 & s & 0 \\ 0 & c & 0 & -s \\ -s & 0 & c & 0 \\ 0 & s & 0 & c \end{pmatrix}, \tag{3.32}$$

*where $c = \cos(\theta)$, $s = \sin(\theta)$.*

Let us look at these centrosymmetric Lie algebras and Lie groups for specific cases of $n$.

$$\mathfrak{sco}(3) = \left\{ \begin{pmatrix} 0 & \omega & 0 \\ -\omega & 0 & -\omega \\ 0 & \omega & 0 \end{pmatrix} : \omega \in \mathbb{R} \right\}, \tag{3.33}$$

$$\mathrm{SCO}(3) = K \begin{pmatrix} \mathrm{SO}(2) & 0 \\ 0 & 1 \end{pmatrix} K = \left\{ \begin{pmatrix} \frac{c+1}{2} & \frac{s}{\sqrt{2}} & \frac{c-1}{2} \\ -\frac{s}{\sqrt{2}} & c & -\frac{s}{\sqrt{2}} \\ \frac{c-1}{2} & \frac{s}{\sqrt{2}} & \frac{c+1}{2} \end{pmatrix} : c^2 + s^2 = 1 \right\}$$

$$\mathfrak{sco}(4) = K \begin{pmatrix} \mathfrak{so}(2) & 0 \\ 0 & \mathfrak{so}(2) \end{pmatrix} K = \left\{ \begin{pmatrix} 0 & \omega_1 & \omega_2 & 0 \\ -\omega_1 & 0 & 0 & -\omega_2 \\ -\omega_2 & 0 & 0 & -\omega_1 \\ 0 & \omega_2 & \omega_1 & 0 \end{pmatrix} : \omega_i \in \mathbb{R} \right\}$$

$$\begin{aligned} \mathrm{SCO}(4) &= K \begin{pmatrix} \mathrm{SO}(2) & 0 \\ 0 & \mathrm{SO}(2) \end{pmatrix} K \\ &= \left\{ \begin{pmatrix} c_1 & s_1 & 0 & 0 \\ -s_1 & c_1 & 0 & 0 \\ 0 & 0 & c_1 & -s_1 \\ 0 & 0 & s_1 & c_1 \end{pmatrix} \begin{pmatrix} c_2 & 0 & s_2 & 0 \\ 0 & c_2 & 0 & -s_2 \\ -s_2 & 0 & c_2 & 0 \\ 0 & s_2 & 0 & c_2 \end{pmatrix} : c_i^2 + s_i^2 = 1 \in \mathbb{R} \right\} \end{aligned}$$

### 3.2.2 Structure of bisymmetric isospectral manifolds

What does the isospectral manifold look like when it is restricted to contain only the bisymmetric matrices that have the same eigenvalues as a given bisymmetric matrix? We denote

$$\mathcal{BI}_{\mathcal{G}}(X_0) = \mathcal{I}_{\mathcal{G}}(X_0) \cap \mathrm{Bisym}(n), \tag{3.34}$$

for any $X_0 \in \mathrm{Bisym}(n)$ and any Lie subgroup $\mathcal{G} \subset \mathrm{SL}(n)$.

**Proposition 3.2.10.** *Let $Y_0 \in \mathrm{Bisym}(n)$. The tangent space $T_Y\mathcal{BI}_{\mathrm{SO}(n)}(Y_0)$ at any point $Y$ is the space*

$$T_Y\mathcal{BI}_{\mathrm{SO}(n)}(Y_0) = \left\{[A, Y] : A \in \mathrm{C}_{\mathfrak{sco}(n)}(Y)^{\perp}\right\}, \tag{3.35}$$

*where $\mathrm{C}_{\mathfrak{sco}(n)}(Y) = \{B \in \mathfrak{sco}(n) : [B, Y] = 0\}$ is the centraliser of $Y$ in $\mathfrak{sco}(n)$ and $\perp$ denotes the orthogonal complement in $\mathfrak{sco}(n)$ with respect to the Frobenius inner product.*

*Furthermore, there is a linear bijection,*

$$\mathrm{ad}_Y : \mathrm{C}_{\mathfrak{sco}(n)}(Y)^{\perp} \to T_Y\mathcal{BI}_{\mathrm{SO}(n)}(Y_0). \tag{3.36}$$

*Proof.* Since $\mathcal{BI}_{\mathcal{G}}(Y_0) = \mathcal{I}_{\mathcal{G}}(Y_0) \cap \mathrm{Bisym}(n)$, Proposition 2.4.1 shows that

$$T_X\mathcal{BI}_{\mathrm{SO}(n)}(Y_0) = \left\{[A, Y] : A \in \mathrm{C}_{\mathfrak{so}(n)}(Y)^{\perp}\right\} \cap \mathrm{Bisym}(n).$$

Suppose that $A \in \mathrm{C}_{\mathfrak{so}(n)}(Y)^{\perp}$ and $[A, Y] \in \mathrm{Bisym}(n)$. Then $[A, Y] = E[A, Y]E = [EAE, EXE] = [EAE, Y]$. Therefore $A = EAE + B$ for some $B \in \mathrm{C}_{\mathfrak{so}(n)}(Y)$. Note that $B - EBE \in \mathrm{C}_{\mathfrak{so}(n)}(Y)$ because $E$ also commutes with $Y$ because it is bisymmetric. Hence

$$\begin{aligned}\langle B, B\rangle_F &= \langle A - EAE, B\rangle_F \\ &= \langle A, EBE - B\rangle_F \\ &= 0\end{aligned}$$

Therefore, $B = 0$, so $A = EAE$. By Lemma 3.2.8 part $(ii)$, $A \in \mathfrak{sco}(n)$. Therefore $A \in \mathrm{C}_{\mathfrak{sco}(n)}(Y)^{\perp}$.

Conversely, if $A \in \mathrm{C}_{\mathfrak{sco}(n)}(Y)^{\perp}$ then $A \in \mathfrak{so}(n)$ and if $B \in C_{\mathfrak{so}}(n)(Y)$ then

$$\langle A, B\rangle_F = \frac{1}{2}\langle A, B + EBE\rangle_F = 0$$

since $B + EBE \in C_{\mathfrak{sco}}(n)(Y)$. Therefore $A \in \mathrm{C}_{\mathfrak{so}(n)}(Y)^{\perp}$ and $E[A, Y]E = [EAE, EXE] = [A, Y]$ because by Lemma 3.2.8 $A \in \mathrm{Centro}(n)$.

To complete the proof and show that $\mathrm{ad}_Y$ is a bijection, note that its kernel in $\mathrm{C}_{\mathfrak{g}}(Y)^{\perp}$ is zero by definition. □

**Theorem 3.2.11.** *Let $X, Y \in \mathrm{Bisym}(n)$. Then there exists $Q \in \mathrm{SCO}(n)$ such that $Y = QXQ^T$ if and only if $X$ and $Y$ have the same eigenvalues with the same parities (taking multiplicity into account).*

*Proof.* Using Table 3.1, we can write $KXK = \mathrm{diag}(X_1, X_2)$ and $KYK = \mathrm{diag}(Y_1, Y_2)$. Then, writing $Q = K\mathrm{diag}(Q_1, Q_2)K$ were $Q_1 \in \mathrm{SO}(p)$, $Q_2 \in \mathrm{SO}(q)$,

$$\begin{aligned}
\exists Q \in \mathrm{SCO}(n) \text{ s.t. } Y = Q^T X Q \iff & \quad \exists Q_1 \in \mathrm{SO}(p), Q_2 \in \mathrm{SO}(q) \text{ s.t.} \\
& \quad Y_i = Q_i^T X_i Q_i,\ i = 1, 2 \\
\iff & \quad X_i \text{ and } Y_i \text{ have the same} \\
& \quad \text{eigenvalues for } i = 1, 2 \\
\iff & \quad X \text{ and } Y \text{ have the same} \\
& \quad \text{eigenvalues with the same parities}
\end{aligned}$$

This last line follows from the fact the even eigenvalues of $X$ and $Y$ are the eigenvalues of $X_1$ and $Y_1$ respectively, and the odd eigenvalues those of $X_2$ and $Y_2$ respectively, as can be seen in the proof of Lemma 3.2.4. □

**Theorem 3.2.12.** *Let $Y_0 \in \mathrm{Bisym}(n)$ have $n$ distinct eigenvalues. Then the bisymmetric isospectral manifold*

$$\mathcal{BI}_{\mathrm{SO}(n)}(X_0) = \{Y \in \mathrm{Bisym}(n) : \mathrm{eigs}(Y) = \mathrm{eigs}(Y_0)\} \tag{3.37}$$

*has $\binom{n}{p}$ connected components, where $p = \lceil \frac{n}{2} \rceil$. The components are each acted upon transitively by $\mathrm{SCO}(n)$, and have dimension $\frac{1}{2}p(p-1) + \frac{1}{2}q(q-1)$.*

*Proof.* Let $X : \mathbb{R} \to \mathcal{BI}_{\mathrm{SO}(n)}(X_0)$ be a $C^1$ path. Then by Proposition 3.2.10, $\dot{X} = [A(t), X]$ for a $C^1$ function $A : \mathbb{R} \to \mathfrak{sco}(n)$. By 2.1.1, $X = QX_0Q^T$, where $\dot{Q} = AQ$, $Q(0) = I$. By Lemma 3.2.8, $Q \in \mathrm{SCO}(n)$. Hence by Theorem 3.2.11 the parity of the eigenvalues of $X(t)$ remains the same for all $t$. Since this path was arbitrary, all the matrices in a connected component have the same parity of eigenvalues.

Conversely, if $X, Y \in \mathcal{BI}_{\mathrm{SO}(n)}(Y_0)$ have the same eigenvalues with the same parity, then by Theorem 3.2.11 there exists $Q \in \mathrm{SCO}(n)$ such that $Y = QXQ^T$. Since $\mathrm{SCO}(n) = \exp(\mathfrak{sco}(n))$ by Lemma 3.2.8, there exists $A$ such that $\exp(A) = Q$. The path $Z(t) = \exp(tA)X\exp(-tA)$ is a continuous path connecting $X$ and $Y$, so they must be in the same connected component.

Therefore, each connected component is in one-to-one correspondence to the set of parity assignments for the eigenvalues.

The dimension of the manifold is that of $\mathrm{SCO}(n)$, which is $\dim(\mathrm{SO}(p))+\dim(\mathrm{SO}(q)) = \frac{1}{2}p(p-1)+\frac{1}{2}q(q-1)$. □

Theorem 3.2.12 can be restated as follows. For $Y \in \mathrm{Bisym}(n)$, the bisymmetric isospectral manifold can be expressed as

$$\mathcal{BI}_{\mathrm{SO}(n)}(Y) = \bigcup_{i=1}^{\binom{n}{p}} \mathcal{I}_{\mathrm{SCO}(n)}(Y_i), \tag{3.38}$$

where $Y_1, \ldots, Y_{\binom{n}{p}}$ are representatives of each connected component of $\mathcal{BI}_{\mathrm{SO}(n)}(Y)$. What is a sensible set of representatives?

**Definition 3.2.13** (Cross matrices)**.** We define the space of *cross matrices* to be the image of $\mathrm{diag}(p) \times \mathrm{diag}(q)$ under conjugation by $K$:

$$\mathrm{Cross}(n) = K \begin{pmatrix} \mathrm{diag}(p) & 0 \\ 0 & \mathrm{diag}(q) \end{pmatrix} K. \tag{3.39}$$

The non-zero structure of a cross matrix is like a cross. For example, for $n = 6$ and using $\mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$ and $\mathrm{diag}(\mu_1, \mu_2, \mu_3)$ in Definition 3.2.13,

$$\frac{1}{2}\begin{pmatrix} \lambda_1+\mu_1 & 0 & 0 & 0 & 0 & \lambda_1-\mu_1 \\ 0 & \lambda_2+\mu_2 & 0 & 0 & \lambda_2-\mu_2 & 0 \\ 0 & 0 & \lambda_3+\mu_3 & \lambda_3-\mu_3 & 0 & 0 \\ 0 & 0 & \lambda_3-\mu_3 & \lambda_3+\mu_3 & 0 & 0 \\ 0 & \lambda_2-\mu_2 & 0 & 0 & \lambda_2+\mu_2 & 0 \\ \lambda_1-\mu_1 & 0 & 0 & 0 & 0 & \lambda_1+\mu_1 \end{pmatrix}. \tag{3.40}$$

The eigenvalues of this matrix are $\lambda_1, \lambda_2, \ldots, \lambda_p, \mu_q, \mu_{q-1}, \ldots, \mu_1$. It is clear from the definition how to produce a cross matrix with prescribed eigenvalues.

The results of this section thus far allow us to solve more than a bisymmetric inverse eigenvalue problem. They allow us to solve a *parity assigned* bisymmetric inverse eigenvalue problem, in which we not only assign the eigenvalues, but what the parity of each eigenvalue should be. How this applies to the symmetric Toeplitz inverse eigenvalue problem and how Landau's Theorem relates to parity of eigenvalues will be discussed in Section 3.3.

### 3.2.3 $3 \times 3$ bisymmetric isospectral manifold

By Theorem 3.2.12, the $3 \times 3$ bisymmetric isospectral manifold consists of three connected components each acted upon transitively by SCO(3). In Figure 3.2 we see that this is indeed the case. Each colour of line represents a distinct set of 3 eigenvalues, each producing three disjoint 1-dimensional manifolds representing a different parity assignment. It is easy to see that there is a cylinder in the 3-dimensional space (which project onto two green circles, one of which is on the back of the sphere) such that all isospectral flows which begin inside this cylinder can never encounter a Toeplitz matrix. We compute an exact formula for this cylinder below.

We have shown above that SCO(3) is a 1-dimensional Lie group with Lie algebra generator

$$\Omega_1 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

Hence any bisymmetric isospectral flow is of the form

$$Y(t) = \exp(\omega(t)\Omega_1) Y_0 \exp(-\omega(t)\Omega_1), \tag{3.41}$$

for some function $\omega : \mathbb{R} \to \mathbb{R}$. Computing $Y(t)$ explicitly for the trace free cross matrix

$$Y_0 = \begin{pmatrix} x & 0 & y \\ 0 & -2x & 0 \\ y & 0 & x \end{pmatrix}, \tag{3.42}$$

we have that all $3 \times 3$ bisymmetric isospectral flows are of the form

$$Y(t) = \frac{x-y}{4} \begin{pmatrix} 1 & 0 & -3 \\ 0 & -2 & 0 \\ -3 & 0 & 1 \end{pmatrix} + \frac{3x+y}{4} \begin{pmatrix} c & -\sqrt{2}s & c \\ -\sqrt{2}s & -2c & -\sqrt{2}s \\ c & -\sqrt{2}s & c \end{pmatrix}, \tag{3.43}$$

where $c(t) = \cos(\omega(t))$ and $s(t) = \sin(\omega(t))$. If $Y(t)$ satisfies an autonomous first order system then so does $\omega(t)$ by Theorem B.3.9. Solutions to scalar autonomous first order system can only be nonincreasing or nondecreasing. Hence $Y(t)$ is either periodic if

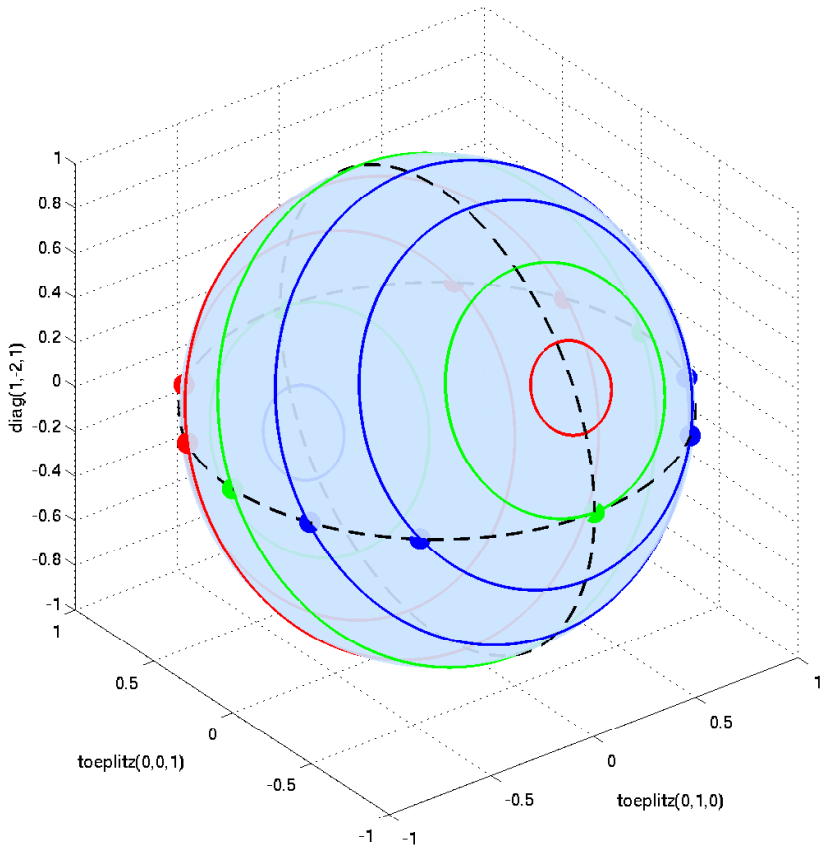



Fig. 3.2 This figure describes the structure of the bisymmetric isospectral manifolds for $3 \times 3$ matrices. The three axes represent the three degrees of freedom for a bisymmetric and trace-free $3 \times 3$ matrix. The surface of the blue sphere corresponds to the matrices with Frobenius norm 1; note that the surface is translucent so we can see lines drawn on its far side. Each line colour (red, blue, green) corresponds to a single bisymmetric isospectral manifold for a choice of 3 eigenvalues. Note that as predicted by Theorem 3.2.12, there are three connected components for each bisymmetric isospectral manifold, each corresponding to a choice of parity for the eigenvalues. Not every parity assignment gives a circle that passes through the $(x, y)$-plane (which represents Toeplitz matrices). Note for example the red circle on the right: any algorithm constrained to lie in that red circle can never encounter a Toeplitz matrix (labeled by a large dot).

$\omega(t)$ is nonconvergent, or $Y(t)$ is convergent in the case that $\omega(t)$ is convergent. The dynamics are very simple.

Note that *all* $3 \times 3$ bisymmetric isospectral flows have any multiple of the following as a fixed point,

$$\begin{pmatrix} 1 & 0 & -3 \\ 0 & -2 & 0 \\ -3 & 0 & 1 \end{pmatrix}. \tag{3.44}$$

### 3.2.4 $4 \times 4$ bisymmetric isospectral manifold

By Theorem 3.2.12, the $4 \times 4$ bisymmetric isospectral manifold consists of six connected components each acted upon transitively by SCO(4).

We have shown above that SCO(4) is a 2-dimensional Lie group with Lie algebra generators

$$\Omega_1 = \begin{pmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & -1 \\ -1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \end{pmatrix}, \quad \Omega_2 = \begin{pmatrix} 0 & 1 & -1 & 0 \\ -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 \end{pmatrix}.$$

Hence any bisymmetric isospectral flow is of the form

$$Y(t) = \exp(\omega_2(t)\Omega_2) \exp(\omega_1(t)\Omega_1) Y_0 \exp(-\omega_1(t)\Omega_1) \exp(\omega_2(t)\Omega_2), \tag{3.45}$$

for some functions $\omega_i : \mathbb{R} \to \mathbb{R}$. It is readily observed that $[\Omega_1, \Omega_2] = 0$, by the isomorphism $\mathrm{SO}(2) \times \mathrm{SO}(2) \cong \mathrm{SCO}(4)$. Hence the ordering of the similarity transformations does not matter.

### 3.2.5 Parity in general

How do these results and principles generalise beyond bisymmetric matrices? Theorem 3.2.11 and Theorem 3.2.12 can be generalised to the following without changing the mechanics of the proofs.

**Definition 3.2.14.** Let $\Sigma \in \mathrm{GL}(n)$ be an involution, i.e. $\Sigma^2 = I$. Equivalently all of its eigenvalues are $\pm 1$. The signature of $\sigma$ is the unique pair of integers $(p, q)$ such that $\sigma$ has a $p$-dimensional 1-eigenspace and $q$-dimensional $-1$-eigenspace.

*Remark* 3.2.15. The exchange matrix $E \in \mathrm{GL}(n)$ is an involution with signature $(\lceil \frac{n}{2} \rceil, \lfloor \frac{n}{2} \rfloor)$.

**Theorem 3.2.16.** *Let $\Sigma \in \mathrm{GL}(n)$ be an involution with signature $(p,q)$ and let $X \in \mathfrak{gl}(n)$ be such that $\Sigma X \Sigma = X$. Define $K \in \mathrm{SL}(n)$ to have the first $p$ columns some $p$ normalised linearly independent $1$-eigenvectors of $\Sigma$ and the last $q$ columns some normalised linearly independent $-1$-eigenvectors of $\Sigma$. Then*

$$K^{-1} X K = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}, \tag{3.46}$$

*where $X_1 \in \mathfrak{gl}(p)$, $X_2 \in \mathfrak{gl}(q)$.*

**Theorem 3.2.17.** *Let $\mathcal{G}$ be a connected Lie subgroup of $\mathrm{SL}(n)$ and $\mathcal{G}_\Sigma$ be the identity component of $\{Q \in \mathcal{G} : \Sigma Q \Sigma = Q\}$. Now let $X \in \mathfrak{gl}(n)$ have distinct eigenvalues and satisfy $\Sigma X \Sigma = X$. Then the constrained $\mathcal{G}$-Adjoint orbit of $X$,*

$$\mathcal{I}_{\mathcal{G}}(X) \cap \{Y \in \mathfrak{gl}(n) : \Sigma Y \Sigma = Y\}, \tag{3.47}$$

*has $\binom{n}{p}$ connected components, where $(p,q)$ is the signature of $\Sigma$, and each component is acted upon by $\mathcal{G}_\Sigma$. Hence the dimension of each component is $\dim(\mathcal{G}_\Sigma)$.*

## 3.3 Bisymmetric isospectral flows for Toeplitz inverse eigenvalue problems

Let us inspect Landau's Theorem more closely [Lan94].

**Definition 3.3.1** (Alternating parity)**.** Let a bisymmetric matrix have even eigenvalues denoted $\lambda_i$ and odd eigenvalues denoted $\mu_i$. The matrix has eigenvalues with alternating parity if the eigenvalues can be arranged in ascending order with parity alternating and the largest even, like

$$\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \mu_2 \leq \cdots \leq \mu_q \leq \lambda_p,$$

or

$$\mu_1 \leq \lambda_1 \leq \lambda_2 \leq \mu_2 \leq \cdots \leq \mu_q \leq \lambda_p$$

where $p = \lceil \frac{n}{2} \rceil$ and $q = n - p$.

**Definition 3.3.2** (Regular Toeplitz matrix)**.** A symmetric Toeplitz matrix is *regular* if it has distinct eigenvalues with alternating parity as above, and every principal submatrix also has these two properties.

Landau showed that the map taking *regular* Toeplitz matrices to their eigenvalues has topological degree 1 modulo 2.

**Theorem 3.3.3** (Landau [Lan94])**.** *Any vector $\boldsymbol{\lambda} \in \mathbb{R}^n$ is the spectrum of an $n \times n$ real symmetric Toeplitz matrix. If the eigenvalues are distinct then there exists a regular Toeplitz matrix with those eigenvalues.*

We derive an apparently new characterisation of existence for the $3 \times 3$ problem.

**Theorem 3.3.4.** *There exists a $3 \times 3$ real symmetric Toeplitz matrix with even eigenvalues $\lambda_1, \lambda_2$ and odd eigenvalue $\mu_1$ if and only if*

$$(\lambda_1 - \mu_1)(\mu_1 - \lambda_2) + 2(\lambda_1 - \lambda_2)^2 \geq 0.$$

*Proof.* Consider the exact orbit the $3 \times 3$ *trace-free* bisymmetric isospectral flow shown in Subsection 3.2.3. This orbit will pass through a Toeplit matrix if and only if $(3x + y)c = x - y$ for some cosine $c$, which is equivalent to $x(x + y) \geq 0$. Using the formula for a cross matrix in terms of its eigenvalues, $x = -\frac{1}{2}\lambda_2$ and $x + y = \frac{1}{2}(\lambda_1 + \mu_1) + \frac{1}{2}(\lambda_1 - \mu_1) = \lambda_1$. Hence the *trace-free* orbit passes through a Toeplitz matrix if and only if $\lambda_1\lambda_2 \leq 0$.

In order to make this apply to non-trace-free orbits, we must make the formula $\lambda_1\lambda_2$ translation invariant. Subtract $\frac{1}{3}(\lambda_1 + \lambda_2 + \mu_1) = 0$ to obtain the condition

$$\left(\lambda_1 - \frac{1}{3}(\lambda_1 + \lambda_2 + \mu_1)\right)\left(\lambda_2 - \frac{1}{3}(\lambda_1 + \lambda_2 + \mu_1)\right) \leq 0.$$

This is now translation invariant and reduces to the condition stated in the theorem. □

In the $3 \times 3$ case, Landau's Theorem states that there exists a symmetric Toeplitz matrix with even eigenvalues $\lambda_1, \lambda_2$ and odd eigenvalues $\mu_1$ if $(\lambda_1 - \mu_1)(\mu_1 - \lambda_2) \geq 0$. Therefore, what we have proved here is stronger.

By Theorem 3.2.10, any isospectral flow which preserves bisymmetry is of the form

$$\dot{Y} = [A(t, Y), Y], \tag{3.48}$$

where $A(t, Y)$ takes values in the Lie algebra $\mathfrak{sco}(n)$. By Theorem 3.2.12, the parity of the eigenvalues remains constant under such a flow. This leads us to an algorithm for solving the symmetric Toeplitz inverse eigenvalue problem. If we are given eigenvalues

$\lambda_1 < \lambda_2 < \ldots < \lambda_n$, then assign them alternating parity as Definition 3.3.1. Construct a cross matrix with those eigenvalues and the desired parity by constructing

$$K \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_p & & & \\ & & & \mu_q & & \\ & & & & \ddots & \\ & & & & & \mu_1 \end{pmatrix} K \tag{3.49}$$

as in Table 3.1. Then numerically solve equation (3.48) with a geometric integrator.

### 3.3.1 Bisymmetric isospectral gradient flows

We must ensure that the gradient flow we discussed in Subsection 3.1.1 preserves bisymmetry. Showing it is of the form in equation (3.48) will suffice. For any $X \in \mathbb{R}^{n\times n}_{\text{sym}}$, for $E$ the exchange matrix as in Definition 3.27,

$$E\left[P_T(X), X\right]E = \left[EP_T(X)E, EXE\right] = \left[P_T(EXE), EXE\right],$$

by inspection of how $P_T$ acts on the entries of $X$. Hence

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}EXE &= E\left[\left[P_T(X), X\right], X\right]E \\ &= \left[\left[P_T(EXE), EXE\right], EXE\right].\end{aligned}$$

Therefore, if $EX(0)E = X(0)$, by uniqueness of solutions to smooth initial value problems, $X(t)$ is bisymmetric for all $t$.

For the $3\times 3$ case, we have reduced the problem from a 6-dimensional flow in $\mathbb{R}^{n\times n}_{\text{sym}}$ to a 1-dimensional flow in $\mathfrak{sco}(3)$. Figure 3.2 shows that the main challenge is ensuring we begin the flow on a connected component of the manifold that contains a Toeplitz matrix, which can be guaranteed by ensuring alternating parity of the eigenvalues or using Theorem 3.3.4.

For the $4\times 4$ case, we reduced the problem considerably, from a 10-dimensional flow in $\mathbb{R}^{n\times n}_{\text{sym}}$ to a 2-dimensional flow in $\mathfrak{sco}(4)$. Using Lemma B.3.9, we have that

$$X(t) = \exp(\Omega(t))X_0\exp(-\Omega(t)),$$

where $\Omega(t) \in \mathfrak{sco}(4)$ satisfies

$$\dot{\Omega}(t) = \mathrm{d}\exp^{-1}_{\Omega(t)}([P_T(X), X]), \quad \Omega(0) = O.$$

However, since $\mathfrak{sco}(4)$ is commutative, $\mathrm{d}\exp^{-1}_{\Omega(t)}$ is the identity operator, so we have the equation

$$\dot{\Omega}(t) = [P_T(X), X]\,.$$

Translating this flow into a flow on the torus $\mathbb{T}^2$ is a simple matter of complicated algebra best left to computer algebra software.

Using a parametrisation of $\mathfrak{sco}(4)$ by $\mathbb{T}^2$ are Figure 3.3, Figure 3.4 and Figure 3.5, all demonstrating the values of $\Psi(Y) = \frac{1}{2}\|Y - P_T(Y)\|_F^2$ on the bisymmetric isospectral manifold. What is actually displayed is $f(\omega_1, \omega_2)$, where

$$f(\omega_1, \omega_2) = \Psi\left(\exp(\omega_2\Omega_2)\exp(\omega_1\Omega_1)X_0\exp(-\omega_1\Omega_1)\exp(-\omega_2\Omega_2)\right), \tag{3.50}$$

where $Y_0$ is a cross matrix with given parity-assigned eigenvalues, and

$$\mathfrak{sco}(4) = \mathrm{Span}\left\{\begin{pmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & -1 \\ -1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & -1 & 0 \\ -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 \end{pmatrix}\right\} = \mathrm{Span}\left\{\Omega_1, \Omega_2\right\}$$

Note that for all figures, Figure 3.3, Figure 3.4 and Figure 3.5, the matrices represented by a 2D point are on the connected component of the bisymmetric isospectral manifold containing matrices with alternating parity. This is done simply by choosing $Y_0$ with alternating parity.

In Figure 3.3, Figure 3.4 and Figure 3.5 the height of $f(\omega_1, \omega_2)$ at each point is represented by a colour. Red means high and blue means low (close to zero). The stationary points are denoted using coloured dots. Yellow dots are the stationary points that are Toeplitz and the red dots are the non-Toeplitz stationary points. The point with a yellow asterisk is a regular Toeplitz matrix as in Definition 3.3.2. The gradient flow will follow the normals of the contour lines.

*Remark* 3.3.5. Interestingly, in Figure 3.3, Figure 3.4 and Figure 3.5, the position of the regular Toeplitz matrix in the Lie algebra appears quite stable to changes in the eigenvalues. This suggests that it may be possible to further constrain the set of potential solutions in order to home in on a regular Toeplitz matrix. For the

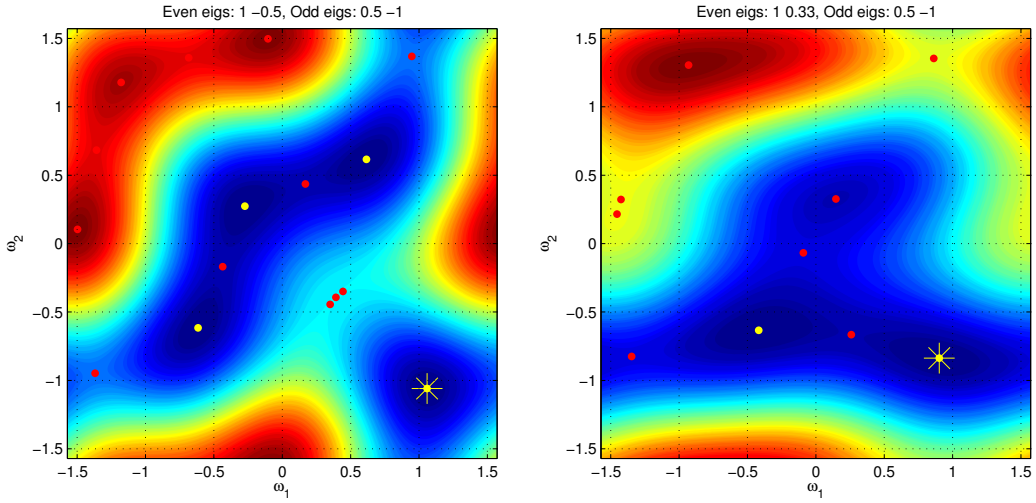



Fig. 3.3 Each contour plot shows the projective Toeplitz error (see equation (3.50)) for an assignment of 2 even eigenvalues and 2 odd eigenvalues to a $4 \times 4$ bisymmetric matrix (written in the title of the plot). In the left plot we see 4 yellow dots representing Toeplitz matrices, of which one is asterisked to represent a regular Toeplitz matrix. Furthermore, all stable stationary points are Toeplitz. In the right plot we see 2 yellow dots representing Toeplitz matrices, one of which is regular. There is now one stable non-Toeplitz stationary point represented by a red dot in a convex blue region, which is problematic for a gradient flow approach. The difference between the two situations appears to be related to how evenly distributed the eigenvalues are.

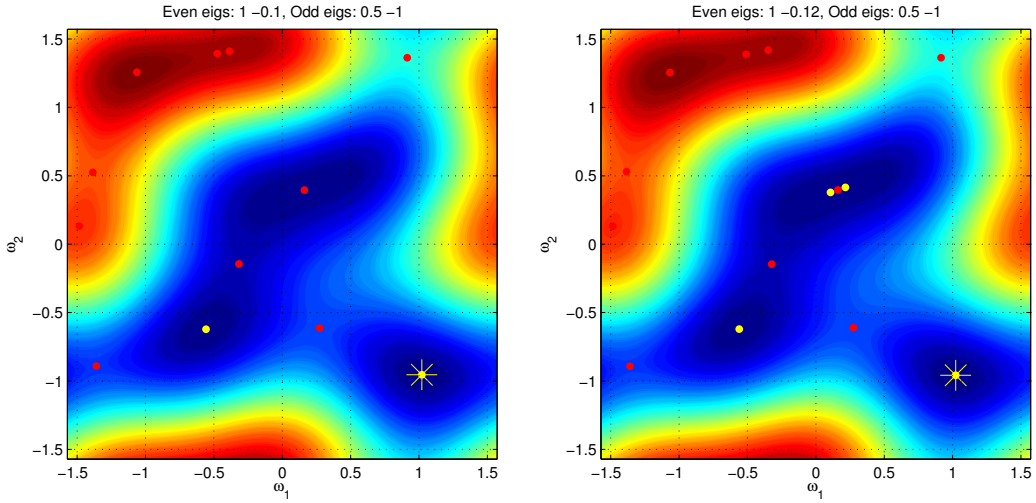



Fig. 3.4 Each contour plot shows the projective Toeplitz error (see equation (3.50)) for an assignment of 2 even eigenvalues and 2 odd eigenvalues to a $4 \times 4$ bisymmetric matrix (written in the title of the plot). One eigenvalue is slightly different between the two plots, and this slight change merges of two Toeplitz (yellow) solutions into a single non-Toeplitz stable stationary point (red).

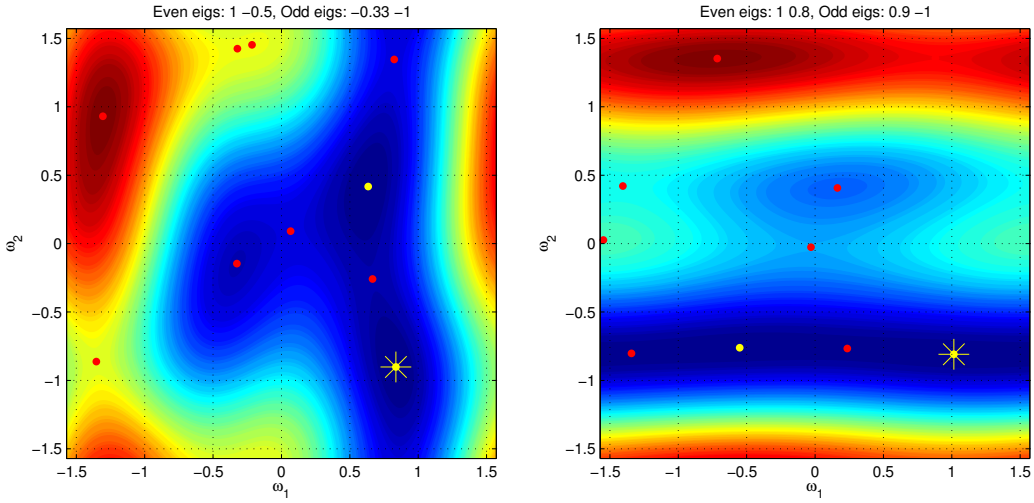



Fig. 3.5 Each contour plot shows the projective Toeplitz error for an assignment of 2 even eigenvalues and 2 odd eigenvalues to a $4 \times 4$ bisymmetric matrix. Here we demonstrate what happens to the error function as eigenvalues get closer together. In the right image, when 3 of the 4 eigenvalues begin to coalesce, the function becomes degenerate.

$4 \times 4$ case we have treated here, clearly starting an isopectral gradient flow in the bottom-right quadrant is the go-to strategy, but any information of where to begin for higher dimensional problems has not been studied.

### 3.3.2 Bisymmetric Chu's flow

One advantage of Chu's flow given in Section 3.1.2 is that it preserves bisymmetry. It has the following symmetry property [Chu93], [DS99],

$$EB(Y)E = B(EYE) \text{ for all } Y \in \mathrm{S}_n, \tag{3.51}$$

where $E$ is the exchange matrix (see Definition 3.27). It follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}(EYE) = [B(EYE), EYE]. \tag{3.52}$$

Therefore, if $EY(0)E = Y(0)$ then $Y(t)$ is bisymmetric throughout the flow by uniqueness of solution to smooth initial value problems.

In Section 3.1.2, we mentioned that Chu's flow has periodic solutions, but that these appear to be unstable, as demonstrated by Figure 3.1. However, this appears to contradict Figure 3.2, including the exact solution found in that section, which bring us to the conclusion that a periodic orbit of $3 \times 3$ bisymmetric matrices is perfectly

stable because the parameter space is 1-dimensional. The reason for the instability behaviour in Figure 3.1 is that in that numerical simulation, bisymmetry is not being preserved. In Figure 3.6 we show an example of a bisymmetric $4 \times 4$ numerical solution to Chu's flow which, when restricted to bisymmetric matrices by design, is periodic, but when errors in bisymmetry are allowed to creep in, the periodic orbit is left and the solution quickly converges to a Toeplitz matrix. The initial datum is

$$Y_0 = \begin{pmatrix} 0.1336 & 0 & 0 & 0.5669 \\ 0 & -0.1336 & 0.3780 & 0 \\ 0 & 0.3780 & -0.1336 & 0 \\ 0.5669 & 0 & 0 & 0.1336 \end{pmatrix}. \tag{3.53}$$

The reasoning behind choosing this initial datum is that we simple want to have a bisymmetric matrix whose eigenvalues *do not* alternate in parity as in Definition 3.3.1. We have also scaled and shifted the matrix so that it is trace free and has Frobenius norm 1. This appears to be a very simple method to find periodic solutions to Chu's flow even for higher dimensions. However, as was just discussed and is demonstrated in Figure 3.6, if the numerical solution is not constrained to stay bisymmetric by design, then this periodic orbit is left and a convergent trajectory is found.

Figure 3.6 is interesting from the point of view of geometric integration. The bottom solution is a worse solution to the isospectral flow than the top solution because it fails to preserve the bisymmetry of the analytical solution and diverges from the periodic orbit rapidly. However, because the eigenvalues are preserved by the geometric integrator used in both cases (except for rounding errors), the bottom solution is a better solution, because it computes a more accurate approximation to a symmetric Toeplitz matrix with the prescribed eigenvalues, unlike the periodic orbit of the top solution. There is a distinction between solving the flow accurately and solving the original problem accurately.

Does this mean we should not restrict Chu's flow to bisymmetric matrices because it would have stable periodic solutions? Figure 3.7 suggests that Chu's flow does not have periodic orbits if we constrain the flow to the connected component of the bisymmetric isospectral manifold such that the parity of the eigenvalues alternates as in Definition 3.3.1. This is done by choosing a cross matrix with alternating parity for the initial datum. In this case Chu's flow appears to converge to a regular Toeplitz matrix.

The numerical solutions in Figure 3.7 were computing by solving the Lie algebra equation (see B.3.9) using Matlab's `ode45` command.

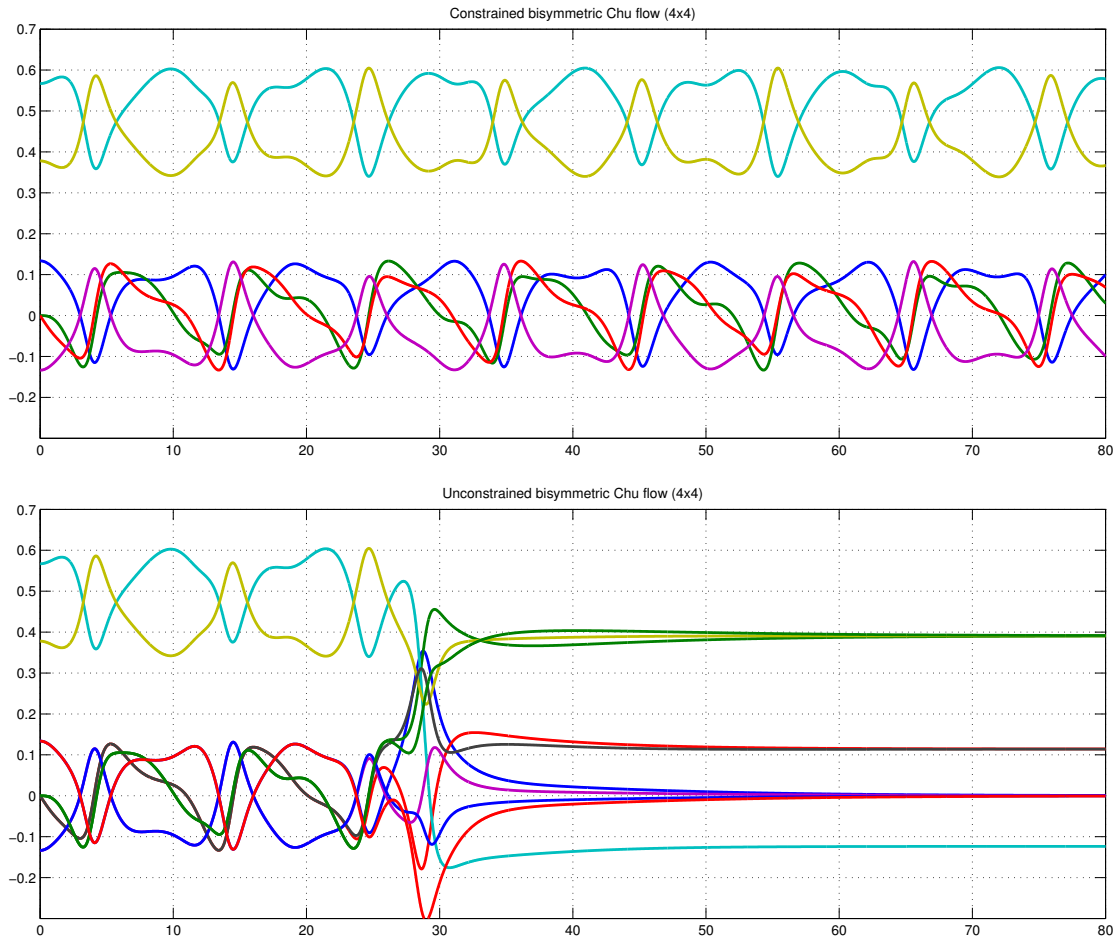



Fig. 3.6 These two figures are numerical solutions to Chu's flow (equation (3.20)) solved using a Cayley Forward Euler method (equation (3.24)), both with a stepsize of $h = 0.001$, and initial condition given in equation (3.53). Each line is an entry of the matrix. In exact arithmetic, the trajectory is periodic with period approximately 1.047 (the periodicity is proven in Subsection 3.2.3). In the top plot, the bisymmetry of the solution is preserved by the numerical method (by design at each step), and the periodic dynamics are respected. However, in the bottom plot bisymmetry is not preserved, merely symmetry, and the numerical solution quickly leaves the orbit due to discretisation and rounding errors. If the step size $h$ is changed then the time $t = hn$ at which the flow leaves the orbit changes. We emphasise that the eigenvalues do are not alternating for this example.

In Figure 3.7 we see that when Chu's flow is not only restricted to bisymmetric matrices, but further restricted to bisymmetric matrices whose eigenvalues have alternating parity, there are no periodic orbits. The flow is clearly not a gradient flow, however, because we can see a curl component in the vector field around the attractive fixed points.

The following is the main open problem in the area.

**Conjecture 3.3.6.** *Given $n$ distinct real eigenvalues, Chu's flow is convergent to a Toeplitz matrix for all initial data which are bisymmetric and whose eigenvalues alternate in parity.*

## 3.4 Computability

For practical purposes, the author recommends using a bisymmetric Chu flow starting from a cross matrix with alternating parity in order to solve the symmetric Toeplitz inverse eigenvalue problem numerically. In practice this has been empirically observed to converge without fail, by this author and the authors of [Chu93], [DS99], [DS02], and [DS03].

However, Chu's flow from these initial data has not yet been proved to converge. The following simple but impractical approach shows that at least in principal there exists a convergent algorithm.

We define for a given matrix $Y_0 \in \mathrm{Bisym}(n)$ whose eigenvalues alternate in parity as in Definition 3.3.1, we parametrise the set of all bisymmetric matrices with the same eigenvalues as $Y_0$ with the same parities by

$$Y : [-\pi, \pi)^m \to \mathrm{Bisym}(n), \quad m = \frac{1}{2}p(p-1) + \frac{1}{2}q(q-1), \quad p = \left\lceil \frac{n}{2} \right\rceil, q = n - p, \tag{3.54}$$

$$Y(\boldsymbol{\omega}) = \exp(\omega_1 \Omega_1) \cdots \exp(\omega_m \Omega_m) X_0 \exp(-\omega_m \Omega_m) \cdots \exp(-\omega_1 \Omega_1), \tag{3.55}$$

where $\Omega_1, \ldots, \Omega_m$ is a basis of $\mathfrak{sco}(n)$ (See Lemma 3.2.7) such that each basis element is a rank 2 matrix with nonzero eigenvalues $\pm i$. This implies that $\{\exp(t\Omega_j) : t \in [-\pi, \pi)\} = \mathfrak{sco}(n)$. This parametrisation covers all matrices with the same eigenvalues with the same parities by Theorem 3.2.11.

**Theorem 3.4.1** (An impractical but convergent algorithm)**.** *The following algorithm, which takes as input $\boldsymbol{\lambda} \in \mathbb{R}^n$, produces bisymmetric matrices $Y_0, Y_1, \ldots$ such that $\sigma(Y_k) = \boldsymbol{\lambda}$ for all $k$ and $\mathrm{dist}(Y_k, \mathcal{T}_{\boldsymbol{\lambda}}) \to 0$ as $k \to \infty$ where $\mathcal{T}_{\boldsymbol{\lambda}}$ is the set of real symmetric Toeplitz matrices with spectrum $\boldsymbol{\lambda}$.*

1. *Assign $p$ even eigenvalues $\lambda_1, \ldots, \lambda_p$ and $q$ odd eigenvalues $\mu_1, \ldots, \mu_q$ so that they alternate in parity: $\lambda_p \geq \mu_q \geq \lambda_{p-1} \geq \mu_{q-1} \geq \ldots$, as in Definition 3.3.1.*
2. *Construct a cross matrix using these parity-assigned eigenvalues as in Definition 3.2.13 and equation (3.40), and set $X_0$ to be this matrix.*

*3.* **for** $k = 1, 2, \ldots$ **do**
*4.* $\quad G_k \leftarrow \left\{\pi 2^{-k}(j_1, \ldots, j_m)^T : j_i = -2^k + 1, \ldots, 2^k\right\}$
*5.* $\quad \omega_k \leftarrow \arg\min_{\boldsymbol{\omega} \in G_k} \frac{1}{2}\|Y(\boldsymbol{\omega}) - P_T(Y(\boldsymbol{\omega}))\|_F^2,$
*6.* $Y_k \leftarrow Y(\omega_k)$
*where in the case that there are multiple minimal values for the argmin, the standard lexicographic ordering of the set $G_k$ determines a unique minimiser by taking the first in the ordering.*

*Proof.* Let us write $\Psi(Y) = \frac{1}{2}\|Y - P_T(Y)\|_F^2$. By Landau's Theorem that there exists real symmetric Toeplitz matrix with $\boldsymbol{\lambda}$ as its spectrum, and with alternating parity. By the parametrisation in equation (3.54), there exists $\boldsymbol{\omega} \in [-\pi, \pi)^m$ such that $\Psi(X(\boldsymbol{\omega})) = 0$.

Since $G_{k+1} \subset G_k$, we have $\Psi(Y_{k+1}) \leq \Psi(Y_k)$. Also, by continuity of $\Psi$,

$$\begin{aligned}\inf_{k\in\mathbb{N}} \Psi(Y_k) &= \inf_{k\in\mathbb{N}} \inf_{\boldsymbol{\omega}\in G_k} \Psi(Y(\boldsymbol{\omega})) \\ &= \inf_{\boldsymbol{\omega}\in[-\pi,\pi)^m} \Psi(Y(\boldsymbol{\omega})) \\ &= 0.\end{aligned}$$

Therefore $\Psi(Y_k) \to 0$ as $k \to \infty$. Suppose for a contradiction that there exists $\varepsilon > 0$ and a subsequence $Y_{k_1}, Y_{k_2}, \ldots$ such that $\operatorname{dist}(Y_{k_j}, \mathcal{T}_{\boldsymbol{\lambda}}) \geq \varepsilon$ for all $j$. Since $[-\pi, \pi)^m$ is compact there exists a further subsequence which converges to $\tilde{Y}$. By continuity of $\Psi$, we have $\Psi(\tilde{Y}) = 0$, so $\tilde{Y}$ is a Toeplitz matrix. By continuity of eigenvalues, $\tilde{Y} \in \mathcal{T}_{\boldsymbol{\lambda}}$, which is a contradiction. □

**Corollary 3.4.2.** *The Solvability Complexity Index of the inverse eigenvalue problem for real symmetric Toeplitz matrices is equal to 1 (see [BAHNS15a] and Section 4.5).*

Note that what we have *not* shown here is that there exists an algorithm with *error control* (see [BAHNS15a] and Section 4.5), which gives a guarantee for any $\varepsilon > 0$ to find an $k$ such that there exists a Toeplitz matrix $T$ such that $\sigma(T) = \sigma(Y_k) = \sigma(Y_0)$ and $\|Y_k - T\|_F < \varepsilon$.

However, if the purpose of the computation is to find a matrix $X$ with prescribed spectrum such that $\|Y - P_T(Y)\|_F^2 < \varepsilon$ (which does not necessarily imply that an isospectral real symmetric Toeplitz is near to $Y$) then of course the algorithm can be terminated only when this condition is satisfied and there is error control.

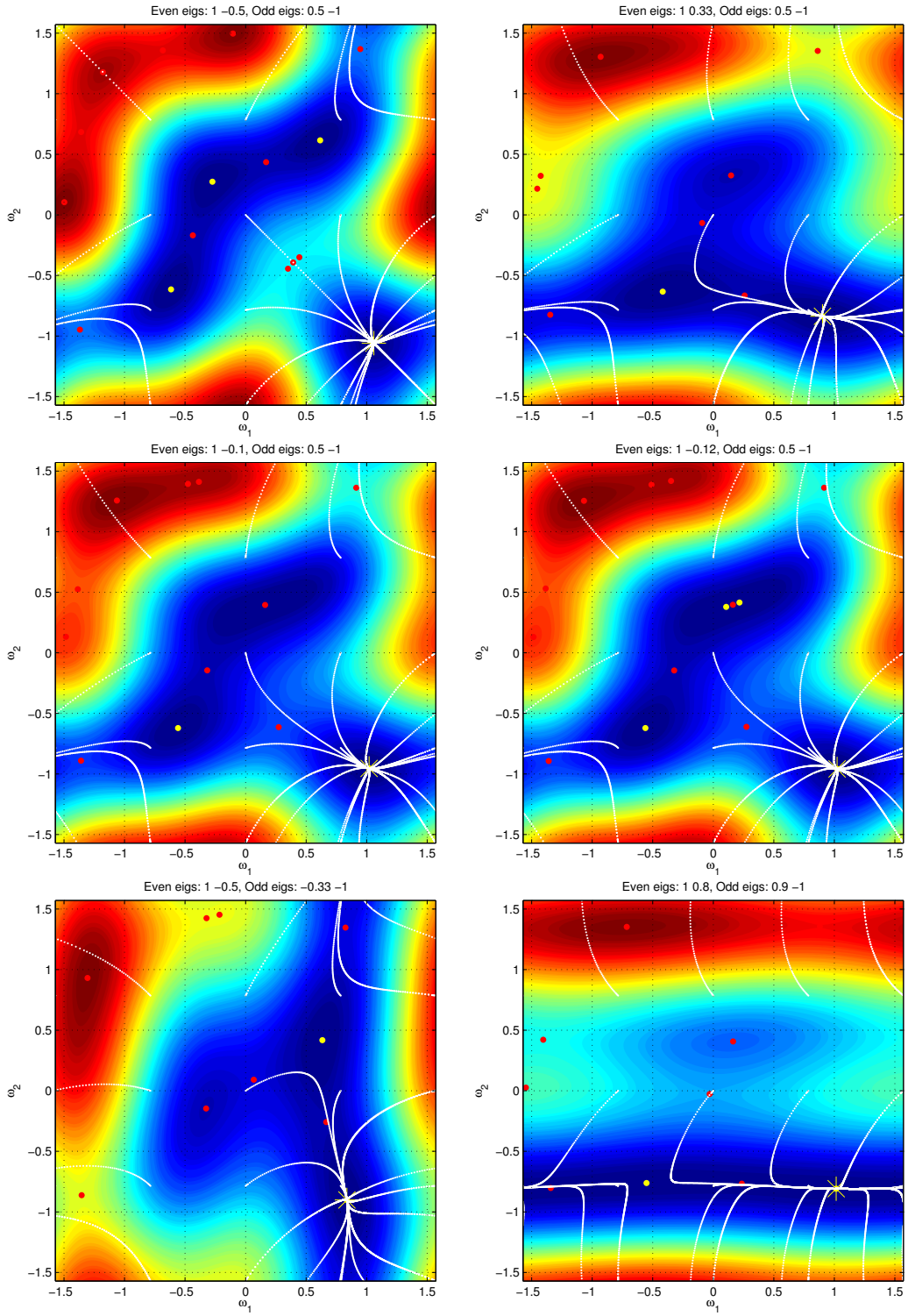



Fig. 3.7 These contour plots are the same as those in Figure 3.3, Figure 3.4, Figure 3.5, except here we have added numerical solutions to Chu's flow starting at points on an equispaced grid of 16 points on the torus. In all situations, the flow converges to the regular Toeplitz matrix in the bottom right of the plot.

画蛇添足，弄巧成拙

(huà shé tiān zú, nòng qiǎo chéng zhuō)

Lit. Drawing legs on a snake turns one from clever to foolish
Fig. To ruin something by adding superfluous parts

# Chapter 4

# Spectra of Jacobi operators via connection coefficients

A (bounded) Jacobi operator is a selfadjoint operator on $\ell^2 = \ell^2(\{0, 1, 2, \ldots\})$, which with respect to the standard orthonormal basis $\{e_0, e_1, e_2, \ldots\}$ has a tridiagonal matrix representation

$$J = \begin{pmatrix} \alpha_0 & \beta_0 & & \\ \beta_0 & \alpha_1 & \beta_1 & \\ & \beta_1 & \alpha_2 & \ddots \\ & & \ddots & \ddots \end{pmatrix}, \tag{4.1}$$

where $\alpha_k$ and $\beta_k$ are real numbers with $\beta_k > 0$. The spectral theorem for Jacobi operators guarantees the existence of a probability measure $\mu$ supported on the spectrum $\sigma(J) \subset \mathbb{R}$, called the *spectral measure*, and a unitary operator $U : \ell^2 \to L^2_\mu(\mathbb{R})$ such that

$$UJU^*[f](s) = sf(s), \tag{4.2}$$

for all $f \in L^2_\mu(\mathbb{R})$ [Dei00]. The coefficients $\alpha_k$ and $\beta_k$ are the three-term-recurrence coefficients of the *orthonormal polynomials* $P_k(s)$ with respect to $\mu$.

In this chapter we show that the computation and theoretical study of the spectra and spectral measure of a Jacobi operator $J$ which is a structured perturbation of another Jacobi operator $D$ whose spectral theory is known, can be conducted using the *connection coefficient matrix* between $J$ and $D$.

**Definition 4.0.1.** Denote the space of complex-valued sequences with finitely many nonzero elements by $\ell_{\mathcal{F}}$, and its algebraic dual, the space of all complex-valued sequences, by $\ell^\star_{\mathcal{F}}$.

Suppose that $D$ is a second, bounded Jacobi operator, and let $Q_k(s)$ denote its orthonormal polynomials. The connection coefficient matrix between $J$ and $D$ is defined as follows.

**Definition 4.0.2.** Define the *connection coefficient matrix* $C = C_{J\to D} = (c_{ij})_{i,j=0}^{\infty}$ to be the upper triangular matrix representing the change of basis between $(P_k)_{k=0}^{\infty}$ and $(Q_k)_{k=0}^{\infty}$ in the following manner:

$$P_k(s) = c_{0k}Q_0(s) + c_{1k}Q_1(s) + \cdots + c_{kk}Q_k(s). \tag{4.3}$$

Note that $C : \ell_{\mathcal{F}} \to \ell_{\mathcal{F}}$ as it is upper triangular and $C^T : \ell_{\mathcal{F}}^\star \to \ell_{\mathcal{F}}^\star$ as it is lower triangular, and thus we may write

$$\begin{pmatrix} P_0(s) \\ P_1(s) \\ P_2(s) \\ \vdots \end{pmatrix} = C^T \begin{pmatrix} Q_0(s) \\ Q_1(s) \\ Q_2(s) \\ \vdots \end{pmatrix} \quad \text{for all } s \in \mathbb{C}. \tag{4.4}$$

Connection coefficient matrices have been well-studied [Ask75, GM09]. Nevertheless, it does not appear to have been noted that the connection coefficients are relevant and useful in the spectral theory of Jacobi operators. Because $C$ is upper triangular with non-zero diagonal it is in fact an invertible operator from $\ell_{\mathcal{F}}$ to $\ell_{\mathcal{F}}$ (the inverse of $C_{J\to D}$ is $C_{D\to J}$). We show that as operators on $\ell_{\mathcal{F}}$, the connection coefficients matrix satisfies

$$J = C^{-1}DC. \tag{4.5}$$

Consequently, when $C$ is a bounded and invertible operator on $\ell^2$, we have $\sigma(J) = \sigma(D)$. More significantly, we further show that when $C$ is neither bounded nor invertible, the matrix entries are still informative about the spectra of $J$ and $D$. For example, if we let $\nu$ denote the spectral measure for $D$, the connection coefficients matrix $C = C_{J\to D}$ determines the existence and certain properties of the Radon–Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ (see Appendix D.1). We prove the following:

- Proposition 4.2.5: $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^2_\mu(\mathbb{R})$ if and only if the first row of $C$ is an $\ell^2$ sequence, in which case
$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = \sum_{k=0}^{\infty} c_{0,k}P_k. \tag{4.6}$$

More generally, if $\sum_{k=0}^{\infty} c_{0,k} P_k$ defines an $L^1_\mu(\mathbb{R})$ function (with the series converging at least in the probabilists' weak sense) then that function is precisely $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.

- Proposition 4.2.8: $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^\infty_\mu(\mathbb{R})$ if and only if $C$ is a bounded operator on $\ell^2$, in which case
$$\|C\|_2^2 = \operatorname*{ess\,sup}_{s\in\sigma(J)} \left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(s)\right|. \tag{4.7}$$
- Corollary 4.2.9: both $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^\infty_\mu(\mathbb{R})$ and $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \in L^\infty_\nu(\mathbb{R})$ if and only if $C$ is bounded and invertible on $\ell^2$.

In this chapter we pay particular attention to the case where $D$ is the so-called *free Jacobi operator*,
$$\Delta = \begin{pmatrix} 0 & \frac{1}{2} & & \\ \frac{1}{2} & 0 & \frac{1}{2} & \\ & \frac{1}{2} & 0 & \ddots \\ & & \ddots & \ddots \end{pmatrix}, \tag{4.8}$$
and $J$ is a Jacobi operator of the form $J = \Delta + K$, where $K$ is compact. This follows many other studies of this class of operators such as those in [DS06a, DS06b, DN86, DE15, GNR16, GC80, KS03, NVA92, VAG89, VA90, VA94, VA91]. One major application of this class of operators is their link to Schrödinger operators in quantum theory. A Jacobi operator of this form is a discrete Schrödinger equation on the natural numbers with a potential which decays to zero at positive infinity [Sim79]. Another reason to study this class of operators is that the Jacobi operators for the classical Jacobi polynomials, useful across numerical analysis, are of this form [olva].

The spectral theory of Toeplitz operators such as $\Delta$ is well understood [BS13]. The spectral measure of $\Delta$ is the semi-circle $\mu_\Delta(s) = \frac{2}{\pi}(1-s^2)^{\frac{1}{2}}$ (restricted to $[-1,1]$), and its orthonormal polynomials are the Chebyshev polynomials of the second kind, $U_k(s)$. We prove the following more specific theorems about the spectra of this class of Jacobi operators $J$, and by an appropriate scaling and shifting by the identity, that of all Jacobi operators which are Toeplitz-plus-compact.

If $J$ is a *finite rank perturbation* of $\Delta$, i.e. there exists $n$ such that
$$\alpha_k = 0, \quad \beta_{k-1} = \frac{1}{2} \text{ for all } k \geq n, \tag{4.9}$$

- Theorem 4.3.8: The connection coefficient matrix $C_{J\to\Delta}$ can be decomposed into $C_{Toe} + C_{fin}$ where $C_{Toe}$ is Toeplitz, upper triangular and has bandwidth $2n-1$, and the entries of $C_{fin}$ are zero outside the $n-1 \times 2n-1$ principal submatrix.

- Theorem 4.3.21: let $c$ be the Toeplitz symbol of $C_{Toe}$. It is a degree $2n-1$ polynomial with $r \leq n$ roots inside the complex unit disc, all of which are simple. The spectrum of $J$ is

$$\sigma(J) = [-1,1] \cup \left\{ \lambda_k := \frac{1}{2}(z_k + z_k^{-1}) : c(z_k) = 0, |z_k| < 1 \right\}, \tag{4.10}$$

  and the spectral measure is given by the formula

$$\mu(s) = \frac{1}{p_C(s)}\mu_\Delta(s) + \sum_{k=1}^{r} \frac{(z_k - z_k^{-1})^2}{z_k c'(z_k) c(z_k^{-1})}\delta_{\lambda_k}(s), \tag{4.11}$$

  where $p_C(s) = \sum_{k=0}^{2n-1} c_{0,k}P_k(s) = \sum_{k=0}^{2n-1}\langle e_k, CC^T e_0\rangle U_k(s)$.

*Remark* 4.0.3. The author, along with Sheehan Olver (University of Sydney) have implemented the algorithms for computing the spectral measure and related functions of Toeplitz-plus-finite-rank Jacobi operators in an open source Julia package called *SpectralMeasures*. It makes use of the extensive open source Julia package called *ApproxFun* [Olvb, OT14], in particular the features for defining and manipulating functions and infinite dimensional operators. More information on the workings of the package is given in Section 4.6.

To extend the results to other Jacobi operators we must make a brief definition. For $R > 0$, define the geometrically weighted Banach space

$$\ell^1_R = \left\{ v \in \ell^\star_{\mathcal{F}} : \|v\|_{\ell^1_R} < \infty \right\}, \text{ where } \|v\|_{\ell^1_R} = \sum_{k=0}^{\infty} |v_k| R^k. \tag{4.12}$$

If $J$ is a *trace class perturbation* of $\Delta$, i.e.

$$\sum_{k=0}^{\infty} |\alpha_k| + \left|\beta_k - \frac{1}{2}\right| < \infty, \tag{4.13}$$

- Theorem 4.4.11: $C = C_{J\to\Delta}$ is bounded as an operator from $\ell^1_{R^{-1}}$ into itself, for all $R \in (0,1)$. Further, we have the decomposition $C = C_{Toe} + C_K$ where $C_{Toe}$ is upper triangular Toeplitz and $C_K$ is compact as an operator from $\ell^1_{R^{-1}}$ into itself.

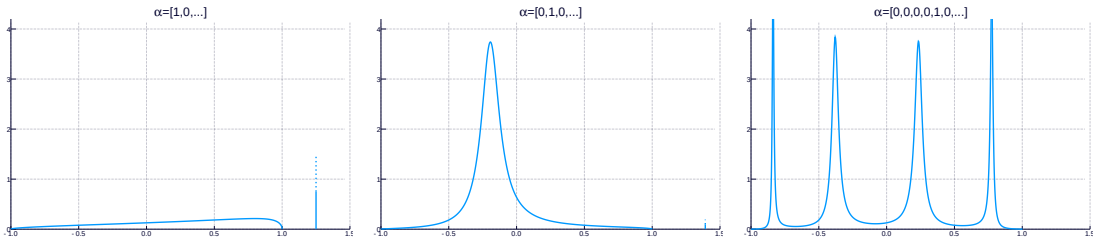



Fig. 4.1 These are the spectral measures of three different Jacobi operators; each differs from $\Delta$ in only one entry. The left plot is of the spectral measure of the Jacobi operator which is $\Delta$ except the $(0,0)$ entry is 1, the middle plot is that except the $(1,1)$ entry is 1, and the right plot is that except the $(4,4)$ entry is 1. This can be interpreted as a discrete Schrödinger operator with a single Dirac potential at different points along the real line. The continuous parts of the measures are given exactly by the computable formula in equation (4.11), and each has a single Dirac delta corresponding to discrete spectrum (the weight of the delta gets progressively smaller in each plot), the location of which can be computed with guaranteed error using interval arithmetic (see Section 4.6)

The transpose operators $C^T$ and $C_K^T$ are bounded and compact (respectively) as operators from $\ell^1_R$ into itself.

- Theorem 4.4.14 and Theorem 4.4.16 : Let $c$ be the Toeplitz symbol of $C_{Toe}$. It is analytic in the unit disc with real inside the complex unit disc. The discrete eigenvalues, as in the Toeplitz-plus-finite-rank case are of the form $\frac{1}{2}(z_k + z_k^{-1})$ where $z_k$ are the roots of $c$ in the open unit disc.

For computation of the spectrum and spectral measure of a Toeplitz-plus-trace-class Jacobi operator $J$, we suggest producing a Toeplitz-plus-finite-rank approximation $J^{[m]}$ and computing the spectrum and spectral measure of that. The question of choosing $J^{[m]}$ so as to guarantee this is a good approximation leads us to computability results. Following the pioneering work of Ben-Artzi–Hansen–Nevanlinna–Seidel on the Solvability Complexity Index [BAHNS15a, BAHNS15b, Han11], we prove the following theorems about computability. We assume real number arithmetic, and the results do not necessarily apply to algorithms using floating point arithmetic.

- Theorem 4.5.7 : If $J$ is a Toeplitz-plus-finite-rank Jacobi operator, then in a finite number of operations, the absolutely continuous part of the spectral measure is computable exactly, and the locations and weights of the discrete part of the spectral measure are computable to any desired accuracy. If the rank is known *a priori* then the algorithm can be designed to terminate with guaranteed error control.

- Theorem 4.5.9 : If $J = \Delta + K$ is a Toeplitz-plus-compact Jacobi operator, then in a finite number of operations, the spectrum of $J$ is computable to any desired accuracy in the Hausdorff metric on subsets of $\mathbb{R}$. If the quantity $\sup_{k\geq m} |\alpha_k| + \sup_{k\geq m} |\beta_k - \frac{1}{2}|$ can be estimated for all $m$, then the algorithm can be designed to terminate with guaranteed error control.

The Toeplitz symbol $c(z)$ defined in this chapter actually appears in the literature under other guises. Killip and Simon describe three different forms [KS03]: the Jost function $u_0$, the perturbation determinant $L$ and the Szegő function $D$ (see also [GC80, DS06a, DS06b]). The related function $p_C(s)$ in equation (4.11) can also be found in disguise as the $\varphi$ function in [VAG89, VA90, NVA92, VA94], and as the $S_n$ function in [DN86]. The novel aspect we present here is that results can be reinterpreted at the operator level, via (formal) similarity transformations of $J$ by $C$, a property that may facilitate extension beyond the Jacobi operators considered in this thesis. In particular, we see the following avenues of research that we intend to pursue in the future. There is also further discussion in the conclusions section.

The connection coefficient matrix can be defined for any two Jacobi operators $J$ and $D$. It is natural to explore what structure $C_{J\to D}$ has when $D$ is a different reference operator to $\Delta$, and $J$ is a finite rank, trace class, or compact perturbation of $D$. For example, the Jacobi operator with periodic entries, as is discussed in [DKS10, GVA86]. Beyond periodic Jacobi operators, it would be interesting from the viewpoint of ergodic theory if we could facilitate the study and computation of almost-periodic Jacobi operators, such as the discrete almost-Mathieu operator [Dei08]. Perturbations of the Jacobi operators for Laguerre polynomials and the Hermite polynomials could also be of interest, but challenges associated with the unboundedness of these operators could hamper progress [olva]. Discrete Schrödinger operators with non-decaying potentials will also be of interest in this direction.

The chapter is structured as follows. In Section 4.1 we cover established results about spectral theory of Jacobi operators. In Section 4.2 we discuss the basic properties of the connection coefficients matrix $C_{J\to D}$ for general Jacobi operators $J$ and $D$, and how they relate to spectra. In Section 4.3 we show how connection coefficient matrices apply to Toeplitz-plus-finite-rank Jacobi operators, and in Section 4.4 we extend these results to the Toeplitz-plus-trace-class case. Section 4.5 is devoted to issues of computability.

## 4.1 Spectral theory of Jacobi operators

In this section we present well known results about the spectra of Jacobi operators. This gives a self contained account of what is required to prove the results later in the chapter, and sets the notation.

### 4.1.1 Resolvents, measures and polynomials

To understand the spectral theory of Jacobi operators, and in particular to understand the spectral measure, we must relay the basic results on the interplay between the resolvent of the operator $J$, the measure whose Stieltjes transform is the resolvent, and certain sequences of polynomials that may be assigned to $J$.

**Definition 4.1.1.** Define the principal resolvent function for $\lambda \in \mathbb{C} \setminus \sigma(J)$,

$$G(\lambda) = \langle e_0, (J - \lambda)^{-1} e_0 \rangle. \tag{4.14}$$

**Theorem 4.1.2** ([Dei00, Tes00, Sim79])**.** *Let $J$ be a bounded Jacobi operator.*

*(i) There exists a unique compactly supported probability measure $\mu$ on $\mathbb{R}$, called the spectral measure of $J$, such that*

$$G(\lambda) = \int (s - \lambda)^{-1} \, \mathrm{d}\mu(s). \tag{4.15}$$

*(ii) For any $s_1 < s_2$ in $\mathbb{R}$,*

$$\frac{1}{2}\mu(\{s_1\}) + \mu((s_1, s_2)) + \frac{1}{2}\mu(\{s_2\}) = \lim_{\varepsilon \searrow 0} \frac{1}{\pi} \int_{s_1}^{s_2} \operatorname{Im} G(s + i\varepsilon) \, \mathrm{d}s. \tag{4.16}$$

*(iii) The spectrum of $J$ is*

$$\sigma(J) = \operatorname{supp}(\mu) = \overline{\{s \in \mathbb{R} : \liminf_{\varepsilon \searrow 0} \operatorname{Im} G(s + i\varepsilon) > 0\}}. \tag{4.17}$$

*The point spectrum $\sigma_p(J)$ of $J$ is the set of points $s \in \mathbb{R}$ such that the limit*

$$\mu(\{s\}) = \lim_{\varepsilon \searrow 0} \frac{\varepsilon}{i} G(s + i\varepsilon) \tag{4.18}$$

*exists and is positive.*

*The continuous spectrum of $J$ is the set of points $s \in \mathbb{R}$ such that $\mu(\{s\}) = 0$ but*

$$\liminf_{\varepsilon \searrow 0} \operatorname{Im} G(s + i\varepsilon) > 0. \tag{4.19}$$

*Remark* 4.1.3. Point (i) says that $G$ is the *Stieltjes transform* of $\mu$ and point (ii) gives the Perron-Stieltjes inversion formula [Sti94].

Let us demonstrate a concrete consequence of Theorem 4.1.2. If the principal resolvent is of the form

$$G(\lambda) = G_{\text{branch}}(\lambda) + \sum_{k=0}^{r} \frac{w_k}{\lambda_k - \lambda}, \tag{4.20}$$

where $G_{\text{branch}}$ is analytic everywhere except a set $B \subset \mathbb{R}$, upon which it has no poles[1]. Then the spectral measure is

$$\mathrm{d}\mu(s) = g(s)\mathrm{d}s + \sum_{k=0}^{r} w_k \delta_{\lambda_k}(s), \tag{4.21}$$

where $g$ is the integrable function

$$g(s) = \frac{1}{\pi} \lim_{\varepsilon \searrow 0} \operatorname{Im} G_{\text{branch}}(s + i\varepsilon), \tag{4.22}$$

which is zero for $s \notin B$.

The measure $\mu$ is the spectral measure that appears in the spectral theorem for self-adjoint operators on Hilbert space [DSBB71], as demonstrated by the following theorem [Dei00, Tes00, Sim79].

**Definition 4.1.4.** The orthonormal polynomials for $J$ are $P_0, P_1, P_2, \ldots$ defined by the three term recurrence

$$sP_k(s) = \beta_{k-1}P_{k-1}(s) + \alpha_k P_k(s) + \beta_k P_{k+1}(s), \tag{4.23}$$

$$P_{-1}(s) = 0, \quad P_0(s) = 1. \tag{4.24}$$

**Theorem 4.1.5** ([Dei00])**.** *Let $J$ be a bounded Jacobi operator and let $P_0, P_1, P_2, \ldots$ be as defined in Definition 4.1.4. Then we have the following.*

[1] By no poles, we mean that for all $\lambda_0 \in B$, $\limsup_{\lambda \to \lambda_0, \lambda \notin B} |(\lambda - \lambda_0)G(\lambda)| = 0$. This allows logarithmic singularities and algebraic singularities with order less than 1.

*(i)* *The polynomials are such that* $P_k(J)e_0 = e_k$.

*(ii)* *The polynomials are orthonormal with respect to the spectral measure of* $J$,

$$\int P_j(s)P_k(s)\,\mathrm{d}\mu(s) = \delta_{jk}. \tag{4.25}$$

*(iii)* *Define the unitary operator* $U : \ell^2 \to L^2_\mu(\mathbb{R})$ *such that* $Ue_k = P_k$. *Then for all* $f \in L^2_\mu(\mathbb{R})$,

$$UJU^*f(s) = sf(s). \tag{4.26}$$

*(iv)* *For all* $f \in L^1_\mu(\mathbb{R})$, *the operator* $f(J) : \ell_{\mathcal{F}} \to \ell^\star_{\mathcal{F}}$ *has entries given by,*

$$\langle e_i, f(J)e_j\rangle = \int f(s)P_i(s)P_j(s)\,\mathrm{d}\mu(s). \tag{4.27}$$

### 4.1.2 First associated polynomials

The following definition is standard in orthogonal polynomial theory (see for example, [Gau04, p. 18], [VA91]). We prove two lemmata about first associated polynomials that we will use later.

**Definition 4.1.6.** The first associated polynomials for $J$ are $P^\mu_0, P^\mu_1, P^\mu_2, \ldots$ defined by the three term recurrence

$$\lambda P^\mu_k(\lambda) = \beta_{k-1}P^\mu_{k-1}(\lambda) + \alpha_k P^\mu_k(\lambda) + \beta_k P^\mu_{k+1}(\lambda), \tag{4.28}$$

$$P^\mu_0(\lambda) = 0, \quad P^\mu_1(\lambda) = \beta_0^{-1}. \tag{4.29}$$

*Remark* 4.1.7. The orthogonal polynomials $P_0, P_1, \ldots$ for $J$ satisfy the following relationship componentwise for any polynomial $f$ and any $s$.

$$f(J)\begin{pmatrix} P_0(s) \\ P_1(s) \\ \vdots \end{pmatrix} = f(s)\begin{pmatrix} P_0(s) \\ P_1(s) \\ \vdots \end{pmatrix}. \tag{4.30}$$

The following lemma describes how the first associated polynomials satisfy the same relationship but with a remainder. The authors are not aware of the following elementary result appearing in the literature.

**Lemma 4.1.8.** *Let $f$ be a polynomial. For all $\lambda \in \mathbb{C} \setminus \sigma(J)$,*

$$f(J)\begin{pmatrix} P_0^\mu(\lambda) \\ P_1^\mu(\lambda) \\ \vdots \end{pmatrix} = f(\lambda)\begin{pmatrix} P_0^\mu(\lambda) \\ P_1^\mu(\lambda) \\ \vdots \end{pmatrix} + \frac{f(J)-f(\lambda)}{J-\lambda}e_0. \tag{4.31}$$

*Proof.* The proof is a straightforward induction on $k = 0, 1, \ldots$ for the functions $f(\lambda) = \lambda^k$, followed by an appeal to linearity. □

The relevance of the first associated polynomials for this work is the following integral formula.

**Lemma 4.1.9.** *([Gau04, pp. 17,18]) The first associated polynomials are given by the integral formula*

$$P_k^\mu(\lambda) = \int \frac{P_k(s)-P_k(\lambda)}{s-\lambda}\,\mathrm{d}\mu(s), \quad \lambda \in \mathbb{C} \setminus \sigma(J). \tag{4.32}$$

For notational convenience we also define the $\mu$-derivative of a general polynomial.

**Definition 4.1.10.** Let $\mu$ be a probability measure compactly supported on the real line and let $f$ be a polynomial. The $\mu$-derivative of $f$ is the polynomial defined by

$$f^\mu(\lambda) = \int \frac{f(s)-f(\lambda)}{s-\lambda}\,\mathrm{d}\mu(s). \tag{4.33}$$

## 4.2 Connection coefficient matrices

In this section we define the connection coefficient matrix and give preliminary results to indicate their relevance to spectral theory.

### 4.2.1 Basic properties

As at the start of the chapter, consider a second bounded Jacobi operator,

$$D = \begin{pmatrix} \gamma_0 & \delta_0 & & \\ \delta_0 & \gamma_1 & \delta_1 & \\ & \delta_1 & \gamma_2 & \ddots \\ & & \ddots & \ddots \end{pmatrix}, \tag{4.34}$$

with principal residual function $H(z)$, spectral measure $\nu$ and orthogonal polynomials denoted $Q_0, Q_1, Q_2, \ldots$. In Definition 4.0.2 we defined the connection coefficient matrix between $J$ and $D$, $C = C_{J\to D}$ to have entries satisfying

$$P_k(s) = c_{0k}Q_0(s) + c_{1k}Q_1(s) + \cdots + c_{kk}Q_k(s). \tag{4.35}$$

By orthonormality of the polynomial sequences the entries can also be interpreted as

$$C_{J\to D} = \begin{pmatrix} \langle P_0, Q_0\rangle_\nu & \langle P_1, Q_0\rangle_\nu & \langle P_2, Q_0\rangle_\nu & \cdots \\ 0 & \langle P_1, Q_1\rangle_\nu & \langle P_2, Q_1\rangle_\nu & \cdots \\ 0 & 0 & \langle P_2, Q_2\rangle_\nu & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{4.36}$$

where $\langle \cdot, \cdot \rangle_\nu$ is the standard inner product on $L^2_\nu(\mathbb{R})$.

**Lemma 4.2.1.** *The entries of the connection coefficients matrix $C_{J\to D}$ satisfy the following 5-point discrete system:*

$$\begin{pmatrix} & -\delta_{i-1}c_{i-1,j} & \\ & + & \\ \beta_{j-1}c_{i,j-1} & + (\alpha_j - \gamma_i)c_{ij} + & \beta_j c_{i,j+1} \\ & + & \\ & -\delta_i c_{i+1,j} & \end{pmatrix} = 0, \text{ for all } 0 \le i < j,$$

*with boundary conditions*

$$c_{ij} = \begin{cases} 1 & \text{if } i = j = 0, \\ 0 & \text{if } j = 0 \text{ and } i \neq 0, \\ 0 & \text{if } j = -1 \text{ or } i = -1. \end{cases} \tag{4.37}$$

*Proof.* Assume by convention that $c_{ij} = 0$ if $i = -1$ or $j = -1$. Now using this boundary condition and the three term recurrences for the polynomial sequences, we see that

$$\begin{aligned} \langle Q_i(s), sP_j(s)\rangle_\nu &= \beta_{j-1}\langle Q_i, P_{j-1}\rangle_\nu + \alpha_j\langle Q_i, P_j\rangle_\nu + \beta_j\langle Q_i, P_{j+1}\rangle_\nu & (4.38)\\ &= \beta_{j-1}c_{i,j-1} + \alpha_j c_{ij} + \beta_j c_{i,j+1}, & (4.39) \end{aligned}$$

and

$$\langle sQ_i(s), P_j(s)\rangle_\nu = \delta_{i-1}\langle Q_{i-1}, P_j\rangle_\nu + \gamma_i\langle Q_i, P_j\rangle_\nu + \delta_i\langle Q_{i+1}, P_j\rangle_\nu \tag{4.40}$$

$$= \delta_{i-1}c_{i-1,j} + \gamma_i c_{ij} + \delta_i c_{i+1,j}. \tag{4.41}$$

Since $\langle sQ_i(s), P_j(s)\rangle_\nu = \langle Q_i(s), sP_j(s)\rangle_\nu$, we have the result for the interior points $0 \le i < j$.

The remaining boundary conditions come from $c_{i0} = \langle Q_i, P_0\rangle_\nu$ which equals 1 if $i = 0$ and 0 otherwise. □

The 5-point discrete system described in Lemma 4.2.1 can be used to find an explicit linear recurrence to compute the entries of $C$,

$$c_{0,0} = 1 \tag{4.42}$$

$$c_{0,1} = (\gamma_0 - \alpha_0)/\beta_0 \tag{4.43}$$

$$c_{1,1} = \delta_0/\beta_0 \tag{4.44}$$

$$c_{0,j} = \left((\gamma_0 - \alpha_{j-1})c_{0,j-1} + \delta_0 c_{1,j-1} - \beta_{j-2}c_{0,j-2}\right)/\beta_{j-1} \tag{4.45}$$

$$c_{i,j} = \left(\delta_{i-1}c_{i-1,j-1} + (\gamma_i - \alpha_{j-1})c_{i,j-1} + \delta_i c_{i+1,j-1} - \beta_{j-2}c_{i,j-2}\right)/\beta_{j-1}. \tag{4.46}$$

The rows and columns of $C$ also satisfy infinite-vector-valued three-term recurrence relations. It is simply the 5-point discrete system rewritten in vector form, but this is the form which we make use of in the proofs later.

**Corollary 4.2.2.** *The columns of $C$ satisfy*

$$c_{*,0} = e_0 \tag{4.47}$$

$$Dc_{*,0} = \alpha_0 c_{*,0} + \beta_0 c_{*,1} \tag{4.48}$$

$$Dc_{*,j} = \beta_{j-1}c_{*,j-1} + \alpha_j c_{*,j} + \beta_j c_{*,j+1}. \tag{4.49}$$

*Consequently the $j$th column can be written $c_{*,j} = P_j(D)e_0$.*

*The rows of $C$ satisfy*

$$c_{0,*}J = \gamma_0 c_{0,*} + \delta_0 c_{1,*}, \tag{4.50}$$

$$c_{i,*}J = \delta_{i-1}c_{i-1,*} + \gamma_i c_{i,*} + \delta_i c_{i+1,*}. \tag{4.51}$$

*Consequently, the $i$th row can be written $c_{i,*} = c_{0,*}Q_i(J)$.*

*Proof.* The recurrence relations for the rows and columns are merely a renotation of equations (4.42)–(4.46). The consequences follow from the uniqueness of solution to second order difference equations with two initial data (adding $c_{-1,*} = 0$ and $c_{*,-1} = 0$). □

### 4.2.2 Connection coefficients and spectral theory

The following theorems give precise results about how the connection coefficients matrix $C$ can be useful for studying and computing the spectra of Jacobi operators.

**Theorem 4.2.3.** *Let $J$ and $D$ be bounded Jacobi operators and $C = C_{J\to D}$ the connection coefficients matrix. For all polynomials $p$, we have the following as operators from $\ell_{\mathcal{F}}$ to $\ell_{\mathcal{F}}$,*

$$Cp(J) = p(D)C.$$

*Remark* 4.2.4. In particular, as operators from $\ell_{\mathcal{F}}$ to $\ell_{\mathcal{F}}$,

$$J = C^{-1}DC.$$

*Proof.* First we begin with the case $p(z) = z$. By definition,

$$\begin{aligned} CJe_0 &= C(\alpha_0 e_0 + \beta_0 e_1) \\ &= \alpha_0 Ce_0 + \beta_0 Ce_1 \\ &= \alpha_0 c_{*,0} + \beta_0 c_{*,1}. \end{aligned}$$

Then by Corollary 4.2.2, this is equal to $Dc_{*,0}$, which is equal to $DCe_0$. Now, for any $j > 0$,

$$\begin{aligned} CJe_j &= C(\beta_{j-1}e_{j-1} + \alpha_j e_j + \beta_j e_{j+1}) \\ &= \beta_{j-1}c_{*,j-1} + \alpha_k c_{*,j} + \beta_j c_{*,j+1}. \end{aligned}$$

Then by Corollary 4.2.2, this is equal to $Dc_{*,j}$, which is equal to $DCe_j$. Hence $CJ = DC$.

Now, when $f(z) = z^k$ for any $k > 0$, $D^kC = D^{k-1}CJ = \cdots = CJ^k$. By linearity $Cf(J) = f(D)C$ for all polynomials $f$. □

**Proposition 4.2.5.** *Let $J$ and $D$ be bounded Jacobi operators with spectral measures $\mu$ and $\nu$ respectively, and connection coefficient matrix $C = C_{J\to D}$. Then*

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^2_\mu(\mathbb{R}) \text{ if and only if } c_{0,*} \in \ell^2,$$

*in which case*

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = \sum_{k=0}^{\infty} c_{0,k} P_k. \tag{4.52}$$

*Proof.* Suppose first that $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^2_\mu(\mathbb{R})$. Then $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = \sum_{k=0}^{\infty} a_k P_k$, for some $a \in \ell^2$. Since the polynomials $P_k$ are orthonormal in $L^2_\mu(\mathbb{R})$,

$$\begin{aligned} a_k &= \int P_k(s) \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(s)\, \mathrm{d}\mu(s) \\ &= \int P_k(s)\, \mathrm{d}\nu(s) \qquad \text{(definition of R–N derivative)} \\ &= c_{0,k} \qquad \text{(equation (4.36)).} \end{aligned}$$

Hence $c_{0,*} \in \ell^2$ and gives the $P_k$ coefficients of $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.

Conversely, suppose that $c_{0,*} \in \ell^2$. Then the function $\sum_{k=0}^{\infty} c_{0,k} P_k$ is in $L^2_\mu(\mathbb{R})$, and by the same manipulations as above its projections onto polynomial subspaces are equal to that of $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$. □

*Remark* 4.2.6. Theorem 4.2.5 can be made more general. If the first row of $C$ is such that the series defined by $\sum_{k=0}^{\infty} c_{0,k} P_k$ converges weakly (in the probabilists' sense) to an $L^1_\mu(\mathbb{R})$ function, then $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ exists and is equal to that limit. However, such a condition on the entries of $C$ is more difficult to check.

If we have a situation in which $c_{0,*} \in \ell^2$, we can by Proposition 4.2.5 and basic properties of the Radon–Nikodym derivative deduce that $\sigma(D) \subset \sigma(J)$ and the function defined by $\sum_{k=0}^{\infty} c_{0,k} P_k$ is zero on $\sigma(J) \setminus \sigma(D)$. This observation translates into a rootfinding problem in Section 4.3.

**Lemma 4.2.7.** *Let $J$ and $D$ be bounded Jacobi operators with spectral measures $\mu$ and $\nu$ respectively, and connection coefficient matrix $C = C_{J\to D}$. If $\nu$ is absolutely continuous with respect to $\mu$, then as operators mapping $\ell_{\mathcal{F}} \to \ell^\star_{\mathcal{F}}$,*

$$C^T C = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(J). \tag{4.53}$$

*Proof.* Note first that since $C : \ell_{\mathcal{F}} \to \ell_{\mathcal{F}}$ and $C^T : \ell^\star_{\mathcal{F}} \to \ell^\star_{\mathcal{F}}$, $C^TC$ is well-defined $\ell_{\mathcal{F}} \to \ell^\star_{\mathcal{F}}$. Then we have,

$$\begin{aligned}
\langle e_i, C^TCe_j\rangle &= \langle e_0, P_i(D)P_j(D)e_0\rangle && \text{(Corollary 4.2.2)}\\
&= \int P_i(s)P_j(s)\,\mathrm{d}\nu(s) && \text{(Theorem 4.1.5)}\\
&= \int P_i(s)P_j(s)\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(s)\,\mathrm{d}\mu(s)\\
&= \left\langle e_i, \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(J)e_j\right\rangle && \text{(Theorem 4.1.5).}
\end{aligned}$$

This completes the proof. □

**Proposition 4.2.8.** *Let $J$ and $D$ be bounded Jacobi operators with spectral measures $\mu$ and $\nu$ respectively, and connection coefficient matrix $C = C_{J\to D}$. Then $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^\infty_\mu(\mathbb{R})$ if and only if $C$ is a bounded operator on $\ell^2$, in which case*

$$\|C\|_2^2 = \operatorname*{ess\,sup}_{s\in\sigma(J)} \left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(s)\right|. \tag{4.54}$$

*Proof.* Suppose first that $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^\infty_\mu(\mathbb{R})$. Then by Lemma 4.2.7,

$$\|C^TC\|_2 = \left\|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(J)\right\|_2 = \operatorname*{ess\,sup}_{s\in\sigma(J)} \left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(s)\right|. \tag{4.55}$$

Hence $C^TC$ is bounded. The relationship $\|C\|_2^2 = \|C^TC\|_2$ completes this direction.

Now suppose that $C$ is bounded. Then by Proposition 4.2.5 $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^2_\mu(\mathbb{R})$. By Lemma 4.2.7, $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(J)$ is a bounded operator on $\ell^2$ (because it is equal to the bounded operator $C^TC$). Since $\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(J)\right\|_2 = \operatorname{ess\,sup}_{s\in\sigma(J)} \left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(s)\right|$, we have that in fact, $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^\infty_\mu(\mathbb{R})$. □

**Corollary 4.2.9.** *Let $J$ and $D$ be bounded Jacobi operators with spectral measures $\mu$ and $\nu$ respectively, and connection coefficient matrix $C = C_{J\to D}$. Then $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^\infty_\mu(\mathbb{R})$ and $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \in L^\infty_\nu(\mathbb{R})$ if and only if $C$ is bounded and invertible on $\ell^2$.*

*Proof.* By Proposition 4.2.8, $C_{J\to D}$ is bounded if and only if $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^\infty_\mu(\mathbb{R})$, and $C_{D\to J}$ is bounded if and only if $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} \in L^\infty_\nu(\mathbb{R})$. Combining this the fact that $C^{-1}_{J\to D} = C_{D\to J}$, as operators from $\ell_{\mathcal{F}}$ to itself, we complete the proof. □

**Lemma 4.2.10.** *Let $J$ and $D$ be Jacobi operators with principal resolvents $G$ and $H$ respectively. If $C_{J\to D}$ is banded with bandwidth $b$, then for all $\lambda \in \mathbb{C} \setminus \sigma(J)$,*

$$H(\lambda) = p_C(\lambda)G(\lambda) + p_C^{\mu}(\lambda), \tag{4.56}$$

*where $p_C(s) = \sum_{k=0}^{b} c_{0k}P_k(s)$ and $p_C^{\mu}$ is the $\mu$-derivative of $p_C$ as in Definition 4.1.10.*

*Remark* 4.2.11. Lemma 4.2.10 may be generalised to some cases in which the connection coefficient matrix is not banded, but one must be careful about which values of $\lambda \in \mathbb{C}$ for which $p_C^{\mu}(\lambda)$ and $p_C(\lambda)$ converge.

*Proof.* Using Theorem 4.1.2 and Proposition 4.2.5,

$$\begin{aligned} H(\lambda) &= \int (s-\lambda)^{-1} \mathrm{d}\nu(s) \\ &= \int (s-\lambda)^{-1} p_C(s) \mathrm{d}\mu(s). \end{aligned}$$

Now, provided both of the following integrals exist, we can split this into

$$H(\lambda) = \int (s-\lambda)^{-1} p_C(\lambda) \mathrm{d}\mu(s) + \int (s-\lambda)^{-1} (p_C(s) - p_C(\lambda)) \mathrm{d}\mu(s).$$

These integrals clearly do exist because $p_C$ is a polynomial, and give the desired result. □

The following definition and lemma are useful later.

**Definition 4.2.12.** Given polynomial sequences $P_0, P_1, P_2, \ldots$ and $Q_0, Q_1, Q_2, \ldots$ for Jacobi operators $J$ and $D$ respectively, we define the matrix $C^{\mu}$ to be the connection coefficients matrix between $P_0^{\mu}, P_1^{\mu}, P_2^{\mu}, \ldots$ and $Q_0, Q_1, Q_2, \ldots$ as in Definition 4.0.2, where $P_0^{\mu}, P_1^{\mu}, P_2^{\mu}, \ldots$ are the first associated polynomials for $J$ as in Definition 4.1.6. Noting that the lower triangular matrix $(C^{\mu})^T$ is a well defined operator fro $\ell^{\star}_{\mathcal{F}}$ into itself, we have

$$\begin{pmatrix} P_0^{\mu}(s) \\ P_1^{\mu}(s) \\ P_2^{\mu}(s) \\ \vdots \end{pmatrix} = (C^{\mu})^{\top} \begin{pmatrix} Q_0(s) \\ Q_1(s) \\ Q_2(s) \\ \vdots \end{pmatrix} \text{ for all } s. \tag{4.57}$$

*Remark* 4.2.13. Note that $C^{\mu}$ is *strictly* upper triangular, because the first associated polynomials have their degrees one less than their indices.

**Lemma 4.2.14.** *The operator $C^\mu$ as defined above for $C_{J\to D}$ is in fact $\beta_0^{-1}\,(0, C_{J^\mu\to D})$, where*

$$J^\mu = \begin{pmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \beta_2 & \\ & \beta_2 & \alpha_3 & \ddots \\ & & \ddots & \ddots \end{pmatrix}. \tag{4.58}$$

*Proof.* The (unique) orthonormal polynomials for $J^\mu$ are $\beta_0 P_1^\mu, \beta_0 P_2^\mu, \beta_0 P_3^\mu, \ldots$, and $P_0^\mu = 0$. □

## 4.3 Toeplitz-plus-finite-rank Jacobi operators

In this section we present several novel results which show how the connection coefficient matrices can be used for computing the spectral measure of a Toeplitz-plus-finite-rank Jacobi operator.

### 4.3.1 Jacobi operators for Chebyshev polynomials

There are two particular Jacobi operators with Toeplitz-plus-finite-rank structure that are of great interest,

$$\Delta = \begin{pmatrix} 0 & \frac{1}{2} & & \\ \frac{1}{2} & 0 & \frac{1}{2} & \\ & \frac{1}{2} & 0 & \ddots \\ & & \ddots & \ddots \end{pmatrix}, \text{ and } \Gamma = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & & & \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{2} & & \\ & \frac{1}{2} & 0 & \frac{1}{2} & \\ & & \frac{1}{2} & 0 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}. \tag{4.59}$$

The spectral measures of $\Delta$ and $\Gamma$ are

$$\mathrm{d}\mu_\Delta(s) = \frac{2}{\pi}\sqrt{1-s^2}\mathrm{d}s, \quad \mathrm{d}\mu_\Gamma(s) = \frac{1}{\pi}\frac{1}{\sqrt{1-s^2}}\mathrm{d}s, \tag{4.60}$$

supported on $[-1, 1]$.

Using results of Stieltjes in his seminal paper [Sti94], [AK65, App.], the principal resolvent can be written elegantly as a continued fraction,

$$G(\lambda) = \cfrac{-1}{\lambda - \alpha_0 - \cfrac{\beta_0^2}{\lambda - \alpha_1 - \cfrac{\beta_1^2}{\lambda - \alpha_2 - \ldots}}}. \tag{4.61}$$

Using this gives explicit expressions for the principal resolvents,

$$G_\Delta(\lambda) = 2\sqrt{\lambda+1}\sqrt{\lambda-1} - 2\lambda, \quad G_\Gamma(\lambda) = \frac{-1}{\sqrt{\lambda+1}\sqrt{\lambda-1}}. \tag{4.62}$$

*Remark* 4.3.1. We must be careful about which branch we refer to when we write the resolvents in this explicit form. Wherever $\sqrt{\ }$ is written above we mean the standard branch that is positive on $(0,\infty)$ with branch cut $(-\infty, 0]$. This gives a branch cut along $[-1,1]$ in both cases, the discontinuity of $G$ across which makes the Perron–Stieltjes inversion formula in Theorem 4.1.2 work. It also ensures the $\mathcal{O}(\lambda^{-1})$ decay resolvents enjoy as $\lambda \to \infty$.

The orthonormal polynomials for $\Delta$ are the Chebyshev polynomials of the second kind, which we denote $U_k(s)$,

$$U_k(s) = \frac{\sin((k+1)\cos^{-1}(s))}{\sin(\cos^{-1}(s))}. \tag{4.63}$$

The orthonormal polynomials for $\Gamma$ are the *normalised* Chebyshev polynomials of the first kind, which we denote $\tilde{T}_k(s)$. Note that these are not the usual Chebyshev polynomials of the first kind (denoted $T_k(s)$) [Gau04, TO16, Dei00]. We in fact have,

$$\tilde{T}_0(s) = 1, \quad \tilde{T}_k(s) = \sqrt{2}\cos(k\cos^{-1}(s)). \tag{4.64}$$

The first associated polynomials have simple relationships with the orthonormal polynomials,

$$U_k^{\mu_\Delta} = 2U_{k-1}, \quad \tilde{T}_k^{\mu_\Gamma} = \sqrt{2}U_{k-1}. \tag{4.65}$$

### 4.3.2 Rank-one perturbations

In this section we demonstrate for two simple, rank-one perturbations of $\Delta$ how the connection coefficient matrix relates properties of the spectrum of the operators. This will give some intuition as to what to expect in more general cases.

**Example 4.3.2** (Basic perturbation 1)**.** Let $\alpha \in \mathbb{R}$, and define

$$J_\alpha = \begin{pmatrix} \frac{\alpha}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ & \frac{1}{2} & 0 & \frac{1}{2} & \\ & & \frac{1}{2} & 0 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}. \tag{4.66}$$

Then the connection coefficient matrix $C_{J_\alpha \to \Delta}$ is the bidiagonal Toeplitz matrix

$$C_{J_\alpha \to \Delta} = \begin{pmatrix} 1 & -\alpha & & & \\ & 1 & -\alpha & & \\ & & 1 & -\alpha & \\ & & & \ddots & \ddots \end{pmatrix}. \tag{4.67}$$

This can be computed using the explicit recurrences (4.42)–(4.46). The connection coefficient matrix $C_{\Delta \to J_\alpha}$ (which is the inverse of $C_{J_\alpha \to \Delta}$ on $\ell_{\mathcal{F}}$) is the full Toeplitz matrix

$$C_{\Delta \to J_\alpha} = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ & 1 & \alpha & \alpha^2 & \cdots \\ & & 1 & \alpha & \cdots \\ & & & \ddots & \ddots \end{pmatrix}. \tag{4.68}$$

From this we see that $C = C_{J_\alpha \to \Delta}$ has a bounded inverse in $\ell^2$ if and only if $|\alpha| < 1$. Hence by Theorem 4.2.3, if $|\alpha| < 1$ then $CJ_\alpha C^{-1} = \Delta$ with each operator bounded on $\ell^2$, so that $\sigma(J_\alpha) = \sigma(\Delta) = [-1, 1]$. We will discuss what happens when $|\alpha| \geq 1$ later in the section.

**Example 4.3.3** (Basic perturbation 2)**.** Let $\beta > 0$, and define

$$J_\beta = \begin{pmatrix} 0 & \frac{\beta}{2} & & & \\ \frac{\beta}{2} & 0 & \frac{1}{2} & & \\ & \frac{1}{2} & 0 & \frac{1}{2} & \\ & & \frac{1}{2} & 0 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}. \tag{4.69}$$

Then the connection coefficient matrix $C_{J_\beta \to \Delta}$ is the banded Toeplitz-plus-rank-1 matrix

$$C_{J_\beta\to\Delta} = \begin{pmatrix} 1 & 0 & \beta^{-1}-\beta & & & \\ & \beta^{-1} & 0 & \beta^{-1}-\beta & & \\ & & \beta^{-1} & 0 & \beta^{-1}-\beta & \\ & & & \beta^{-1} & 0 & \ddots \\ & & & & \ddots & \ddots \end{pmatrix}. \tag{4.70}$$

Just as in Example 4.3.2, this can be computed using the explicit recurrences (4.42)–(4.46). The connection coefficient matrix $C_{\Delta\to J_\beta}$ (which is the inverse of $C_{J_\beta\to\Delta}$ on $\ell_{\mathcal{F}}$) is the Toeplitz-plus-rank-1 matrix

$$C_{\Delta\to J_\beta} = \begin{pmatrix} 1 & 0 & \beta^2-1 & 0 & (\beta^2-1)^2 & 0 & (\beta^2-1)^3 & \cdots \\ & \beta & 0 & \beta(\beta^2-1) & 0 & \beta(\beta^2-1)^2 & 0 & \cdots \\ & & \beta & 0 & \beta(\beta^2-1) & 0 & \beta(\beta^2-1)^2 & \cdots \\ & & & \beta & 0 & \beta(\beta^2-1) & 0 & \cdots \\ & & & & \ddots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{4.71}$$

From this we see that $C = C_{J_\beta\to\Delta}$ has a bounded inverse on $\ell^2$ if and only if $\beta < \sqrt{2}$. Hence by Theorem 4.2.3, if $\beta < \sqrt{2}$ then $CJ_\beta C^{-1} = \Delta$ with each operator bounded on $\ell^2$, so that $\sigma(J_\beta) = \sigma(\Delta) = [-1, 1]$. We will discuss what happens when $\beta \geq \sqrt{2}$ later in the section. Note that the case $\beta = \sqrt{2}$ gives the Jacobi operator $\Gamma$ in equation (4.59).

### 4.3.3 Fine properties of the connection coefficients

The two basic perturbations of $\Delta$ discussed above give connection coefficient matrices that are highly structured. The following lemmata and theorems prove that this is no coincidence; in fact, if Jacobi operator $J$ is a finite-rank perturbation of $\Delta$ then $C_{J\to\Delta}$ is also a finite-rank perturbation of Toeplitz.

*Remark* 4.3.4. Note for the following results that all vectors and matrices are indexed starting from 0.

**Lemma 4.3.5.** *If $\delta_j = \beta_j$ for $j \geq n$ then $c_{jj} = c_{nn}$ for all $j \geq n$.*

*Proof.* By the recurrence in Lemma 4.2.1, $c_{jj} = (\delta_{j-1}/\beta_{j-1})c_{j-1,j-1}$. The result follows by induction. □

**Lemma 4.3.6.** *Let $J$ and $D$ be Jacobi operators with coefficients $\{\alpha_k, \beta_k\}$ and $\{\gamma_k, \delta_k\}$ respectively, such that there exists an $n$ such that* [2]

$$\alpha_k = \gamma_k = \alpha_n, \quad \beta_{k-1} = \delta_{k-1} = \beta_{n-1} \text{ for all } k \geq n. \tag{4.72}$$

*Then the entries of the connection coefficient matrix $C = C_{J\to D}$ satisfy*

$$c_{i,j} = c_{i-1,j-1} \text{ for all } i, j > 0 \text{ such that } i \geq n. \tag{4.73}$$

*Remark* 4.3.7. This means that $C$ is of the form $C = C_{\text{Toe}} + C_{\text{fin}}$ where $C_{\text{Toe}}$ is Toeplitz and $C_{\text{fin}}$ is zero except in the first $n-1$ rows. For example, when $n = 4$, we have the following structure

$$C = \begin{pmatrix} t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & \cdots \\ & t_0 & t_1 & t_2 & t_3 & t_4 & \ddots \\ & & t_0 & t_1 & t_2 & t_3 & \ddots \\ & & & t_0 & t_1 & t_2 & \ddots \\ & & & & \ddots & \ddots & \ddots \end{pmatrix} + \begin{pmatrix} f_{00} & f_{01} & f_{02} & f_{03} & f_{04} & \cdots \\ & f_{11} & f_{12} & f_{13} & f_{14} & \cdots \\ & & f_{22} & f_{23} & f_{24} & \cdots \\ & & & & & \\ & & & & & \end{pmatrix}.$$

*Proof.* We prove by induction on $k = 0, 1, 2, \ldots$ that

$$c_{i,i+k} = c_{i-1,i+k-1} \text{ for all } i \geq n. \tag{4.74}$$

We use the recurrences in Lemma 4.2.1 and equations (4.42)–(4.46). The base case $k = 0$ is proved in Lemma 4.3.5. Now we deal with the second base case, $k = 1$. For any $i \geq n$, we have $\beta_i = \delta_i = \beta_{i-1} = \delta_{i-1}$, and $\alpha_i = \gamma_i$, so

$$\begin{aligned} c_{i,i+1} &= \left(\delta_{i-1}c_{i-1,i} + (\gamma_i - \alpha_i)c_{i,i} + \delta_i c_{i+1,i} - \beta_{i-1}c_{i,i-1}\right)/\beta_i \\ &= 1 \cdot c_{i-1,i} + 0 \cdot c_{i,i} + 1 \cdot 0 - 1 \cdot 0 \\ &= c_{i-1,i}. \end{aligned}$$

Now we deal with the case $k > 1$. For any $i \geq n$, we have $\delta_i = \delta_{i-1} = \beta_{i+k-2} = \beta_{i+k-1}$, and $\alpha_{i+k-1} = \gamma_i$, so

$$c_{i,i+k} = \left(\delta_{i-1}c_{i-1,i+k-1} + (\gamma_i - \alpha_{i+k-1})c_{i,i+k-1} + \delta_i c_{i+1,i+k-1} - \beta_{i+k-2}c_{i,i+k-2}\right)/\beta_{i+k-1}$$

[2] More intuitively, the entries of $J$ and $D$ are both equal and Toeplitz, except in the principal $n \times n$ submatrix, where neither statement necessarily holds.

$$
\begin{aligned}
&= 1 \cdot c_{i-1,i+k-1} + 0 \cdot c_{i,i+k-1} + 1 \cdot c_{i+1,i+k-1} - 1 \cdot c_{i,i+k-2} \\
&= c_{i-1,i+k-1} + c_{i+1,i+k-1} - c_{i,i+k-2} \\
&= c_{i-1,i+k-1}.
\end{aligned}
$$

The last line follows from the induction hypothesis for the case $k-2$ (hence why we needed two base cases). □

The special case in which $D$ is Toeplitz gives even more structure to $C$, as demonstrated by the following theorem. We state the results for a finite-rank perturbation of the free Jacobi operator $\Delta$, but they apply to general Toeplitz-plus-finite rank Jacobi operators because the connection coefficients matrix $C$ is unaffected by a scaling and shift by the identity applied to both $J$ and $D$.

**Theorem 4.3.8.** *Let $J$ be a Jacobi operator such that there exists an $n$ such that*

$$\alpha_k = 0, \quad \beta_{k-1} = \frac{1}{2} \text{ for all } k \geq n, \tag{4.75}$$

*i.e. it is equal to the free Jacobi operator $\Delta$ outside the $n \times n$ principal submatrix. Then the entries of the connection coefficient matrix $C = C_{J \to \Delta}$ satisfy*

$$c_{i,j} = c_{i-1,j-1} \text{ for all } i,j > 0 \text{ such that } i + j \geq 2n \tag{4.76}$$

$$c_{0,j} = 0 \text{ for all } j \geq 2n. \tag{4.77}$$

*Remark* 4.3.9. This means that $C$ is of the form $C = C_{\text{Toe}} + C_{\text{fin}}$ where $C_{\text{Toe}}$ is Toeplitz with bandwidth $2n-1$ and $C_{\text{fin}}$ zero except for entries in the $(n-1) \times (2n-2)$ principal submatrix. For example when $n = 4$, we have the following structure,

$$C = \begin{pmatrix}
t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & & \\
 & t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & \\
 & & t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & \ddots \\
 & & & t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & \ddots \\
 & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots
\end{pmatrix} + \begin{pmatrix}
f_{0,0} & f_{0,1} & f_{0,2} & f_{0,3} & f_{0,4} & f_{0,5} \\
 & f_{1,1} & f_{1,2} & f_{1,3} & f_{1,4} & \\
 & & f_{2,2} & f_{2,3} & & \\
 & & & & & \\
 & & & & &
\end{pmatrix}.$$

*Proof.* First we prove (4.76). Fix $i, j$ such that $i + j \geq 2n$. Note that the case $i \geq n$ is proven in Lemma 4.3.6. Hence we assume $i < n$. Hence $j > n$. Using Lemma 4.2.1 and equations (4.42)–(4.46) we find the following recurrence. Substituting $\delta_i = \frac{1}{2}$, $\gamma_i = 0$

for all $i$, and $\alpha_k = 0$, $\beta_{k-1} = \frac{1}{2}$ for $k \geq n$ into the recurrence, we have

$$\begin{aligned} c_{i,j} &= \left(\delta_{i-1}c_{i-1,j-1} + (\gamma_i - \alpha_{j-1})c_{i,j-1} + \delta_i c_{i+1,j-1} - \beta_{j-2}c_{i,j-2}\right)/\beta_{j-1}. \\ &= \left(\frac{1}{2}c_{i-1,j-1} - \alpha_{j-1}c_{i,j-1} + \frac{1}{2}c_{i+1,j-1} - \beta_{j-2}c_{i,j-2}\right)/\beta_{j-1} \\ &= c_{i-1,j-1} + c_{i+1,j-1} - c_{i,j-2}. \end{aligned}$$

Repeating this process on $c_{i+1,j-1}$ in the above expression gives

$$c_{i,j} = c_{i-1,j-1} + c_{i+2,j-2} - c_{i+1,j-3}.$$

Repeating the process on $c_{i+2,j-2}$ and so on eventually gives

$$c_{i,j} = c_{i-1,j-1} + c_{n,i+j-n} - c_{n-1,i+j-n-1}.$$

By Lemma 4.3.6, $c_{n,i+j-n} = c_{n-1,i+j-n-1}$, so we are left with $c_{i,j} = c_{i-1,j-1}$. This completes the proof of (4.76).

Now we prove (4.77). Let $j \geq 2n$. Then

$$\begin{aligned} c_{0,j} &= \left((\gamma_0 - \alpha_{j-1})c_{0,j-1} + \delta_0 c_{1,j-1} - \beta_{j-2}c_{0,j-2}\right)/\beta_{j-1} \\ &= \left(-\alpha_{j-1}c_{0,j-1} + \frac{1}{2}c_{1,j-1} - \beta_{j-2}c_{0,j-2}\right)/\beta_{j-1} \\ &= c_{1,j-1} - c_{0,j-2}. \end{aligned}$$

This is equal to zero by (4.76), because $1 + (j-1) \geq 2n$. □

**Corollary 4.3.10.** *Let $C^\mu$ be as defined in Definition 4.2.12 for $C$ as in Theorem 4.3.8. Then $C^\mu = C^\mu_{\text{Toe}} + C^\mu_F$, where $C^\mu_{\text{Toe}}$ is Toeplitz with bandwidth $2n-2$ and $C^\mu_F$ is zero outside the $(n-2) \times (2n-1)$ principal submatrix.*

*Proof.* This follows from Theorem 4.3.8 applied to $J^\mu$ as defined in Lemma 4.2.14. □

*Remark* 4.3.11. A technical point worth noting for use in proofs later is that for Toeplitz-plus-finite-rank Jacobi operators like $J$ and $D$ occurring in Theorem 4.3.8 and Corollary 4.3.10, the operators $C$, $C^T$, $C^\mu$ and $(C^\mu)^T$ all map $\ell_\mathcal{F}$ to $\ell_\mathcal{F}$. Consequently, combinations such as $CC^T$, $C^\mu C^T$ are all well defined operators from $\ell_\mathcal{F}$ to $\ell_\mathcal{F}$.

### 4.3.4 Properties of the resolvent

When the Jacobi operator $J$ is Toeplitz-plus-finite rank, as a consequence of the structure of the connection coefficients matrix proved in subsection 4.3.3, the principal resolvent $G$ (see Definition 4.1.1) and spectral measure (see Theorem 4.1.2) are also highly structured. As usual these proofs are stated for a finite-rank perturbation of the free Jacobi operator $\Delta$, but apply to general Toeplitz-plus-finite rank Jacobi operators by applying appropriate scaling and shifting.

**Theorem 4.3.12.** *Let $J$ be a Jacobi operator such that there exists an $n$ such that*

$$\alpha_k = 0, \quad \beta_{k-1} = \frac{1}{2} \text{ for all } k \geq n, \tag{4.78}$$

*i.e. it is equal to the free Jacobi operator $\Delta$ outside the $n \times n$ principal submatrix. Then the principal resolvent for $J$ is*

$$G(\lambda) = \frac{G_\Delta(\lambda) - p_C^\mu(\lambda)}{p_C(\lambda)}, \tag{4.79}$$

*where*

$$p_C(\lambda) = \sum_{k=0}^{2n-1} c_{0,k} P_k(\lambda) = \sum_{k=0}^{2n-1} \langle e_k, CC^T e_0 \rangle U_k(\lambda), \tag{4.80}$$

$$p_C^\mu(\lambda) = \sum_{k=1}^{2n-1} c_{0,k} P_k^\mu(\lambda) = \sum_{k=0}^{2n-1} \langle e_k, C^\mu C^T e_0 \rangle U_k(\lambda), \tag{4.81}$$

*$P_k$ are the orthonormal polynomials for $J$, $P_k^\mu$ are the first associated polynomials for $J$ as in Definition 4.1.6, and $U_k$ are the Chebyshev polynomials of the second kind.*

*Remark* 4.3.13. $p_C^\mu$ is the $\mu$-derivative of $p_C$ as in Definition 4.1.10.

*Proof.* By Lemma 4.2.10 with bandwidth $b = 2n - 1$,

$$G_\Delta(\lambda) = p_C(\lambda) G(\lambda) + p_C^\mu(\lambda). \tag{4.82}$$

The equation can be immediately rearranged to obtain (4.79). To see the equality in equations (4.80), note that by the definition of the connection coefficient matrix $C$,

$$\sum_{k=0}^{2n-1} c_{0,k} P_k(\lambda) = \sum_{k=0}^{2n-1} c_{0,k} \sum_{j=0}^{2n-1} c_{j,k} U_j(\lambda) \tag{4.83}$$

$$= \sum_{j=0}^{2n-1} \left( \sum_{k=0}^{2n-1} c_{0,k} c_{j,k} \right) U_j(\lambda) \tag{4.84}$$

$$= \sum_{j=0}^{2n-1} \langle e_j, CC^T e_0 \rangle U_j(\lambda). \tag{4.85}$$

Equation (4.81) follows by the same algebra. □

**Theorem 4.3.14.** *Let $J$ be a Jacobi operator such that there exists an $n$ such that*

$$\alpha_k = 0, \quad \beta_{k-1} = \frac{1}{2} \text{ for all } k \geq n, \tag{4.86}$$

*i.e. it is equal to the free Jacobi operator $\Delta$ outside the $n \times n$ principal submatrix. Then the spectral measure for $J$ is*

$$\mu(s) = \frac{1}{p_C(s)} \mu_\Delta(s) + \sum_{k=1}^{r} w_k \delta_{\lambda_k}(s), \tag{4.87}$$

*where $\lambda_1, \ldots, \lambda_r$ are the roots of $p_C$ in $\mathbb{R} \setminus \{1, -1\}$ such that*

$$w_k = \lim_{\varepsilon \searrow 0} \frac{\varepsilon}{i} G(\lambda_k + i\varepsilon) \neq 0. \tag{4.88}$$

*Furthermore, there are no roots of $p_C$ inside $(-1, 1)$, and the number of roots of $p_C$ for which $w_k \neq 0$ is at most $n$ (i.e. $r \leq n$).*

*Proof.* Let $G$ and $\mu$ be the principal resolvent and spectral measure of $J$ respectively. By Theorem 4.3.12,

$$G(\lambda) = \frac{G_\Delta(\lambda) - p_C^\mu(\lambda)}{p_C(\lambda)}.$$

Letting $\lambda_1, \ldots, \lambda_{2n-1}$ be the roots of $p_C$ in the complex plane, define the set

$$S = [-1, 1] \cup (\{\lambda_1, \ldots, \lambda_{2n-1}\} \cap \mathbb{R}). \tag{4.89}$$

By inspection of the above formula for $G$, and because resolvents of selfadjoint operators are analytic off the real line, we have that $G$ is continuous outside of $S$. Therefore, for any $s \in \mathbb{R}$ such that $\operatorname{dist}(s, S) > 0$, we have

$$\lim_{\varepsilon \searrow 0} \operatorname{Im} G(s + i\varepsilon) = \operatorname{Im} G(s) = 0. \tag{4.90}$$

Hence by Theorem 4.1.2 part (ii), for any interval $(s_1, s_2)$ such that $\operatorname{dist}(S, (s_1, s_2)) > 0$, we have $\mu((s_1, s_2)) + \frac{1}{2}\mu(\{s_1\}) + \frac{1}{2}\mu(\{s_2\}) = 0$. Therefore the essential support of $\mu$ is contained within $S$.

We are interested in the real roots of $p_C$. Can any roots of $p_C$ lie in the interval $[-1, 1]$? By Proposition 4.2.5, $\mathrm{d}\mu_\Delta(s) = p_C(s)\mathrm{d}\mu(s)$ for all $s \in \mathbb{R}$. For any $s \in [-1, 1]$ such that $p_C(s) \neq 0$, it follows that $\mathrm{d}\mu(s) = \frac{2}{\pi}\frac{\sqrt{1-s^2}}{p_C(s)}\mathrm{d}s$. From this we have

$$1 \geq \mu((-1, 1)) = \int_{-1}^{1} \frac{2}{\pi}\frac{\sqrt{1-s^2}}{p_C(s)}\,\mathrm{d}s. \tag{4.91}$$

This integral is only finite if $p_C$ has no roots in $(-1, 1)$ and only simple roots at $\pm 1$. Since $S$ is a disjoint union of $[-1, 1]$ and a finite set $S'$ we can write

$$\mu(s)\frac{1}{p_C(s)}\mu_\Delta(s) + \sum_{\lambda_k \in S'} \mu(\{\lambda_k\})\delta_{\lambda_k}(s). \tag{4.92}$$

By Theorem 4.1.2 part (iii),

$$\mu(\{s\}) = \lim_{\varepsilon \searrow 0} \frac{\varepsilon}{i} G(s + i\varepsilon) \text{ for all } s \in \mathbb{R}. \tag{4.93}$$

This gives the desired formula for $w_k$.

Finally, let us prove that the number of eigenvalues (denoted $r$ in the statement of the theorem) must be at most $n$. We proceed by induction on $n = 0, 1, 2, \ldots$ that the Jacobi operator $J = \Delta + \sum_{k=1}^{n} v_k \otimes v_k$ has at most $n$ eigenvalues, where $(v \otimes w)u = \langle w, u\rangle v$ for $u, v, w \in \ell^2$.

For $n = 0$ this is already known because $\Delta$ only has continuous spectrum. Now consider $n > 0$. If $\lambda$ is an eigenvalue of $J$ with eigenvector $w \in \ell^2$, then

$$(\Delta - \lambda + \sum_{k=1}^{n} v_k \otimes v_k)w = 0. \tag{4.94}$$

If $\langle v_k, w\rangle = 0$ for all $k$ then $w$ is an eigenvector of $\Delta$, which is impossible because $\Delta$ has only continuous spectrum. Hence, without loss of generality $\langle v_n, w\rangle \neq 0$. Write $\tilde{J} = \Delta + \sum_{k=1}^{n-1} v_k \otimes v_k$. Then we have

$$(\tilde{J} - \lambda)w + \langle v_n, w\rangle v_n = 0. \tag{4.95}$$

This implies that $\lambda$ is not an eigenvalue of $\tilde{J}$. Hence we may write

$$w + (\tilde{J} - \lambda)^{-1}\langle v_n, w\rangle v_n = 0, \tag{4.96}$$

which after multiplying by $v_n^T/\langle v_n, w\rangle$ gives

$$1 + \langle v_n, (\tilde{J} - \lambda)^{-1} v_n\rangle = 0. \tag{4.97}$$

If we write the left hand side as $f(\lambda)$, then for $\lambda \in \mathbb{R}\backslash\sigma(\tilde{J})$, $f'(\lambda) = \langle v_n, (\tilde{J}-\lambda)^{-2}v_n\rangle > 0$ since $(\tilde{J} - \lambda)$ is invertible and $v_n \neq 0$. Hence on intervals of continuity $f$ is strictly increasing and therefore injective. Therefore there can be at most one root of $f$ in each interval of continuity. See Figure 4.2 for a demonstration of the intervals of continuity for an example $f(\lambda)$.

By inductive hypothesis $\langle v_n, (\tilde{J} - \lambda)^{-1} v_n\rangle$ has at most $n-1$ poles and a branch cut along $[-1, 1]$. $f$ thus has at most $n+1$ intervals of continuity. However, for the left-most interval, $(-\infty, a)$, where $a$ is either the left-most pole of $f$ or $-1$, $f(\lambda) \geq 1$, since $\lim_{\lambda\to-\infty} f(\lambda) = 1$ and $f$ is increasing. Hence there are at most $n$ roots of the equation $f(\lambda) = 0$. Hence there are at most $n$ eigenvalues of $J$. □

*Remark* 4.3.15. Theorem 4.3.14 gives an explicit formula for the spectral measure of $J$, when $J$ is Toeplitz-plus-finite-rank Jacobi operator. The entries of $C$ can be computed in $\mathcal{O}(n^2)$ operations (for an $n \times n$ perturbation of Toeplitz). Hence, the absolutely continuous part of the measure can be computed *exactly* in finite time. It would appear at first that we may compute the locations of the point spectrum by computing the roots of $p_C$, but in fact we find that not all real roots of $p_C$ have $w_k \neq 0$. Hence we rely on cancellation between the numerator and denominator in the formula for $G(\lambda)$, which numerically is a dangerous game. Subsection 4.3.5 remedies this situation.

**Example 4.3.16** (Basic perturbation 1 revisited)**.** The polynomial $p_C$ in Theorem 4.3.12 is

$$p_C(\lambda) = c_{0,0}P_0(\lambda) + c_{0,1}P_1(\lambda) = 1 - \alpha(2\lambda - \alpha) = 2\alpha\left(\frac{1}{2}(\alpha + \alpha^{-1}) - \lambda\right), \tag{4.98}$$

and the $\mu$-derivative is $p_C^{\mu}(\lambda) = -2\alpha$. Theorem 4.3.12 gives

$$G(\lambda) = \frac{G_\Delta(\lambda) + 2\alpha}{2\alpha\left(\frac{1}{2}(\alpha + \alpha^{-1}) - \lambda\right)}. \tag{4.99}$$

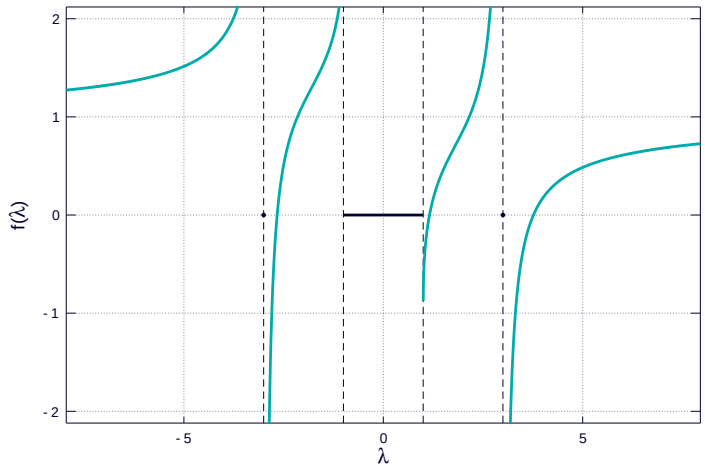



Fig. 4.2 In order to demonstrate the second part of the proof of Theorem 4.3.14, this figure shows the Herglotz function $f(\lambda) = 1 + 2\sqrt{\lambda-1}\sqrt{\lambda+1} - 2\lambda + \lambda(9-\lambda^2)^{-1}$. There are four intervals of continuity for $f$; they are $(-\infty,-3)$, $(-3,-1)$, $(1,3)$ and $(3,\infty)$, within each of which $f$ is strictly increasing. As in the proof, the left-most interval cannot possibly contain a root because there $f(\lambda) > 1$. Hence $f$ has at most three real roots.

Consider the case $|\alpha| \leq 1$. Then a brief calculation reveals $G_\Delta(\frac{1}{2}(\alpha+\alpha^{-1})) = -2\alpha$. Hence the root $\lambda = \frac{1}{2}(\alpha+\alpha^{-1})$ of the denominator is always cancelled out. Hence $G$ has no poles, and so $J$ has no eigenvalues.

In the case where $|\alpha| > 1$, we have a different situation. Here $G_\Delta(\frac{1}{2}(\alpha+\alpha^{-1})) = -2\alpha^{-1}$. Therefore the root $\lambda = \frac{1}{2}(\alpha+\alpha^{-1})$ of the denominator is *never* cancelled out. Hence there is always a pole of $G$ at $\lambda = \frac{1}{2}(\alpha+\alpha^{-1})$, and therefore also an eigenvalue of $J$ there.

Notice a heavy reliance on cancellations in the numerator and denominator for the existence of eigenvalues. The approach in subsection 4.3.5 avoids this.

**Example 4.3.17** (Basic perturbation 2 revisited)**.** The polynomial $p_C$ in Theorem 4.3.12 is

$$p_C(\lambda) = c_{0,0}P_0(\lambda) + c_{0,2}P_2(\lambda) = 1 + (\beta^{-1}-\beta)(4\beta^{-1}\lambda^2 - \beta). \tag{4.100}$$

This simplifies to $p_C(\lambda) = 4(1-\beta^{-2})\left(\frac{\beta^4}{4(\beta^2-1)} - \lambda^2\right)$. Using Definition 4.1.6, the $\mu$-derivative is $p_C^\mu(\lambda) = c_{0,2}P_2^\mu(\lambda) = 4\beta^{-1}\lambda$. Theorem 4.3.12 gives

$$G(\lambda) = \frac{G_\Delta(\lambda) + 4\beta^{-1}\lambda}{4(1-\beta^{-2})\left(\frac{\beta^4}{4(\beta^2-1)} - \lambda^2\right)}. \tag{4.101}$$

Clearly the only points $G$ may have a pole is at $\lambda = \pm\frac{\beta^2}{2\sqrt{\beta^2-1}}$. However, it is difficult to see whether there would be cancellation on the numerator. In the previous discussion on this example we noted that there would not be any poles when $|\beta| < \sqrt{2}$, which means that the numerator must be zero at these points, but it is far from clear here. The techniques we develop in the sequel illuminate this issue, especially for examples which are much more complicated than the two trivial ones given so far.

### 4.3.5 The Joukowski transformation

The following two lemmata and two theorems prove that Theorem 4.3.12 and Theorem 4.3.14 can be simplified drastically by making the change of variables

$$\lambda(z) = \frac{1}{2}(z + z^{-1}) \tag{4.102}$$

This map is known as the Joukowski map. It is an analytic bijection from $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$ to $\mathbb{C} \setminus [-1, 1]$, sending the unit circle to two copies of the interval $[-1, 1]$.

The Joukowski map has special relevance for the principal resolvent of $\Delta$. A brief calculation reveals that for $z \in \mathbb{D}$,

$$G_\Delta(\lambda(z)) = -2z. \tag{4.103}$$

Further, we will see that the polynomials $p_C(\lambda)$ and $p_C^\mu(\lambda)$ occurring in our formula for $G$ can be expressed neatly as polynomials in $z$ and $z^{-1}$. This is a consequence of a special property of the Chebyshev polynomials of the second kind, that for any $k \in \mathbb{Z}$ and $z \in \mathbb{D}$

$$\frac{U_{m-k}(\lambda(z))}{U_m(\lambda(z))} \to z^k \text{ as } m \to \infty. \tag{4.104}$$

These convenient facts allow us to remove any square roots involved in the formulae in Theorem 4.3.12.

**Lemma 4.3.18.** *Let $p_C(\lambda) = \sum_{k=0}^{2n-1} \langle e_0, CC^T e_k\rangle U_k(\lambda)$ as in Theorem 4.3.12 and let $c$ be the symbol of $C_{\text{Toe}}$, the Toeplitz part of $C$ as guaranteed by Theorem 4.3.8. Then*

$$p_C(\lambda(z)) = c(z)c(z^{-1}), \tag{4.105}$$

*where $\lambda(z) = \frac{1}{2}(z + z^{-1})$.*

*Proof.* The key quantity to observe for this proof is

$$\frac{1}{U_m(\lambda)}\left\langle \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, CC^T e_m \right\rangle \tag{4.106}$$

as $m \to \infty$. We will show it is equal to both sides of equation (4.105). Recall from Remark 4.3.11 that $CC^T$ maps from $\ell_{\mathcal{F}}$ to $\ell_{\mathcal{F}}$.

By Corollary 4.2.2 on the $m$th row of $C$, we have $C^T e_m = U_m(J)C^T e_0$. By Theorem 4.2.3, $CU_m(J) = U_m(\Delta)C$. Hence,

$$CC^T e_m = CU_m(J)C^T e_0 = U_m(\Delta)CC^T e_0. \tag{4.107}$$

Using this and the relationship (4.30) for $f = U_m$ and $J = \Delta$, we have

$$\begin{aligned}\left\langle \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, CC^T e_m \right\rangle &= \left\langle \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, U_m(\Delta)CC^T e_0 \right\rangle = \left\langle U_m(\Delta) \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, CC^T e_0 \right\rangle \\ &= U_m(\lambda)\left\langle \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, CC^T e_0 \right\rangle = U_m(\lambda)p_C(\lambda).\end{aligned}$$

Now, by Theorem 4.3.8, $C = C_{\text{Toe}} + C_{\text{fin}}$, where $C_{\text{fin}}$ is zero outside the $(n-1) \times (2n-2)$ principal submatrix. Hence for $m$ sufficiently large we have $c_{m,m+k} = t_k$ for a sequence $(t_k)_{k\in\mathbb{Z}}$ such that $t_k = 0$ for $k \notin \{0, 1, \ldots, 2n-1\}$. The Toeplitz symbol of $C_{\text{Toe}}$ is $c(z) = \sum_{k=0}^{2n-1} t_k z^k$. Hence we have for $m$ sufficiently large,

$$\begin{aligned}\frac{1}{U_m(\lambda)}\left\langle \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, CC^T e_m \right\rangle &= \sum_{k=1-2n}^{2n-1} \langle C^T e_{m+k}, C^T e_m\rangle U_{m+k}(\lambda)/U_m(\lambda) \\ &= \sum_{k=1-2n}^{2n-1}\sum_{j=0}^{2n-1} c_{m+k,m+j}c_{m,m+j}U_{m+k}(\lambda)/U_m(\lambda) \\ &= \sum_{k=1-2n}^{2n-1}\sum_{j=0}^{2n-1} t_{j-k}t_j U_{m+k}(\lambda)/U_m(\lambda)\end{aligned}$$

$$= \sum_{k=0}^{2n-1} \sum_{j=0}^{2n-1} t_k t_j U_{m+k-j}(\lambda)/U_m(\lambda).$$

By equation (4.104), this tends to $\sum_{k=0}^{2n-1} \sum_{j=0}^{2n-1} t_k t_j z^{j-k}$ as $m \to \infty$. This is equal to $c(z)c(z^{-1})$. □

**Lemma 4.3.19.** *Let $p_C^\mu(\lambda) = \sum_{k=0}^{2n-1} \langle e_k, C^\mu C^T e_0 \rangle U_k(\lambda)$ as in Theorem 4.3.12 and let $c_\mu$ be the symbol of $C^\mu_{\mathrm{Toe}}$, the Toeplitz part of $C^\mu$ as guaranteed by Corollary 4.3.10. Then*

$$p_C^\mu(\lambda(z)) = c(z^{-1})c_\mu(z) - 2z, \tag{4.108}$$

*where $\lambda(z) = \frac{1}{2}(z + z^{-1})$ and $z \in \mathbb{D}$.*

*Proof.* The key quantity to observe for this proof is

$$\frac{1}{U_m(\lambda)} \left\langle \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, C^\mu C^T e_m \right\rangle \tag{4.109}$$

as $m \to \infty$. We will compute two equivalent expressions for this quantity to derive equation (4.108).

By Corollary 4.2.2 on the $m$th row of $C$, we have $C^T e_m = U_m(J) C^T e_0$. Hence, using the definition of $C^\mu$ for the second line,

$$\left\langle \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, C^\mu C^T e_m \right\rangle = \left\langle \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, C^\mu U_m(J) C^T e_0 \right\rangle \tag{4.110}$$

$$= \left\langle \begin{pmatrix} P_0^\mu(\lambda) \\ P_1^\mu(\lambda) \\ \vdots \end{pmatrix}, U_m(J) C^T e_0 \right\rangle. \tag{4.111}$$

By Lemma 4.1.8 with $f = U_m$,

$$U_m(J) \begin{pmatrix} P_0^\mu(\lambda) \\ P_1^\mu(\lambda) \\ \vdots \end{pmatrix} = U_m(\lambda) \begin{pmatrix} P_0^\mu(\lambda) \\ P_1^\mu(\lambda) \\ \vdots \end{pmatrix} + \frac{U_m(J) - U_m(\lambda)}{J - \lambda} e_0. \tag{4.112}$$

Combining equation (4.111) and equation (4.112) gives

$$\left\langle \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, C^\mu C^T e_m \right\rangle = \underbrace{U_m(\lambda) \left\langle \begin{pmatrix} P_0^\mu(\lambda) \\ P_1^\mu(\lambda) \\ \vdots \end{pmatrix}, C^T e_0 \right\rangle}_{(1)} + \underbrace{\left\langle e_0, \frac{U_m(J) - U_m(\lambda)}{J - \lambda} C^T e_0 \right\rangle}_{(2)}. \tag{4.113}$$

Part (1) yields,

$$(1) = U_m(\lambda) \sum_{k=1}^{2n-1} c_{0,k} P_k^\mu(\lambda) = U_m(\lambda) p^\mu(\lambda). \tag{4.114}$$

By Theorem 4.2.3 with $f(\zeta) = (U_m(\zeta) - U_m(\lambda))/(\zeta - \lambda)$, part (2) yields

$$(2) = \left\langle Ce_0, (U_m(\Delta) - U_m(\lambda))(\Delta - \lambda)^{-1} e_0 \right\rangle \tag{4.115}$$

Since $Ce_0 = e_0$ and using Theorem 4.1.5, we further have

$$(2) = \int \frac{U_m(s) - U_m(\lambda)}{s - \lambda} \mathrm{d}\mu_\Delta(s). \tag{4.116}$$

This is the $\mu_\Delta$-derivative of $U_m$ as in Definition 4.1.10. We noted in equation (4.65) that $U_m^{\mu_\Delta}(\lambda) = 2U_{m-1}(\lambda)$.

Combining (1) and (2) we have as $m \to \infty$,

$$\frac{1}{U_m(\lambda)} \left\langle \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, C^\mu C^T e_m \right\rangle = p_C^\mu(\lambda) + 2U_{m-1}(\lambda)/U_m(\lambda) \to p_C^\mu(\lambda) + 2z.$$

Now we compute the quantity in equation (4.109) in a different direction. By Corollary 4.3.10, $C^\mu = C^\mu_{\text{Toe}} + C^\mu_{\text{fin}}$, where $C^\mu_{\text{fin}}$ is zero outside the principal $(n-2) \times (2n-1)$ submatrix. Hence for $m$ sufficiently large we have $(c_\mu)_{m,m+k} = t_k^\mu$ for a sequence $(t_k^\mu)_{k\in\mathbb{Z}}$ such that $t_k = 0$ for $k \notin \{1, \ldots, 2n-2\}$. The Toeplitz symbol of $C^\mu_{\text{Toe}}$ is $c_\mu(z) = \sum_{k=1}^{2n-2} t_k^\mu z^k$. Hence we have for $m$ sufficiently large,

$$\frac{1}{U_m(\lambda)} \left\langle \begin{pmatrix} U_0(\lambda) \\ U_1(\lambda) \\ \vdots \end{pmatrix}, C^\mu C^T e_m \right\rangle = \sum_{k=1-2n}^{2n-1} \langle (C^\mu)^T e_{m+k}, C^T e_m \rangle U_{m+k}(\lambda)/U_m(\lambda)$$

$$
\begin{aligned}
&= \sum_{k=1-2n}^{2n-1} \sum_{j=0}^{2n-1} (c_\mu)_{m+k,m+j} c_{m,m+j} U_{m+k}(\lambda)/U_m(\lambda) \\
&= \sum_{k=1-2n}^{2n-1} \sum_{j=0}^{2n-1} t^\mu_{j-k} t_j U_{m+k}(\lambda)/U_m(\lambda) \\
&= \sum_{i=1}^{2n-2} \sum_{j=0}^{2n-1} t^\mu_i t_j U_{m+j-i}(\lambda)/U_m(\lambda).
\end{aligned}
$$

By equation (4.104), this tends to $\sum_{i=1}^{2n-2} \sum_{j=0}^{2n-1} t^\mu_i t_j z^{i-j}$ as $m \to \infty$. This is equal to $c(z^{-1})c_\mu(z)$.

Equating these two quantities gives $p^\mu(\lambda(z)) = c(z^{-1})c_\mu(z) - 2z$ as required. □

The following theorem describes Theorem 4.3.12 under the change of variables induced by the Joukowski map.

**Theorem 4.3.20.** *Let $J$ be a Jacobi operator such that there exists an $n$ such that*

$$\alpha_k = 0, \quad \beta_{k-1} = \frac{1}{2} \text{ for all } k \geq n, \tag{4.117}$$

*i.e. it is equal to the free Jacobi operator $\Delta$ outside the $n \times n$ principal submatrix. By Theorem 4.3.8 the connection coefficient matrix can be decomposed into $C = C_{\text{Toe}} + C_{\text{fin}}$. By Corollary 4.3.10, we similarly have $C^\mu = C^\mu_{\text{Toe}} + C^\mu_{\text{fin}}$. If $c$ and $c_\mu$ are the Toeplitz symbols of $C_{\text{Toe}}$ and $C^\mu_{\text{Toe}}$ respectively, then for $\lambda(z) = \frac{1}{2}(z + z^{-1})$ with $z \in \mathbb{D}$, the principal resolvent $G$ is given by the rational function*

$$G(\lambda(z)) = -\frac{c_\mu(z)}{c(z)}. \tag{4.118}$$

*Proof.* Combining Theorem 4.3.12, equation (4.103) and Lemmata 4.3.18 and 4.3.19, we have

$$
\begin{aligned}
G(\lambda(z)) &= \frac{G_\Delta(\lambda(z)) - p^\mu(\lambda(z))}{p(\lambda(z))} && (4.119) \\
&= \frac{-2z - (c(z^{-1})c_\mu(z) - 2z)}{c(z)c(z^{-1})} && (4.120) \\
&= -\frac{c_\mu(z)}{c(z)}. && (4.121)
\end{aligned}
$$

This completes the proof. □

The following theorem gives a better description of the weights $w_k$ in Theorem 4.3.14, utilising the Joukowski map and the Toeplitz symbol $c$.

**Theorem 4.3.21.** *Let $J$ be a Jacobi operator such that there exists an $n$ such that*

$$\alpha_k = 0, \quad \beta_{k-1} = \frac{1}{2} \text{ for all } k \geq n, \tag{4.122}$$

*i.e. it is equal to the free Jacobi operator $\Delta$ outside the $n \times n$ principal submatrix. By Theorem 4.3.8 the connection coefficient matrix can be written $C = C_{\text{Toe}} + C_{\text{fin}}$. If $c$ is the Toeplitz symbol of $C_{\text{Toe}}$, then the spectral measure of $J$ is*

$$\mu(s) = \frac{1}{p_C(s)}\mu_\Delta(s) + \sum_{k=1}^{r} \frac{(z_k - z_k^{-1})^2}{z_k c'(z_k) c(z_k^{-1})}\delta_{\lambda(z_k)}(s). \tag{4.123}$$

*Here $z_i$ are the roots of $c$ that lie in the open unit disk, which are all real and simple. The only roots of $c$ on the unit circle are $\pm 1$, which can also only be simple. Further, $r \leq n$.*

*Proof.* By Theorem 4.3.14,

$$\mu(s) = \frac{1}{p_C(s)}\mu_\Delta(s) + \sum_{k=1}^{r} w_k \delta_{\lambda_k}(s), \tag{4.124}$$

where $r \leq n$. Hence we just need to prove something more specific about the roots of $c$, $\lambda_1, \ldots, \lambda_r$, and $w_1, \ldots, w_r$.

By Theorem 4.3.20, $G(\lambda(z)) = -c_\mu(z)/c(z)$ for $z \in \mathbb{D}$. By Lemma 4.3.19, $c(z^{-1})c_\mu(z) - 2z = p^\mu(\lambda(z)) = p^\mu(\lambda(z^{-1})) = c(z)c_\mu(z^{-1}) - 2z^{-1}$, so

$$c(z^{-1})c_\mu(z) - c(z)c_\mu(z^{-1}) = 2(z - z^{-1}). \tag{4.125}$$

Therefore $c$ and $c_\mu$ cannot simultaneously be zero unless $z = z^{-1}$, which only happens at $z = \pm 1$. By the same reasoning, $c(z)$ and $c(z^{-1})$ also cannot be simultaneously zero unless $z = \pm 1$. Since the Joukowski map $\lambda$ is a bijection from $\mathbb{D}$ to $\mathbb{C} \setminus [-1, 1]$, this shows that the (simple and real) poles of $G$ in $\mathbb{C} \setminus [-1, 1]$ are precisely $\lambda(z_1), \ldots, \lambda(z_r)$, where $z_1, \ldots, z_r$ are the (necessarily simple and real) roots of $c$ in $\mathbb{D}$.

What are the values of the weights of the Dirac deltas, $w_1, \dots, w_r$? By Theorem 4.3.14,

$$\begin{aligned} w_k &= \lim_{\varepsilon \searrow 0} \frac{\varepsilon}{i} G(\lambda(z_k) + i\varepsilon) \\ &= \lim_{\lambda \to \lambda(z_k)} (\lambda(z_k) - \lambda) G(\lambda) \\ &= \lim_{z \to z_k} \frac{1}{2}(z_k + z_k^{-1} - z - z^{-1})(-1)\frac{c_\mu(z)}{c(z)} \\ &= \lim_{z \to z_k} \frac{1}{2} z^{-1}(z - z_k)(z - z_k^{-1})\frac{c_\mu(z)}{c(z)} \\ &= \frac{1}{2} z_k^{-1}(z_k - z_k^{-1}) c_\mu(z_k) \lim_{z \to z_k} \frac{(z - z_k)}{c(z)} \\ &= \frac{1}{2} z_k^{-1}(z_k - z_k^{-1})\frac{c_\mu(z_k)}{c'(z_k)}. \end{aligned}$$

By equation (4.125), since $c(z_k) = 0$, we have $c_\mu(z_k) = 2(z_k - z_k^{-1})/c(z_k^{-1})$. This gives

$$w_k = \frac{(z_k - z_k^{-1})^2}{z_k c(z_k^{-1}) c'(z_k)}.$$

Note that if $c(z) = 0$ then $c(\overline{z}) = 0$ because $c$ has real coefficients. If $c$ has a root $z_0$ on the unit circle, then $c(z_0) = c(z_0^{-1}) = 0$ because $\overline{z_0} = z_0^{-1}$, which earlier in the proof we showed only occurs if $z_0 = \pm 1$. Hence $c$ does not have roots on the unit circle except possibly $\pm 1$. □

**Example 4.3.22** (Basic perturbation 1 re-revisited)**.** Considering the connection coefficient matrix in equation (4.67), we see that the Toeplitz symbol $c$ is $c(z) = 1 - \alpha z$. By Theorem 4.3.21 the roots of $c$ in the unit disc correspond to eigenvalues of $J_\alpha$. As is consistent with our previous considerations, $c$ has a root in the unit disc if and only if $|\alpha| > 1$, and those eigenvalues are $\lambda(\alpha^{-1}) = \frac{1}{2}(\alpha + \alpha^{-1})$. See Section 4.6 for figures depicting the spectral measure and the resolvent.

**Example 4.3.23** (Basic perturbation 2 re-revisited)**.** Considering the connection coefficient matrix in equation (4.70), we see that the Toeplitz symbol $c$ is $c(z) = \beta^{-1} + (\beta^{-1} - \beta)z^2$. By Theorem 4.3.21 the roots of $c$ in the unit disc correspond to eigenvalues of $J_\beta$. The roots of $c$ are $\pm\frac{1}{\sqrt{\beta^2 - 1}}$. If $\beta \in \left(0, \sqrt{2}\right] \setminus \{1\}$ then $\left|\pm\frac{1}{\sqrt{\beta^2-1}}\right| \geq 1$ so there are no roots of $c$ in the unit disc, as is consistent with the previous observations. What was difficult to see before is, if $\beta > \sqrt{2}$ then $\left|\pm\frac{1}{\sqrt{\beta^2-1}}\right| < 1$, so there is a root of

$c$ inside $\mathbb{D}$, and it corresponds to an eigenvalue,

$$\lambda\left(\pm\frac{1}{\sqrt{\beta^2-1}}\right)=\pm\frac{1}{2}\left(\frac{1}{\sqrt{\beta^2-1}}+\sqrt{\beta^2-1}\right)=\pm\frac{\beta^2}{2\sqrt{\beta^2-1}}. \tag{4.126}$$

See Section 4.6 for figures depicting the spectral measure and the resolvent.

## 4.4 Toeplitz-plus-trace-class Jacobi operators

In this section we extend the results of the previous section to the case where the Jacobi operator is Toeplitz-plus-trace-class. This cannot be done as a direct extension of the work in the previous section as the formulae obtained depended on the fact that some of the functions involved were merely polynomials in order to have a function defined for all $\lambda$ in an a priori known region of the complex plane. We admit that it may be possible to perform the analysis directly, but state that it is not straightforward. We are interested in feasible (finite) computation so are content to deal directly with the Toeplitz-plus-finite-rank case and perform a limiting process. The crucial question for computation is, can we approximate the spectral measure of a Toeplitz-plus-trace-class Jacobi operator whilst reading only finitely many entries of the matrix?

Here we make clear the definition of a Toeplitz-plus-trace-class Jacobi operator.

**Definition 4.4.1.** An operator $K:\ell^2\to\ell^2$ is said to be trace class if $\sum_{k=0}^{\infty}e_k^T(K^TK)^{1/2}e_k<\infty$. Hence we say that a Jacobi operator $J$ such that $\alpha_k\to 0$, $\beta_k\to\frac{1}{2}$ as $k\to\infty$ is Toeplitz-plus-trace-class if

$$\sum_{k=0}^{\infty}\left|\beta_k-\frac{1}{2}\right|+|\alpha_k|<\infty. \tag{4.127}$$

### 4.4.1 Jacobi operators for Jacobi polynomials

The most well known class of orthogonal polynomials is the Jacobi polynomials, whose measure of orthogonality is

$$\mathrm{d}\mu(s)=\left(2^{\alpha+\beta+1}B(\alpha+1,\beta+1)\right)^{-1}(1-s)^{\alpha}(1+s)^{\beta}\Big|_{s\in[-1,1]}\mathrm{d}s, \tag{4.128}$$

where $\alpha$,$\beta>-1$ and $B$ is Euler's Beta function. The Jacobi operator for the normalised Jacobi polynomials with respect to this probability measure, and hence the three term

recurrence coefficients, are given by [olva],

$$\alpha_k = \frac{\beta^2 - \alpha^2}{(2k+\alpha+\beta)(2k+\alpha+\beta+2)} \tag{4.129}$$

$$\beta_{k-1} = 2\sqrt{\frac{k(k+\alpha)(k+\beta)(k+\alpha+\beta)}{(2k+\alpha+\beta-1)(2k+\alpha+\beta)^2(2k+\alpha+\beta+1)}} \tag{4.130}$$

Note that $|\alpha_k| = \mathcal{O}(k^{-2})$ and

$$\beta_{k-1} = \frac{1}{2}\sqrt{1 + \frac{(4-8\alpha^2-8\beta^2)k^2 + \mathcal{O}(k)}{(2k+\alpha+\beta-1)(2k+\alpha+\beta)^2(2k+\alpha+\beta+1)}} = \frac{1}{2} + \mathcal{O}(k^{-2}). \tag{4.131}$$

Hence the Jacobi operators for the Jacobi polynomials are Toeplitz-plus-trace-class for all $\alpha, \beta > -1$.

The Chebyshev polynomials $T_k$ and $U_k$ discussed in the previous section are specific cases of Jacobi polynomials, with $\alpha, \beta = -\frac{1}{2}, -\frac{1}{2}$ for $T_k$ and $\alpha, \beta = \frac{1}{2}, \frac{1}{2}$ for $U_k$.

In Section 4.6 numerical computations of the spectral measures and resolvents of these Jacobi operators are presented.

### 4.4.2 Toeplitz-plus-finite-rank truncations

We propose to use the techniques from Section 4.3. Therefore for a Jacobi operator $J$, we can define the Toeplitz-plus-finite-rank truncations $J^{[m]}$, where

$$J^{[m]}_{i,j} = \begin{cases} J_{i,j} & \text{if } 0 \le i,j < m \\ \Delta_{i,j} & \text{otherwise.} \end{cases} \tag{4.132}$$

Each Jacobi operator $J^{[m]}$ has a spectral measure $\mu^{[m]}$ which can be computed using Theorem 4.3.21. The main question for this section is: how do the computable measures $\mu^{[m]}$ approximate the spectral measure $\mu$ of $J$?

**Proposition 4.4.2.** *Let $J$ a Jacobi operator (bounded, but with no assumed structure imposed) and let $\mu$ be its spectral measure. Then the measures $\mu^{[1]}, \mu^{[2]}, \ldots$ which are the spectral measures of $J^{[1]}, J^{[2]}, \ldots$ converge to $\mu$ in a weak sense. Precisely,*

$$\lim_{m\to\infty} \int f(s)\, \mathrm{d}\mu^{[m]}(s) = \int f(s)\, \mathrm{d}\mu(s), \tag{4.133}$$

*for all $f \in C_b(\mathbb{R})$.*

*Proof.* Each spectral measure $\mu^{[m]}$ and $\mu$ are supported on the spectra of $J^{[m]}$ and $J$, each of which are contained within $[-\|J^{[m]}\|_2, \|J^{[m]}\|_2]$ and $[-\|J\|_2, \|J\|_2]$. Since $\|J^{[m]}\|_2$ and $\|J\|_2$ are less than

$$M = 3\left(\sup_{k\geq 0}|\alpha_k| + \sup_{k\geq 0}|\beta_k|\right), \tag{4.134}$$

we have that all the spectral measures involved are supported within the interval $[-M, M]$. Hence we can consider integrating functions $f \in C([-M, M])$ without ambiguity.

By Weierstrass' Theorem, polynomials are dense in $C([-M, M])$, so we only need to consider polynomials as test functions, and by linearity we only need to consider the orthogonal polynomials for $J$. The first polynomial $P_0$ has immediate convergence, since the measures are all probability measures. Now consider $P_k$ for some $k > 0$, which satisfies $\int P_k(s)\,\mathrm{d}\mu(s) = 0$. For $m > k$, $P_k$ is also the $k$th orthogonal polynomial for $J^{[m]}$, hence $\int P_k(s)\,\mathrm{d}\mu^{[m]}(s) = 0$. This completes the proof. □

### 4.4.3 Asymptotics of the connection coefficients

Here we formulate a block operator equation $\mathcal{L}\underline{c} = e_0^0$ satisfied by the entries of the connection coefficient matrices encoded into a vector $\underline{c}$. For Toeplitz-plus-trace-class Jacobi operators we give appropriate Banach spaces upon which the operator $\mathcal{L}$ is bounded and invertible, enabling precise results about the asymptotics of the connection coefficients to be derived.

**Lemma 4.4.3.** *Let $J$ and $D$ be Jacobi operators with entries $\{\alpha_k, \beta_k\}_{k=0}^{\infty}$ and $\{\gamma_k, \delta_k\}_{k=0}^{\infty}$ respectively. If we decompose the upper triangular part of $C_{J\to D}$ into a sequence of sequences, stacking each diagonal on top of each other, we get the following block linear system,*

$$\begin{pmatrix} B_{-1} & & & & \\ A_0 & B_0 & & & \\ B_0^T & A_1 & B_1 & & \\ & B_1^T & A_2 & B_2 & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} c_{*,*} \\ c_{*,*+1} \\ c_{*,*+2} \\ c_{*,*+3} \\ \vdots \end{pmatrix} = \begin{pmatrix} e_0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \tag{4.135}$$

*where for each $i$,*

$$B_i = 2\begin{pmatrix} \beta_i & & & \\ -\delta_0 & \beta_{i+1} & & \\ & -\delta_1 & \beta_{i+2} & \\ & & \ddots & \ddots \end{pmatrix}, \quad A_i = 2\begin{pmatrix} \alpha_i - \gamma_0 & & & \\ & \alpha_{i+1} - \gamma_1 & & \\ & & \alpha_{i+2} - \gamma_2 & \\ & & & \ddots \end{pmatrix}.$$

*For $B_{-1}$ to make sense we define $\beta_{-1} = 1/2$.*

*Proof.* This is simply the 5-point discrete system in Lemma 4.2.1 rewritten. □

We write the infinite dimensional block infinite dimensional system (4.135) in the form,

$$\mathcal{L}\underline{c} = e_0^0. \tag{4.136}$$

For general Jacobi operators $J$ and $D$, the operators $A_i$ and $B_i$ are well defined linear operators from $\ell^\star_{\mathcal{F}}$ to $\ell^\star_{\mathcal{F}}$. The block operator $\mathcal{L}$ is whence considered as a linear operator from the space of sequences of real sequences, $\ell^\star_{\mathcal{F}}(\ell^\star_{\mathcal{F}})$ to itself. We will use this kind of notation for other spaces as follows.

**Definition 4.4.4** (Vector-valued sequences)**.** If $\ell_X$ is a vector space of scalar-valued sequences, and $Y$ is another vector space then we let $\ell_X(Y)$ denote the vector space of sequences of elements of $Y$. In many cases in which $\ell_X$ and $Y$ are both normed spaces, then $\ell_X(Y)$ naturally defines a normed space in which the norm is derived from that of $\ell_X$ by replacing all instances of absolute value with the norm on $Y$. For example, $\ell^p(\ell^\infty)$ is a normed space with norm $\|(a_k)_{k=0}^\infty\|_{\ell^p(\ell^\infty)} = (\sum_{k=0}^\infty \|a_k\|_\infty^p)^{\frac{1}{p}}$.

The following two spaces are relevant for the Toeplitz-plus-trace-class Jacobi operators.

**Definition 4.4.5** (Sequences of bounded variation)**.** Following [DSBB71, Ch. IV.2.3], denote by $bv$ the Banach space of all sequences with bounded variation, that is sequences such that the norm

$$\|a\|_{bv} = |a_0| + \sum_{k=0}^{\infty} |a_{k+1} - a_k|, \tag{4.137}$$

is finite.

The following result is immediate from the definition of the norm on $bv$.

**Lemma 4.4.6.** *There is a continuous embedding $bv \subset c$ the Banach space of convergent sequences i.e. for all $(a_k)_{k=0}^{\infty} \in bv$, $\lim_{k\to\infty} a_k$ exists, and $\sup_k |a_k| \leq \|(a_k)_{k=0}^{\infty}\|_{bv}$. Furthermore, $\lim_{k\to\infty} |a_k| \leq \|a\|_{bv}$.*

**Definition 4.4.7** (Geometrically weighted $\ell^1$)**.** We define the Banach space $\ell^1_R$ to be the space of sequences such that the norm

$$\|v\|_{\ell^1_R} = \sum_{k=0}^{\infty} R^k |v_k|, \tag{4.138}$$

is finite.

**Proposition 4.4.8.** *The operator norm on $\ell^1_R$ is equal to*

$$\|A\|_{\ell^1_R \to \ell^1_R} = \sup_j \sum_i R^{i-j} |a_{ij}|. \tag{4.139}$$

The following Lemma and its Corollary show that it is natural to think of $\underline{c}$ as lying in the space $\ell^1_R(bv)$.

**Lemma 4.4.9.** *Let $J = \Delta + K$ be a Jacobi operator where $K$ is trace class and let $D = \Delta$. Then for any $R \in (0,1)$ the operator $\mathcal{L}$ in equation (4.136) is bounded and invertible as an operator from $\ell^1_R(bv)$ to $\ell^1_R(\ell^1)$. Furthermore, if $\mathcal{L}^{[m]}$ is the operator in equation (4.136) generated by the Toeplitz-plus-finite-rank truncation $J^{[m]}$, then*

$$\|\mathcal{L} - \mathcal{L}^{[m]}\|_{\ell^1_R(bv) \to \ell^1_R(\ell^1)} \to 0 \text{ as } m \to \infty. \tag{4.140}$$

*Proof.* We can write $\mathcal{L}$ in equation (4.136) in the form $\mathcal{L} = \mathcal{T} + \mathcal{K}$ where

$$\mathcal{T} = \begin{pmatrix} T & & & & \\ 0 & T & & & \\ T^T & 0 & T & & \\ & T^T & 0 & T & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \quad T = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & -1 & 1 & \\ & & \ddots & \ddots \end{pmatrix}, \tag{4.141}$$

and

$$\mathcal{K} = \begin{pmatrix} K_{-1} & & & & \\ A_0 & K_0 & & & \\ K_0 & A_1 & K_1 & & \\ & K_1 & A_2 & K_2 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \qquad \begin{aligned} A_i &= 2\mathrm{diag}(\alpha_i, \alpha_{i+1}, \ldots), \\ K_i &= \mathrm{diag}(2\beta_i - 1, 2\beta_{i+1} - 1, \ldots). \end{aligned} \tag{4.142}$$

This decomposition will allow us to prove that $\mathcal{L}$ is bounded and invertible as follows. We will show that as operators from $\ell^1_R(bv)$ to $\ell^1_R(\ell^1)$, $\mathcal{T}$ is bounded and invertible, and $\mathcal{K}$ is compact. This implies that $\mathcal{L}$ is a Fredholm operator with index 0. Therefore, $\mathcal{L}$ is invertible if and only if it is injective. It is indeed injective, because it is block lower triangular with invertible diagonal blocks, so forward substitution on the system $\mathcal{L}\underline{v} = \underline{0}$ implies that each entry of $\underline{v}$ must be zero.

First let us prove that $\mathcal{T}$ is bounded and invertible. It is elementary that $T$ is an isometric isomorphism from $bv$ to $\ell^1$ and $T^T$ is bounded with norm at most 1. Hence using Proposition 4.4.8 we have

$$\|\mathcal{T}\|_{\ell^1_R(bv)\to\ell^1_R(\ell^1)} = R^0\|T\|_{bv\to\ell^1} + R^2\|T^T\|_{bv\to\ell^1} \le 1 + R^2. \tag{4.143}$$

Because each operator is lower triangular, the left and right inverse of $\mathcal{T} : \ell_{\mathcal{F}}(\ell_{\mathcal{F}}) \to \ell_{\mathcal{F}}(\ell_{\mathcal{F}})$ is

$$\mathcal{T}^{-1} = \begin{pmatrix} T^{-1} & & & & & \\ 0 & T^{-1} & & & & \\ -T^{-1}T^TT^{-1} & 0 & T^{-1} & & & \\ 0 & -T^{-1}T^TT^{-1} & 0 & T^{-1} & & \\ T^{-1}(T^TT^{-1})^2 & 0 & -T^{-1}T^TT^{-1} & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{4.144}$$

This matrix is block lower triangular and block Toeplitz with first column having $2i$th block of the form $T^{-1}(-T^TT^{-1})^i$ and $(2i+1)$th block zero. We must check that this matrix is bounded in the norms on $\ell^1_R(\ell^1)$ to $\ell^1_R(bv)$ so that it may be extended to those spaces from $\ell_{\mathcal{F}}$. Again using Proposition 4.4.8 we have

$$\|\mathcal{T}^{-1}\|_{\ell^1_R(\ell^1)\to\ell^1_R(bv)} = \sup_j \sum_{i=j}^{\infty} R^{2(i-j)}\|T^{-1}(-T^TT^{-1})^{i-j}\|_{\ell^1\to bv}$$

$$
\begin{aligned}
&= \sum_{k=0}^{\infty} R^{2k} \|T^{-1}(-T^T T^{-1})^k\|_{\ell^1 \to bv} \\
&\leq \sum_{k=0}^{\infty} R^{2k} \|T^{-1}\|_{\ell^1 \to bv} \left(\|T^T\|_{bv \to \ell^1} \|T^{-1}\|_{\ell^1 \to bv}\right)^k \\
&\leq \sum_{k=0}^{\infty} R^{2k} = (1-R^2)^{-1} < \infty.
\end{aligned}
$$

Now let us prove that $\mathcal{K} : \ell^1_R(bv) \to \ell^1_R(\ell^1)$ is compact. Consider the finite rank operator $\mathcal{K}^{[m]}$, where all elements are the same as in $\mathcal{K}$, except that all occurrences of $\alpha_i$ and $2\beta_i - 1$ are replaced by 0 for $i \geq m$. Using Proposition 4.4.8 we have

$$
\|\mathcal{K} - \mathcal{K}^{[m]}\|_{\ell^1_R(bv) \to \ell^1_R(\ell^1)} = \sup_j R^0 \|K_{j-1} - K_{j-1}^{[m]}\|_{bv \to \ell^1} + R^1 \|A_j - A_j^{[m]}\|_{bv \to \ell^1} + R^2 \|K_j - K_j^{[m]}\|_{bv \to \ell^1}. \tag{4.145}
$$

By the continuous embedding in Lemma 4.4.6, $\|\cdot\|_{bv \to \ell^1} \leq \|\cdot\|_{\ell^\infty \to \ell^1}$. Hence

$$
\begin{aligned}
\|\mathcal{K} - \mathcal{K}^{[m]}\|_{\ell^1_R(bv) \to \ell^1_R(\ell^1)} &\leq \sum_{k=m}^{\infty} R^0 |2\beta_{k-1} - 1| + R^1 |\alpha_k| + R^2 |2\beta_k - 1| \\
&\to 0 \text{ as } m \to \infty.
\end{aligned}
$$

Since $\mathcal{K}$ is a norm limit of finite rank operators it is compact. This completes the proof that $\mathcal{L}$ is bounded and invertible.

Now consider the operator $\mathcal{L}^{[m]}$, which is equal to $\mathcal{T} + \mathcal{K}^{[m]}$ (where $\mathcal{K}^{[m]}$ is precisely that which was considered whilst proving $\mathcal{K}$ is compact). Hence,

$$
\|\mathcal{L} - \mathcal{L}^{[m]}\|_{\ell^1_R(bv) \to \ell^1_R(\ell^1)} = \|\mathcal{K} - \mathcal{K}^{[m]}\|_{\ell^1_R(bv) \to \ell^1_R(\ell^1)} \to 0 \text{ as } m \to \infty. \tag{4.146}
$$

This completes the proof. □

**Corollary 4.4.10.** *Let $J = \Delta + K$ be a Jacobi operator where $K$ is trace class and let $\underline{c} \in \ell^\star_{\mathcal{F}}(\ell^\star_{\mathcal{F}})$ be the vector of diagonals of $C_{J \to \Delta}$ as in equation (4.136). Then $\underline{c} \in \ell^1_R(bv)$. If $J$ has Toeplitz-plus-finite-rank truncations $J^{[m]}$ and $\underline{c}^{[m]}$ denotes the vector of diagonals of $C^{[m]}$, then*

$$
\|\underline{c} - \underline{c}^{[m]}\|_{\ell^1_R(bv)} \to 0 \text{ as } m \to \infty. \tag{4.147}
$$

*Proof.* By equation (4.136)

$$\underline{c} - \underline{c}^{[m]} = (\mathcal{L}^{-1} - (\mathcal{L}^{[m]})^{-1})e_0^0. \tag{4.148}$$

Since $\|e_0^0\|_{\ell^1_R(\ell^1)} = 1$, the proof is completed if we show $\|\mathcal{L}^{-1} - (\mathcal{L}^{[m]})^{-1}\|_{\ell^1_R(\ell^1)\to\ell^1_R(bv)} \to 0$ as $m \to \infty$.

Suppose that $m$ is sufficiently large so that $\|\mathcal{L} - \mathcal{L}^{[m]}\| < \|\mathcal{L}^{-1}\|^{-1}$. Note that $\mathcal{L}^{-1}$ is bounded by the Inverse Mapping Theorem and Lemma 4.4.9. Then by a well-known result (see for example, [TO16], [AH05]),

$$\|\mathcal{L}^{-1} - (\mathcal{L}^{[m]})^{-1}\| \leq \frac{\|\mathcal{L}^{-1}\|^2\|\mathcal{L} - \mathcal{L}^{[m]}\|}{1 - \|\mathcal{L}^{-1}\|\|\mathcal{L} - \mathcal{L}^{[m]}\|}. \tag{4.149}$$

This tends to zero as $m \to \infty$, by Lemma 4.4.9. □

**Theorem 4.4.11.** *Let $J = \Delta + K$ be a Jacobi operator where $K$ is trace class. Then $C = C_{J\to\Delta}$ can be decomposed into*

$$C = C_{\text{Toe}} + C_{\text{com}}, \tag{4.150}$$

*where $C_{\text{Toe}}$ is upper triangular, Toeplitz and bounded as an operator from $\ell^1_{R^{-1}}$ to $\ell^1_{R^{-1}}$, and $C_{\text{com}}$ is compact as an operator from $\ell^1_{R^{-1}}$ to $\ell^1_{R^{-1}}$, for all $R \in (0,1)$. Also, if $J$ has Toeplitz-plus-finite-rank truncations $J^{[m]}$ with connection coefficient matrices $C^{[m]} = C^{[m]}_{\text{Toe}} + C^{[m]}_{\text{com}}$, then*

$$C^{[m]} \to C, \quad C^{[m]}_{\text{Toe}} \to C_{\text{Toe}}, \quad C^{[m]}_{\text{com}} \to C_{\text{com}} \text{ as } m \to \infty, \tag{4.151}$$

*in the operator norm topology over $\ell^1_{R^{-1}}$.*

*Proof.* By Lemma 4.4.9, for each $k$ the sequence $(c_{0,0+k}, c_{1,1+k}, c_{2,2+k}, \ldots)$ is an element of $bv$. By Lemma 4.4.6 each is therefore a convergent sequence, whose limits we call $t_k$. Hence we can define an upper triangular Toeplitz matrix $C_{\text{Toe}}$ whose $(i,j)$th element is $t_{j-i}$, and define $C_{\text{com}} = C - C_{\text{Toe}}$.

The Toeplitz matrix $C_{\text{Toe}}$ is bounded from $\ell^1_{R^{-1}}$ to $\ell^1_{R^{-1}}$ by the following calculation.

$$\|C_{\text{Toe}}\|_{\ell^1_{R^{-1}}\to\ell^1_{R^{-1}}} = \sup_j \sum_{i=0}^{j} R^{j-i}|t_{j-i}| \tag{4.152}$$

$$= \sum_{k=0}^{\infty} R^k |t_k| \tag{4.153}$$

$$\leq \sum_{k=0}^{\infty} R^k \|c_{*,*+k}\|_{bv} \tag{4.154}$$

$$= \|\underline{c}\|_{\ell^1_R(bv)}. \tag{4.155}$$

By Lemma 4.4.9 this quantity is finite.

Now we show convergence results. The compactness of $C_{\mathrm{com}}$ will follow at the end.

$$\|C - C^{[m]}\|_{\ell^1_{R^{-1}} \to \ell^1_{R^{-1}}} = \sup_j \sum_{i=0}^{j} R^{j-i} |c_{i,j} - c^{[m]}_{i,j}| \tag{4.156}$$

$$= \sup_j \sum_{k=0}^{j} R^k |c_{j-k,j} - c^{[m]}_{j-k,j}| \tag{4.157}$$

$$\leq \sup_j \sum_{k=0}^{j} R^k \|c_{*,*+k} - c^{[m]}_{*,*+k}\|_{bv} \tag{4.158}$$

$$= \sum_{k=0}^{\infty} R^k \|c_{*,*+k} - c^{[m]}_{*,*+k}\|_{bv} \tag{4.159}$$

$$= \|\underline{c} - \underline{c}^{[m]}\|_{\ell^1_R(bv)}. \tag{4.160}$$

For the third line, note that for fixed $k$, $c_{0,k} - c^{[m]}_{0,k}, c_{1,1+k} - c^{[m]}_{1,1+k}, c_{2,2+k} - c^{[m]}_{2,2+k}, \ldots$ is a $bv$ sequence, and refer to Lemma 4.4.6.

$$\|C_{\mathrm{Toe}} - C^{[m]}_{\mathrm{Toe}}\|_{\ell^1_{R^{-1}} \to \ell^1_{R^{-1}}} = \sup_j \sum_{i=0}^{j} R^{j-i} |t_{j-i} - t^{[m]}_{j-i}| \tag{4.161}$$

$$= \sum_{k=0}^{j} R^k |t_k - t^{[m]}_k| \tag{4.162}$$

$$= \sum_{k=0}^{\infty} R^k \|c_{*,*+k} - c^{[m]}_{*,*+k}\|_{bv} \tag{4.163}$$

$$= \|\underline{c} - \underline{c}^{[m]}\|_{\ell^1_R(bv)}. \tag{4.164}$$

For the third line, note that $t_k - t^{[m]}_k$ is the limit of the $bv$ sequence $c_{*,*+k} - c^{[m]}_{*,*+k}$, and refer Lemma 4.4.6.

$$\|C_{\mathrm{com}} - C^{[m]}_{\mathrm{com}}\|_{\ell^1_{R^{-1}} \to \ell^1_{R^{-1}}} \leq \|C - C^{[m]}\| + \|C_{\mathrm{Toe}} - C^{[m]}_{\mathrm{Toe}}\| \leq 2\|\underline{c} - \underline{c}^{[m]}\|_{\ell^1_R(bv)}. \tag{4.165}$$

Using Corollary 4.4.10, that $\|\underline{c} - \underline{c}^{[m]}\|_{\ell^1_R(bv)} \to 0$ as $m \to \infty$, we have the convergence results.

By Theorem 4.3.8, $C^{[m]}_{\mathrm{com}}$ has finite rank. Therefore, since $C_{\mathrm{com}} = \lim_{m\to\infty} C^{[m]}_{\mathrm{com}}$ in the operator norm topology over $\ell^1_{R^{-1}}$, we have that $C_{\mathrm{com}}$ is compact in that topology. □

*Remark* 4.4.12. The transposed matrices $C^T_{\mathrm{Toe}}$ and $C^T_{\mathrm{com}}$ are bounded and compact (respectively) as operators from $\ell^1_R$ to $\ell^1_R$.

**Corollary 4.4.13.** *Let $C^\mu$ be as defined in Definition 4.2.12 for $C$ as in Theorem 4.4.11. Then $C^\mu$ can be decomposed into $C^\mu = C^\mu_{\mathrm{Toe}} + C^\mu_{\mathrm{com}}$ where $C^\mu_{\mathrm{Toe}}$ is upper triangular, Toeplitz and bounded as an operator from $\ell^1_{R^{-1}}$ to $\ell^1_{R^{-1}}$, and $C^\mu_{\mathrm{com}}$ is compact as an operator from $\ell^1_{R^{-1}}$ to $\ell^1_{R^{-1}}$, for all $R \in (0,1)$. Furthermore, if $J$ has Toeplitz-plus-finite-rank truncations $J^{[m]}$ with connection coefficient matrices $(C^\mu)^{[m]} = (C^\mu_{\mathrm{Toe}})^{[m]} + (C^\mu_{\mathrm{com}})^{[m]}$, then*

$$(C^\mu)^{[m]} \to C^\mu, \quad (C^\mu_{\mathrm{Toe}})^{[m]} \to C^\mu_{\mathrm{Toe}}, \quad (C^\mu_{\mathrm{com}})^{[m]} \to C^\mu_{\mathrm{com}} \text{ as } m \to \infty, \tag{4.166}$$

*in the operator norm topology over $\ell^1_{R^{-1}}$.*

*Proof.* This follows from Theorem 4.4.11 applied to $J^\mu$ as defined in Lemma 4.2.14. □

**Theorem 4.4.14.** *Let $J$ be a Jacobi operator such that $J = \Delta + K$ where $K$ is trace class. The Toeplitz symbols $c$ and $c_\mu$ of the Toeplitz parts of $C_{J\to\Delta}$ and $C^\mu_{J\to\Delta}$ are both analytic in the unit disc. Furthermore, if $J$ has Toeplitz-plus-finite-rank truncations $J^{[m]}$ with Toeplitz symbols $c^{[m]}$ and $c^{[m]}_\mu$, then $c^{[m]} \to c$ and $c^{[m]}_\mu \to c_\mu$ as $m \to \infty$ uniformly on compact subsets of $\mathbb{D}$.*

*Proof.* Let $R \in (0,1)$, and let $0 \le |z| \le R < 1$. Then by Lemma 4.4.6 we have

$$\left|\sum_{k=0}^{\infty} t_k z^k\right| \le \sum_{k=0}^{\infty} |t_k| R^k \le \sum_{k=0}^{\infty} \|c_{*,*+k}\|_{bv} R^k = \|\underline{c}\|_{\ell^1_R(bv)},$$

where $\underline{c}$ is as defined in equation (4.136). By Lemma 4.4.9 this quantity is finite. Since $R$ is arbitrary, the radius of convergence of the series is 1. The same is true for $c_\mu$ by Lemma 4.2.14.

Now we prove that the Toeplitz symbols corresponding to the Toeplitz-plus-finite-rank truncations converge.

$$\sup_{|z|\le R} |c(z) - c^{[m]}(z)| = \sup_{|z|\le R} \left|\sum_{k=0}^{\infty} (t_k - t^{[m]}_k) z^k\right| \le \sum_{k=0}^{\infty} |t_k - t^{[m]}_k| R^k$$

$$\leq \sum_{k=0}^{\infty} \|c_{*,*+k} - c^{[m]}_{*,*+k}\|_{bv} R^k = \|\underline{c} - \underline{c}^{[m]}\|_{\ell^1_R(bv)},$$

To go between the first and second lines, note that for each $k$, $c_{*,*+k} - c^{[m]}_{*,*+k}$ is a $bv$ sequence whose limit is $t_k - t^{[m]}_k$ and refer to Lemma 4.4.6. Now, $\|\underline{c} - \underline{c}^{[m]}\|_{\ell^1_R(bv)} \to 0$ as $m \to \infty$ by Corollary 4.4.10. The same is true for $\sup_{|z|\leq R} |c_\mu(z) - c^{[m]}_\mu(z)|$ by Lemma 4.2.14. □

**Theorem 4.4.15** (See [Kat95])**.** *Let $A$ and $B$ be bounded self-adjoint operators on $\ell^2$. Then*

$$\operatorname{dist}(\sigma(A), \sigma(B)) \leq \|A - B\|_2. \tag{4.167}$$

**Theorem 4.4.16.** *Let $J = \Delta + K$ be a Toeplitz-plus-trace-class Jacobi operator, and let $c$ and $c_\mu$ be the analytic functions as defined in Theorem 4.4.11 and Corollary 4.4.13. Then for $\lambda(z) = \frac{1}{2}(z + z^{-1})$ with $z \in \mathbb{D}$ such that $\lambda(z) \notin \sigma(J)$, the principal resolvent $G$ is given by the meromorphic function*

$$G(\lambda(z)) = -\frac{c_\mu(z)}{c(z)}. \tag{4.168}$$

*Therefore, all eigenvalues of $J$ are of the form $\lambda(z_k)$, where $z_k$ is a root of $c$ in $\mathbb{D}$.*

*Proof.* Let $z \in \mathbb{D}$ such that $\lambda(z) \notin \sigma(J)$, and let $J^{[m]}$ denote the Toeplitz-plus-finite-rank truncations of $J$ with principal resolvents $G^{[m]}$. Then $J^{[m]} \to J$ as $m \to \infty$, so by Theorem 4.4.15 there exists $M$ such that for all $m \geq M$, $\lambda(z) \notin \sigma(J^{[m]})$. For such $m$, both $G(\lambda(z))$ and $G^{[m]}(\lambda(z))$ are well defined, and using a well-known result on the difference of inverses (see for example, [TO16], [AH05]), we have

$$\begin{aligned} G^{[m]}(\lambda) - G(\lambda) &= \left\langle e_0, \left((J^{[m]} - \lambda)^{-1} - (J - \lambda)^{-1}\right) e_0 \right\rangle \\ &\leq \|(J^{[m]} - \lambda)^{-1} - (J - \lambda)^{-1}\|_2 \\ &\leq \frac{\|(J - \lambda)^{-1}\|_2^2 \|J - J^{[m]}\|_2}{1 - \|(J - \lambda)^{-1}\|_2 \|J - J^{[m]}\|_2} \\ &\to 0 \text{ as } m \to \infty. \end{aligned}$$

Theorem 4.4.14 shows that $\lim_{m\to\infty} c^{[m]}_\mu(z)/c^{[m]}(z) = c_\mu(z)/c(z)$. Therefore by Theorem 4.3.20 these limits are the same and we have equation (4.168). □

## 4.5 Computability aspects

In this section we discuss computability questions à la Ben-Artzi–Hansen–Nevanlinna–Seidel [BAHNS15a, BAHNS15b, Han11]. This involves an informal definition of the Solvability Complexity Index (SCI), a recent development that rigorously describes the extent to which various scientific computing problems can be solved. It is in contrast to classical computability theory à la Turing, in which problems are solvable *exactly* in finite time. In scientific computing we are often interested in problems which we can only *approximate* the solution in finite time, such that in an ideal situation this approximation can be made as accurate as desired. For example, the solution to a differential equation, the roots of a polynomial, or the spectrum of a linear operator.

Throughout this section we will consider only real number arithmetic, and the results do not necessarily apply to algorithms using floating point arithmetic.

The following is a modified definition of the Solvability Complexity Index (SCI). It is slightly stronger than the Ben-Artzi–Hansen–Nevanlinna–Seidel definition, which can be found in [BAHNS15a]; we do this to avoid a lengthy self-contained account of the theory where this simpler but stronger definition suffices for our purposes.

**Definition 4.5.1.** [(Strong) Solvability Complexity Index] A function $\Gamma$ which takes inputs to elements in a metric space $\mathcal{M}$ has Solvability Complexity Index at most $k$ if for each $n_1, \ldots, n_k \in \mathbb{N}$ there exists a Turing computable function $\Gamma_{n_1,\ldots,n_k}$ taking inputs to elements of $\mathcal{M}$, such that the function output for input $A$ is equal to the $\mathcal{M}$-limit

$$\Gamma(A) = \lim_{n_k \to \infty} \lim_{n_{k-1} \to \infty} \ldots \lim_{n_1 \to \infty} \Gamma_{n_1,\ldots,n_k}(A). \tag{4.169}$$

In other words, the output of $\Gamma$ can be computed using a sequence of $k$ limits.

*Remark* 4.5.2. The requirement we use here that the functions $\Gamma_{n_1,\ldots,n_k}$ are Turing computable is much stronger than what is used in the formal setting of [BAHNS15a], where the authors merely assume that these functions $\Gamma_{n_1,\ldots,n_k}$ depend only on, and are determined by, finitely many evaluable elements of the input datum. Informally, this suffices for our needs because we want to prove *positive* results about computability in this chapter; we prove our problem is computable in this stronger regime and that implies computability in the weaker regime of [BAHNS15a].

We require a metric space for the SCI. This

**Definition 4.5.3.** The Hausdorff metric for two compact subsets of the complex plane $A$ and $B$ is defined to be

$$d_H(A,B) = \max\left\{\sup_{a\in A}\operatorname{dist}(a,B), \sup_{b\in B}\operatorname{dist}(b,A)\right\}. \tag{4.170}$$

If a sequence of sets $A_1, A_2, A_3, \ldots$ converges to $A$ in the Hausdorff metric, we write $A_n \xrightarrow{H} A$ as $n \to \infty$.

**Theorem 4.5.4** ([BAHNS15a])**.** *The Solvability Complexity Index of the problem of computing the spectrum of a self-adjoint operator $A \in \mathcal{B}(\ell^2)$ is equal to 2 with respect to the Hausdorff metric on $\mathbb{R}$. For compact operators and banded self-adjoint operators the SCI reduces to 1.*

Theorem 4.5.4 implies that the SCI of computing the spectrum of bounded Jacobi operators in the Hausdorff metric is 1. In loose terms, the problem is solvable using only one limit of computable outputs. What more can we prove about the computability?

The results of Section 4.3 reduce the computation of the spectrum of a Toeplitz-plus-finite-rank Jacobi operator to finding the roots of a polynomial. From an uninformed position, one is lead to believe that polynomial rootfinding is a solved problem, with many standard approaches used every day. One common method is to use the QR algorithm to find the eigenvalues of the companion matrix for the polynomial. This can be done stably and efficiently in practice [AMVW15]. However, the QR algorithm is not necessarily convergent for non-normal matrices. Fortunately, the SCI of polynomial rootfinding with respect to the Hausdorff metric in for subsets of $\mathbb{C}$ is 1, but if you require the multiplicities of these roots then the SCI is not yet known [BAHNS15a].

A globally convergent polynomial rootfinding algorithm is given in [HSS01]. For any degree $d$ polynomial the authors describe a procedure guaranteed to compute fewer than $1.11d(\log d)^2$ points in the complex plane, such that for each root of the polynomial, a Newton iteration starting from at least one of these points will converge to this root.

Let $\varepsilon > 0$. If a polynomial $p$ of degree $d$ has $r$ roots, how do we know when to stop so that we have $r$ points in the complex plane each within $\varepsilon$ of a distinct root of $p$? This leads us to the concept of error control.

**Definition 4.5.5.** [Error control] A function $\Gamma$ which takes inputs to elements in a metric space $\mathcal{M}$ is computable with error control if it has solvability complexity index

1, and for each $\varepsilon$ we can compute $n$ to guarantee that

$$d_{\mathcal{M}}(\Gamma_n(A), \Gamma(A)) < \varepsilon. \tag{4.171}$$

In other words, the output of $\Gamma$ can be computed using a single limit, and an upper bound for the error committed by each $\Gamma_n$ is known.

Besides providing $\mathcal{O}(d(\log d)^2)$ initial data for the Newton iteration (to find the complex roots of a degree $d$ polynomial), the authors of [HSS01] discuss stopping criteria. In Section 9 of [HSS01], it is noted therein that for Newton iterates $z_1, z_2, \ldots$, if $|z_k - z_{k-1}| < \varepsilon/d$, then there exists a root $\xi$ of the polynomial in question such that $|z_k - \xi| < \varepsilon$. It is then noted, however, that if there are multiple roots then it is in general impossible to compute their multiplicities with complete certainty. This is because the Newton iterates can pass arbitrarily close to a root to which this iterate does not, in the end, converge. Another consequence of this possibility is that roots could be missed out altogether because all of the iterates can be found to be close to a strict subset of the roots.

To salvage the situation, we give the following lemma, which adds some assumptions to the polynomial in question.

**Lemma 4.5.6.** *Let $p$ be a polynomial and $\Omega \subset \mathbb{C}$ an open set such that, a priori, the degree $d$ is known and it is known that there are $r$ distinct roots of $p$ in $\Omega$ and no roots on the boundary of $\Omega$. Then the roots of $p$ in $\Omega$ is computable with error control in the Hausdorff metric (see Definition 4.5.3 and Definition 4.5.5).*

*Proof.* Use Newton's method with the $\mathcal{O}(d(\log d)^2)$ complex initial data given in [HSS01]. Using the stopping criteria in the discussion preceding this Lemma, the algorithm at each iteration produces $\mathcal{O}(d(\log d)^2)$ discs in the complex plane, within which all roots of $p$ must lie. Let $R_k \subset \Omega$ denote the union of the discs which lie entirely inside $\Omega$, with radius less than $\epsilon$ (the desired error).

Because the Newton iterations are guaranteed to converge from these initial data, we must have (for some sufficiently large $k$) that $R_k$ has $r$ connected components each with diameter less than $\epsilon$. Terminate when this verifiable condition has been fulfilled. □

**Theorem 4.5.7.** *Let $J = \Delta + F$ be a Toeplitz-plus-finite-rank Jacobi operator such that the rank of $F$ is known a priori. Then its point spectrum $\sigma_p(J)$ is computable with error control in the Hausdorff metric (see Definition 4.5.3 and Definition 4.5.5).*

*Remark* 4.5.8. Note that the full spectrum is simply $[-1,1] \cup \sigma_p(J)$.

*Proof.* Suppose $F$ is zero outside the $n \times n$ principal submatrix. The value of $n$ can be computed given that we know the rank of $F$. Compute the principal $2n \times 2n$ submatrix of the connection coefficients matrix $C_{J\to\Delta}$ using formulae (4.42)–(4.46). The entries in the final column of this $2n \times 2n$ matrix give the coefficients of the Toeplitz symbol $c$, which is a degree $2n-1$ polynomial.

Decide if $\pm 1$ are roots by evaluating $p(\pm 1)$. Divide by the linear factors if necessary to obtain a polynomial $\tilde{p}$ such that $\tilde{p}(\pm 1) \neq 0$. Noting that all roots inside $(-1,1)$ are simple (although the roots outside are not necessarily), use Sturm's Theorem to determine the number of roots of $\tilde{p}$ in $(-1,1)$, which we denote $r$ [RS02]. Since all roots in $\overline{\mathbb{D}}$ are real, there are $r$ roots of $\tilde{p}$ in the open unit disc $\mathbb{D}$ and none on the boundary.

By Lemma 4.5.6, the roots $z_1, \ldots, z_r$ of this polynomial $c$ which lie in $(-1,1)$ can be computed with error control. By Theorem 4.3.21, for the point spectrum of $J$ we actually require $\lambda_k = \frac{1}{2}(z_k + z_k^{-1})$ to be computed with error control. Note that since $|\lambda_k| \leq \|J\|_2$ for each $k$, we have that $|z_k| \geq (1+2\|J\|_2)^{-1}$. We should ensure that this holds for the computed roots $\hat{z}_k \in \mathbb{D}$ too. By the mean value theorem,

$$\begin{aligned}
|\lambda(z_k) - \lambda(\hat{z}_k)| &\leq \sup_{|z|\geq(1+2\|J\|_2)^{-1}} |\lambda'(z)||z_k - \hat{z}_k| \\
&= \frac{1}{2}\left((1+2\|J\|_2)^2 - 1\right)|z_k - \hat{z}_k| \\
&= 2\|J\|_2(1+\|J\|_2)|z_k - \hat{z}_k| \\
&\leq 2(1+\|F\|_2)(2+\|F\|_2)|z_k - \hat{z}_k|.
\end{aligned}$$

Therefore it suffices to compute $\hat{z}_k$ such that $|z_k - \hat{z}_k| \leq \frac{\varepsilon}{2}(1+\|F\|_2)^{-1}(2+\|F\|_2)^{-1}$, where $\varepsilon$ is the desired error in the eigenvalues. □

**Theorem 4.5.9.** *Let $J = \Delta + K$ be a Toeplitz-plus-compact Jacobi operator. If for all $\epsilon > 0$ an integer $m$ can be computed such that*

$$\sup_{k\geq m} |\alpha_k| + \sup_{k\geq m} \left|\beta_k - \frac{1}{2}\right| < \epsilon, \tag{4.172}$$

*then the spectrum can be computed with error control in the Hausdorff metric.*

*Proof.* Let $\epsilon > 0$. By the oracle assumed in the statement of the Theorem, compute $m$ such that

$$\sup_{k\geq m} |\alpha_k| + \sup_{k\geq m} \left|\beta_k - \frac{1}{2}\right| < \frac{\epsilon}{6}. \tag{4.173}$$

Now compute the point spectrum of Toeplitz-plus-finite-rank truncation $J^{[m]}$ such that $d_H(\Sigma, \sigma(J^{[m]})) < \varepsilon/2$, where $\Sigma$ denotes the computed set. Then, using Theorem 4.4.15, we have

$$\begin{aligned} d_H(\Sigma, \sigma(J)) &\leq d_H(\Sigma, \sigma(J^{[m]})) + d_H(\sigma(J^{[m]}), \sigma(J)) && (4.174)\\ &\leq \frac{\varepsilon}{2} + \|J^{[m]} - J\|_2 && (4.175)\\ &\leq \frac{\varepsilon}{2} + 3\frac{\varepsilon}{6} && (4.176)\\ &= \varepsilon. && (4.177) \end{aligned}$$

Here we used the fact that for a self-adjoint tridiagonal operator $A$, $\|A\|_2 \leq 3(\sup_{k\geq 0} |a_{k,k}| + \sup_{k\geq 0} |a_{k,k+1}|)$. This completes the proof. □

## 4.6 Numerical results and the *SpectralMeasures* package

In this section we demonstrate some of the features of the Julia package that present authors have written to implement the ideas in the paper. The package is called *SpectralMeasures* and is part of the *JuliaApproximation* project, whose main package is *ApproxFun*. `ApproxFun` is an extensive piece of software influenced by the Chebfun package in Matlab, which can represent functions and operators [OT14],[Olvb], [DHT14]. All of the packages are open source and available to download at http://www.github.com/JuliaApproximation. The code is subject to frequent changes and updates.

Given a Jacobi operator $J$ which is a finite-rank perturbation of the free Jacobi operator $\Delta$, there are four things from this paper which we would like to compute. Let $n$ be an integer such that $\alpha_k = 0$, $\beta_{k-1} = \frac{1}{2}$ for all $k \geq n$.

(i) The connection coefficients matrix $C_{J\to\Delta}$: This is computed using the recurrences in equation (4.42)–(4.42). By Theorem 4.3.8, we only need to compute $n(n+1)$ entries of $C$ to have complete knowledge of all entries. In *SpectralMeasures*, there is a class of operator called PertToeplitz, which allows such an operator to be stored and manipulated as if it were the full infinite-dimensional operator.

(ii) The spectral measure $\mu(s)$: By Theorem 4.3.14, this measure has the form

$$\mathrm{d}\mu(s) = \frac{1}{p_C(s)}\frac{2}{\pi}\sqrt{1-s^2}\mathrm{d}s + \sum_{k=1}^{r} w_k \delta_{\lambda_k}(s),$$

where $p_C$ is the polynomial given by the computable formula $p_C(s) = \sum_{k=0}^{2n-1}\langle e_k, CC^T e_0\rangle U_k(s)$ and $r \leq n$. By Theorem 4.3.21, the numbers $\lambda_k$ are found by finding the distinct real roots $z_k$ of $c$ (the Toeplitz symbol of the Toeplitz part of $C$, which here is a polynomial of degree $2n-1$) in the interval $(-1,1)$. Also by Theorem 4.3.21, the weights $w_k$ can be computed using the formula

$$w_k = \frac{1}{2} z_k^{-1}(z_k - z_k^{-1})\frac{c_\mu(z_k)}{c'(z_k)}.$$

(iii) The principal resolvent $G(\lambda)$: For any $\lambda \in \mathbb{C}\setminus\sigma(J)$, by Theorem 4.3.12, this function can be defined by the formula

$$G(\lambda) = \frac{G_\Delta(\lambda) - p_C^\mu(\lambda)}{p_C(\lambda)},$$

where $p_C$ is as above and $p_C^\mu(\lambda) = \sum_{k=0}^{2n-1}\langle e_k, C^\mu C^T e_0\rangle U_k(\lambda)$.

(iv) The mapped principal resolvent $G(\lambda(z))$: which is the principal resolvent mapped to $z$ in the unit disc by the Joukowski map $\lambda : z \to \frac{1}{2}(z+z^{-1})$. This is computed using the simple formula from Theorem 4.3.20,

$$G(\lambda(z)) = -\frac{c_\mu(z)}{c(z)},$$

where $c$ and $c_\mu$ are the Toeplitz symbols of the Toeplitz parts of $C$ and $C^\mu$ respectively (these are polynomials of degree $2n-1$ and $2n-2$ respectively).

Consider a concrete example of a Toeplitz-plus-finite-rank Jacobi operator,

$$J = \begin{pmatrix} \frac{3}{4} & 1 & & & & \\ 1 & -\frac{1}{4} & \frac{3}{4} & & & \\ & \frac{3}{4} & \frac{1}{2} & \frac{1}{2} & & \\ & & \frac{1}{2} & 0 & \frac{1}{2} & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{4.178}$$

The connection coefficients operator $C = C_{J\to\Delta}$ is

$$C = \begin{pmatrix} 1 & -\frac{3}{4} & -\frac{5}{4} & \frac{49}{24} & -\frac{1}{12} & -\frac{1}{3} & \\ & \frac{1}{2} & -\frac{1}{3} & -\frac{4}{3} & \frac{41}{24} & -\frac{1}{12} & -\frac{1}{3} & \\ & & \frac{1}{3} & -\frac{2}{3} & -\frac{4}{3} & \frac{41}{24} & -\frac{1}{12} & \ddots \\ & & & \frac{1}{3} & -\frac{2}{3} & -\frac{4}{3} & \frac{41}{24} & \ddots \\ & & & & \frac{1}{3} & -\frac{2}{3} & -\frac{4}{3} & \ddots \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{4.179}$$

As was noted above, we only need to explicitly compute $3 \cdot 4 = 12$ entries and the rest are defined to be equal to the entry above and left one space.

In Figure 4.3 we present plots of the spectral measure, the principal resolvent and the mapped principal resolvent. This format of figure is repeated for other Jacobi operators below.

The plot on left in Figure 4.3 is the spectral measure. There is a continuous part supported in the interval $[-1, 1]$ and Dirac deltas a represented by vertical lines whose heights are precisely their weights. As noted in Theorem 4.5.7, it is possible to compute the eigenvalues of $J$ with guaranteed error control. Computations with guaranteed error control are made quite straightforward and flexible using the ValidatedNumerics package [BS], in which computations are conducted using interval arithmetic, and the desired solution is rigorously guaranteed to lie within the interval the algorithm gives the user [Tuc11]. Using this open source Julia package, we can compute the two eigenvalues for this operator to be $-1.1734766767874558$ and $1.5795946563898884$ with a guaranteed maximum error of $8.9 \times 10^{-16}$. This can be replicated using the command `validated_spectrum([.75;-.25;.5],[1;.75])` in *SpectralMeasures*.

The coloured plots in the middle and the right of Figure 4.3 are *Wegert plots* (sometimes called phase portraits) [Weg12],[Kro]. For a function $f : \mathbb{C} \to \mathbb{C}$, a Wegert plot assigns a colour to every point $z \in \mathbb{C}$ by the argument of $f(z)$. Specifically, if $f(z) = re^{i\theta}$, then $\theta = 0$ corresponds to the colour red, then cycles upwards through yellow, green, blue, purple as $\theta$ increases until at $\theta = 2\pi$ it returns to red. This makes zeros and poles very easy to see, because around them the argument cycles through all the colours the same number of times as the degree of the root or pole. In these particular Wegert plots, we also plot lines of constant modulus as shadowed steps.

The middle plot in Figure 4.3 is the principal resolvent $G(\lambda)$, which always has a branch cut along the interval $[-1, 1]$ and roots and poles along the real line. The poles correspond to Dirac delta measures in the spectral measure.

The third plot is the principal resolvent of $J$ mapped to the unit disc by the Joukowski map. Poles and roots of this resolvent in the unit disc correspond to those of the middle plot outside $[-1, 1]$.

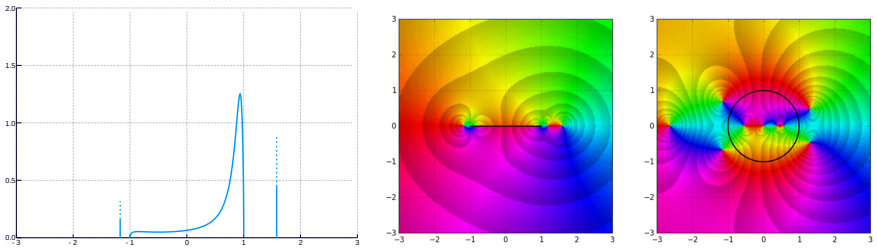


Fig. 4.3 The left plot is the spectral measure $\mu(s)$ of the Jacobi operator in equation (4.178). The middle plot is a Wegert plot (explained in the text above) depicting the principal resolvent of the same Jacobi operator, and the right plot is the principal resolvent under the Joukowski mapping. The two Dirac deltas in the spectral measure correspond to two poles along the real line for the middle plot and two poles inside the unit disc for the right plot.

In Figure 4.4 we have plotted the spectral measure and principal resolvent of the Basic Perturbation 1 (see Examples 4.3.2, 4.3.16, 4.3.22) in which the top-left entry of the operator has been set to $\alpha/2$ for values $\alpha = 0, 0.15, 0.35, 0.5, 0.75, 1$. For the first four cases, the perturbation from the free Jacobi operator is small, and so the spectrum is purely continuous, which corresponds to no poles in the principal resolvent, and in the mapped resolvent there are only poles *outside* the unit disc. For the cases $\alpha = 0.75, 1$, the Jacobi operator has a single isolated point of discrete spectrum. This is manifested as a Dirac delta in the spectral measure and a single pole in the principal resolvent.

In Figure 4.5 we have plotted the spectral measure and principal resolvent of the Basic Perturbation 2 (see Examples 4.3.3, 4.3.17, 4.3.23) in which the $(0, 1)$ and $(1, 0)$ entries have been set to $\beta/2$ for values $\beta = 0.5, 0.707, 0.85, 1.0, 1.2, 1.5$. The effect is similar to that observed in Figure 4.4. For small perturbations the spectrum remains purely continuous, but for larger perturbations here two discrete eigenvalues emerge corresponding to Dirac deltas in the spectral measure and poles in the principal resolvent.

In Figure 4.6 we have plotted a sequence of approximations to the Jacobi operator for the Legendre polynomials, which has entries $\alpha_k = 0$ for $k = 0, 1, 2, \ldots$ and

$$\beta_{k-1} = \frac{1}{\sqrt{4k^2 - 1}}, \text{ for } k = 1, 2, 3, \ldots.$$

This is a Toeplitz-plus-trace-class Jacobi operator because $\beta_k = \frac{1}{2} + \mathcal{O}(k^{-2})$, and by taking Toeplitz-plus-finite-rank approximations $J^{[n]}$ as in equation (4.132), we can compute approximations to the spectral measure and principal resolvent. Figure 4.6 depicts the spectral measure and the principal resolvent for the Toeplitz-plus-finite-rank Jacobi operators $J^{[n]}$ for the values $n = 1, 2, 3, 10, 30, 100$. For the spectral measures, we see that there is no discrete part for any $n$, and as $n$ increases, the spectral measure converges to the scaled Lebesgue measure $\frac{1}{2}\mathrm{d}s$ restricted to $[-1, 1]$. The convergence is at least weak by Proposition 4.4.2, but it would be interesting (as mentioned in the conclusions) to determine if there is a stronger form of convergence at play due to the perturbation of $\Delta$ lying in the space of trace class operators. There is a Gibbs effect occurring at the boundaries, which suggests that this convergence, if it occurs pointwise, can only do so everywhere up to the boundary of $[-1, 1]$. For the principal resolvents, the middle plots do not show much interesting, as the difference between the functions in the complex plane is not major. However, in right plots, there are hidden pole-root pairs in the resolvent lying outside the unit disc which coalesce around the unit disc and form a barrier. The meaning of this barrier is unknown to the authors.

Figures 4.7 and 4.8 demonstrate similar features to Figure 4.6, except that the polynomials sequences they correspond to are the ultraspherical polynomials with parameter $\gamma = 0.6$ (so that the spectral measure is proportional to $(1 - s^2)^{1.1}$) and the Jacobi polynomials with parameters $(\alpha, \beta) = (0.4, 1.9)$ (so that the spectral measure is proportional to $(1 - s)^{0.4}(1 + s)^{1.9}$). Similar barriers of pole-root pairs outside the unit disc occur for these examples as well.

Figure 4.9 presents a Toeplitz-plus-trace-class Jacobi operator with pseudo-randomly generated entries. With a random vector $\mathbf{r}$ containing entries uniformly distributed in the interval $[0, 1)$, the following entries were used

$$\alpha_k = 3\frac{2r_k - 1}{(k+1)^2}, \qquad \beta_k = \frac{1}{2}.$$

Then Toeplitz-plus-finite-rank truncations $J^{[n]}$ (see equation (4.132)) of this operator were taken for values $n = 1, 2, 3, 10, 50, 100$. Since the off-diagonal elements are constant,

this is a scaled and shifted version of a discrete Schrödinger operator with a random, decaying potential.

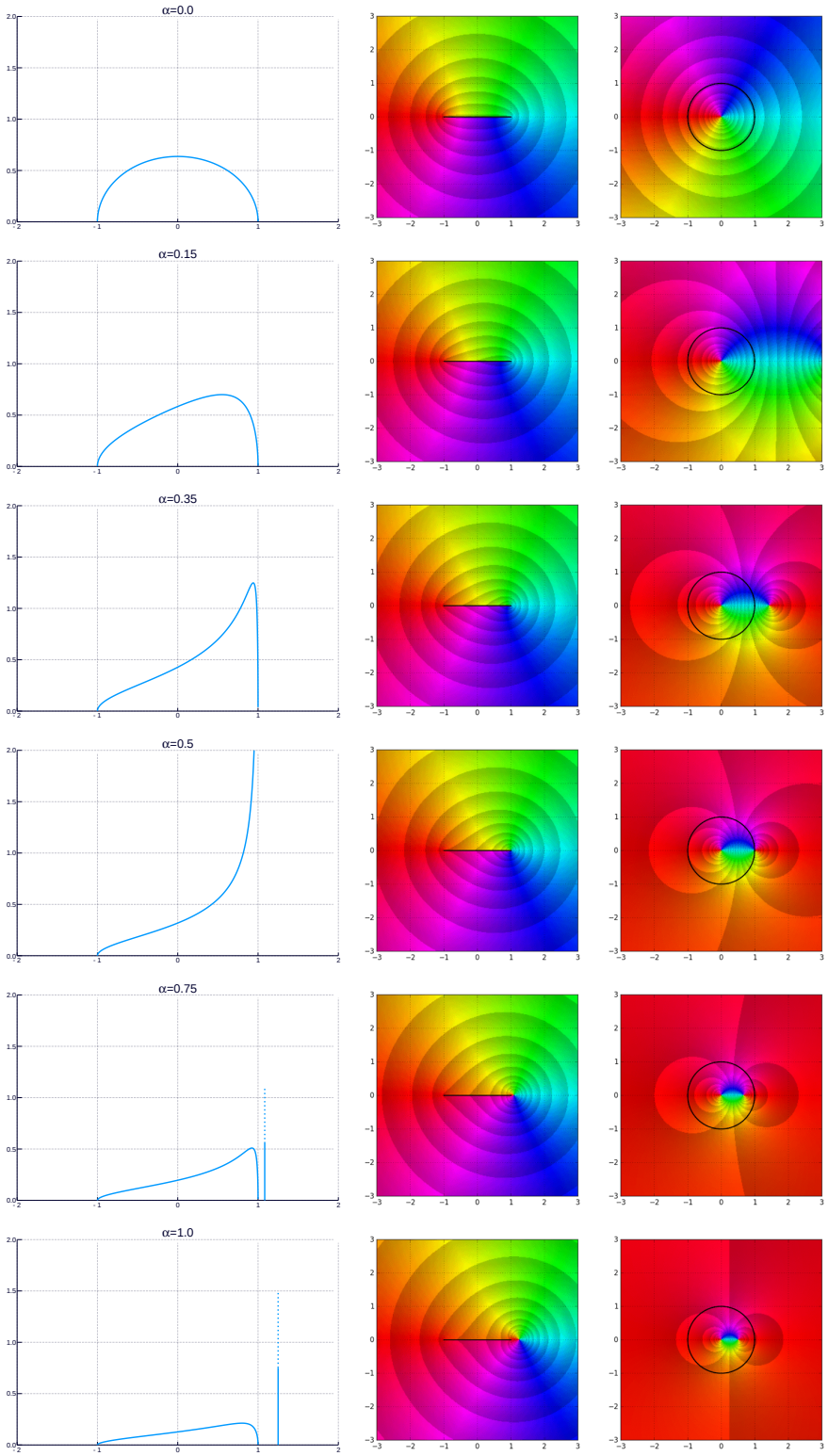



Fig. 4.4 The left hand, centre and right hand figures show the spectral measures $\mu(s)$, principal resolvents $G(\lambda)$ and disc resolvents $G(\lambda(z))$ (analyticially continued outside the disc) respectively for $J_\alpha$, the Basic perturbation 1 example, with $\alpha = 0, 0.15, 0.35, 0.5, 0.75, 1$. We see that a Dirac mass in the measure corresponds to a pole of the disc resolvent *inside* the unit disc, which corresponds to a pole in the principal resolvent outside the interval $[-1, 1]$.

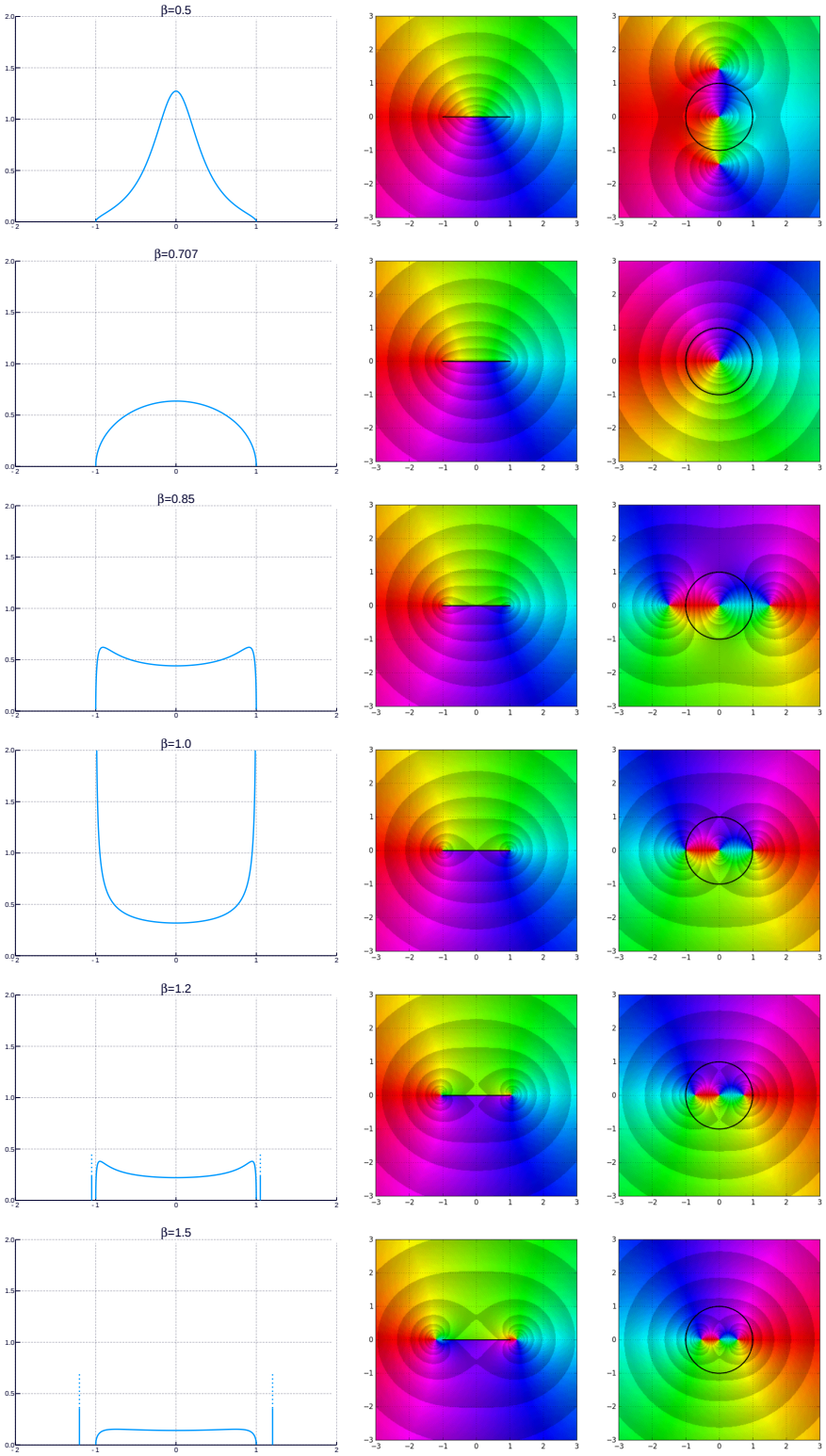



Fig. 4.5 The left hand, centre and right hand figures show the spectral measures $\mu(s)$, principal resolvents $G(\lambda)$ and disc resolvents $G(\lambda(z))$ (analyticially continued outside the disc) respectively for $J_\beta$, the Basic perturbation 2 example, with $\beta = 0.5, 0.707, 0.85, 1, 1.2, 1.5$. Again, we see that a Dirac mass in the measure corresponds to a pole of the disc resolvent *inside* the unit disc, which corresponds to a pole in the principal resolvent outside the interval $[-1, 1]$.

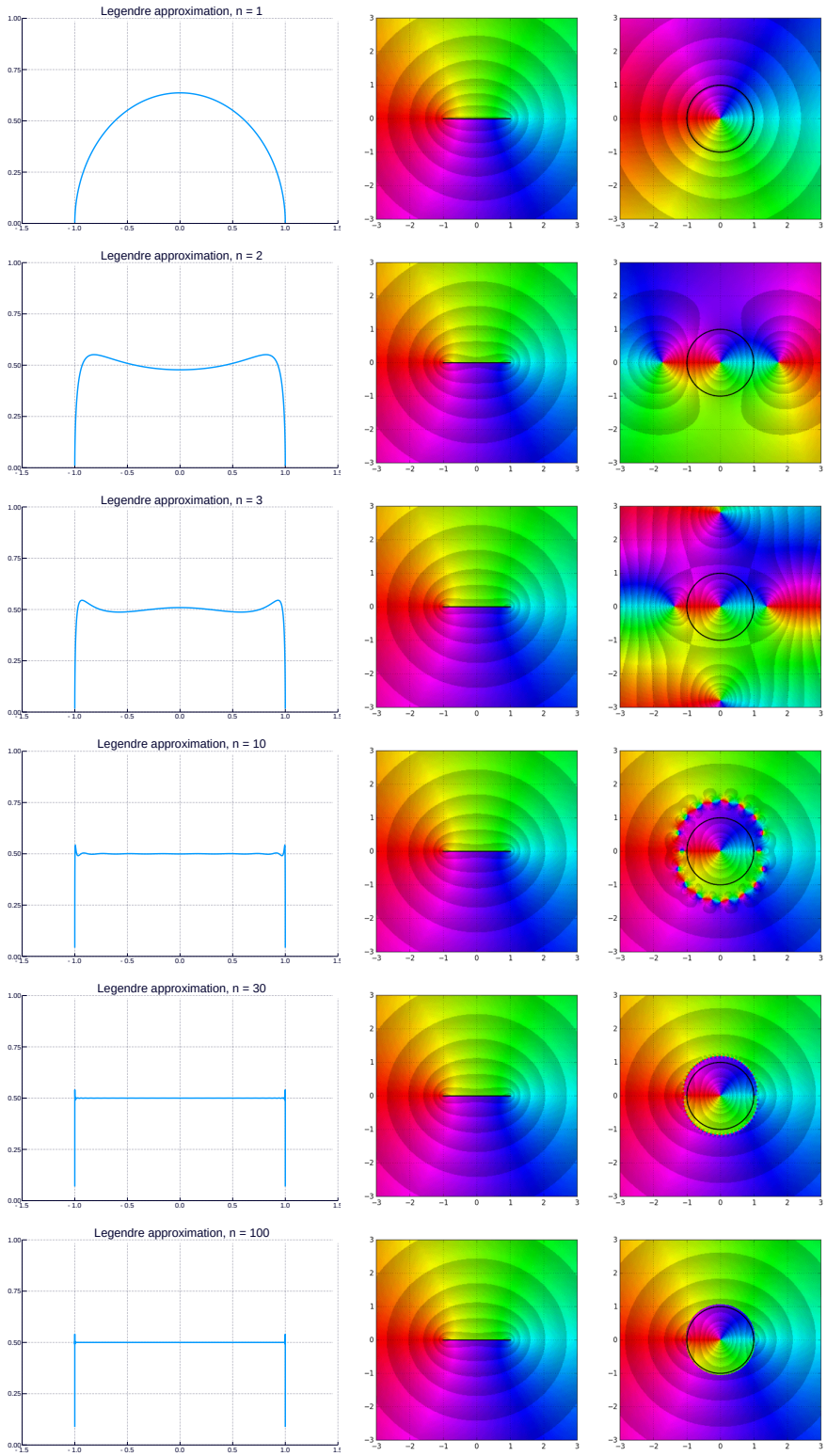



Fig. 4.6 These plots are of approximations to the spectral measure and principal resolvents of the Legendre polynomials, which has a Toeplitz-plus-trace-class Jacobi operator. The Jacobi operator can be found in Subsection 4.4.1. As the parameter $n$ of the approximation increases, a barrier around the unit circle forms. Also notice that a Gibbs phenomenon forms at the end points, showing that there are limitations to how good these approximations can be to the final measure.

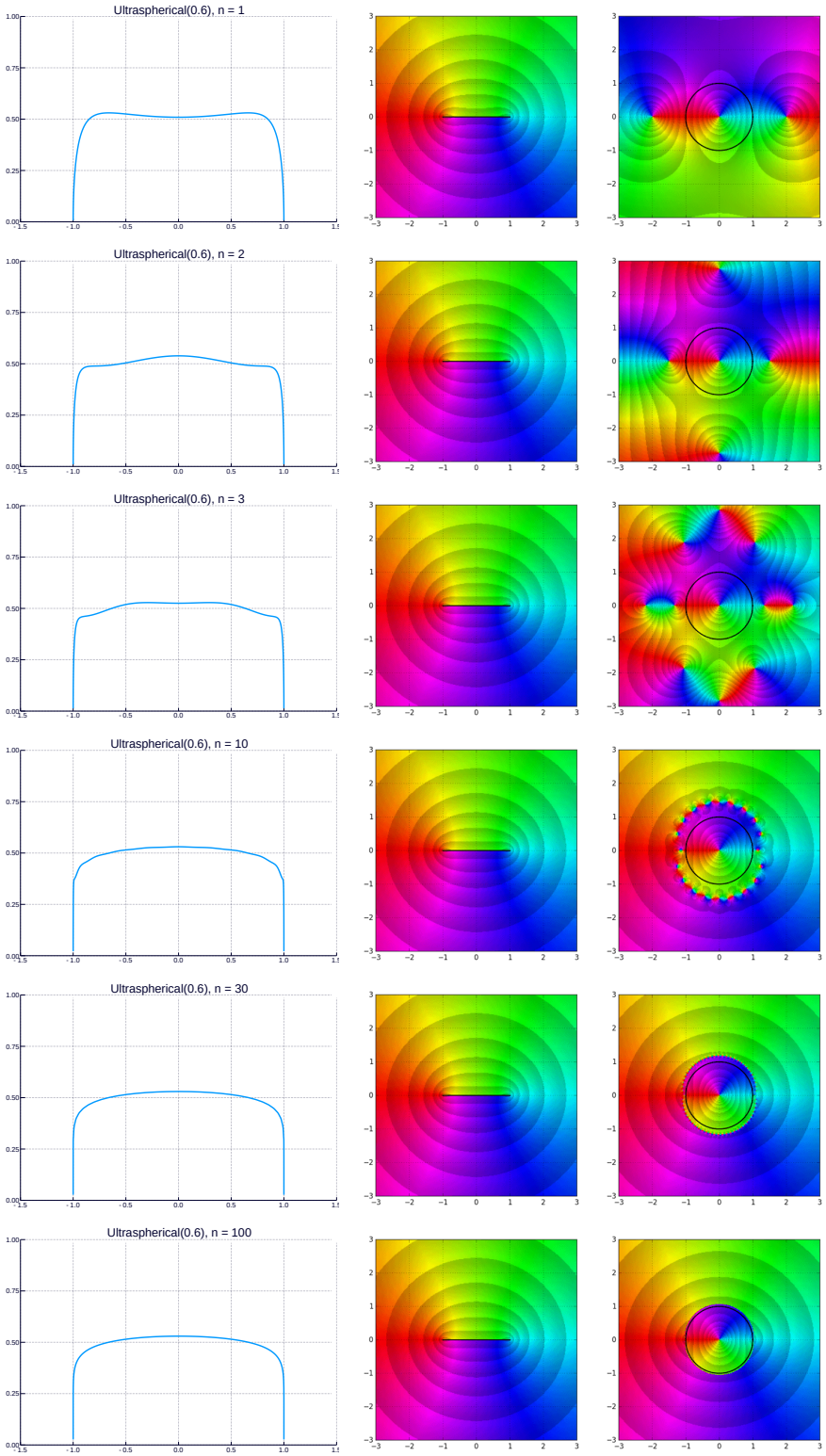



Fig. 4.7 These plots are of approximations to the spectral measure and principal resolvents of the Ultraspherical polynomials with parameter $\gamma = 0.6$, which has a Toeplitz-plus-trace-class Jacobi operator. The Jacobi operator can be found in Subsection 4.4.1. As the parameter $n$ of the approximation increases, a barrier around the unit circle forms.

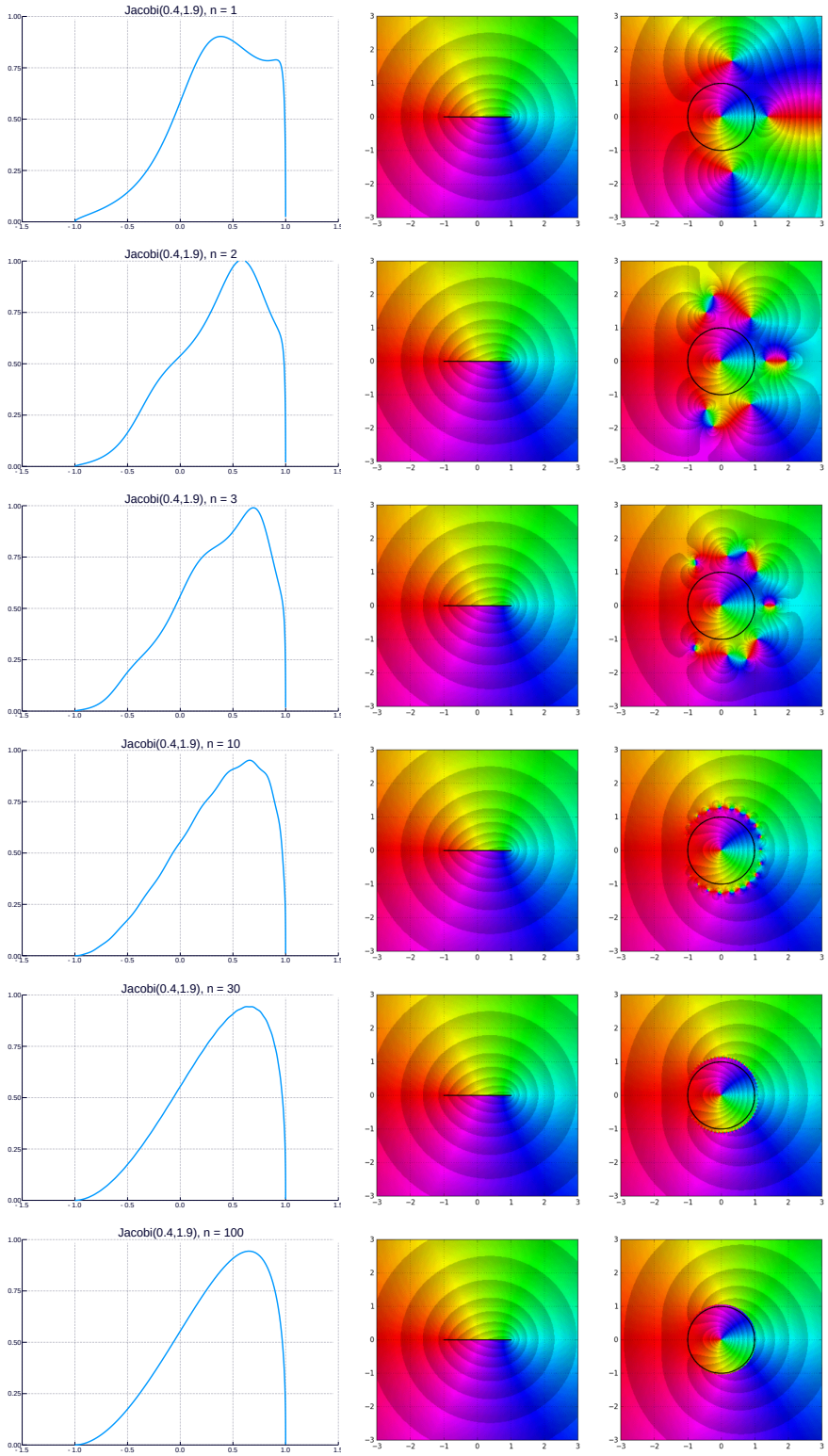



Fig. 4.8 These plots are of approximations to the spectral measure and principal resolvents of the Jacobi polynomials with parameter $\alpha, \beta = 0.4, 1.9$, which has a Toeplitz-plus-trace-class Jacobi operator. The Jacobi operator can be found in Subsection 4.4.1. As the parameter $n$ of the approximation increases, a barrier around the unit circle forms.

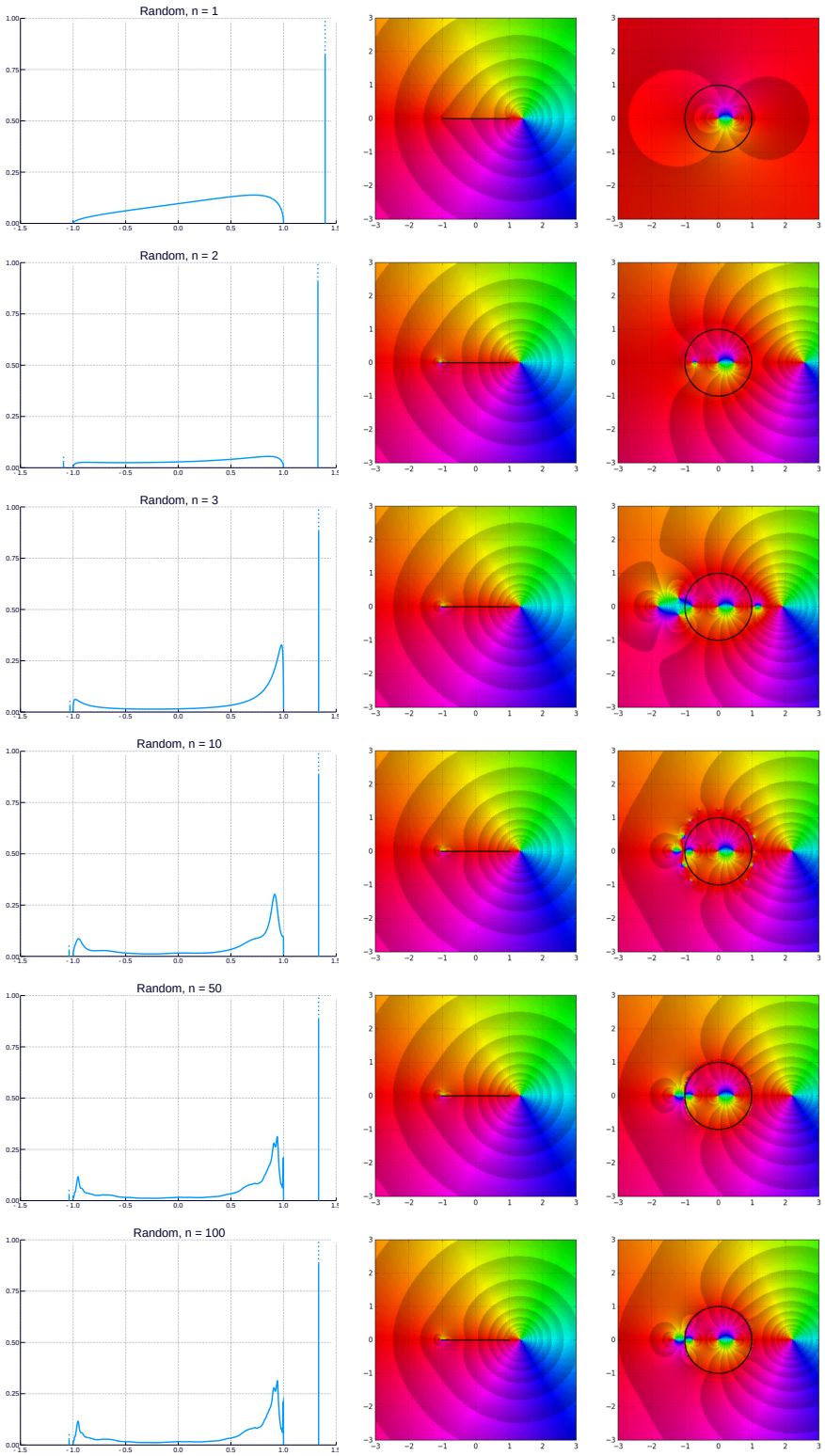



Fig. 4.9 These plots are of approximations to the spectral measure and principal resolvents of a trace-class pseudo-random diagonal perturbation of the free Jacobi operator.

开门见山

(kāi mén jiànshān)

Lit. Open the door and see the mountain

Fig. To get straight to the point

# Chapter 5

# The infinite dimensional QL algorithm

In this Chapter we discuss the infinite dimensional QL algorithm. The QL algorithm is the same as the QR algorithm (See Section 2.2), except that instead of computing QR factorisations we compute QL factorisations where $L$ is lower triangular.

In finite dimensions the QR and QL algorithms are equivalent in the following sense. Take $E$ to be the exchange matrix

$$E = \begin{pmatrix} & & & & 1 \\ & & & 1 & \\ & & \iddots & & \\ & 1 & & & \\ 1 & & & & \end{pmatrix}.$$

If the sequence of matrices $X_0, X_1, X_2, \ldots$ is produced by the QR algorithm, then it may also be produced by setting $Y_0 = EX_0E$, performing the QL algorithm on $Y_0$ to produce $Y_0, Y_1, Y_2, \ldots$ and setting $X_k = EY_kE$. Since the operation $X \mapsto EXE$ simply rotates the entries of $X$ by $\pi$ radians, the QL algorithm is equivalent to the QR algorithm operating on a rotated matrix. Hence in finite dimensions there is no real difference between the QR and QL algorithms in terms of convergence.

When considering the QR algorithm applied to Hermitian matrices, one usually restricts to tridiagonal matrices, for two reasons. First, tridiagonalisation can be performed in $\mathcal{O}(n^2)$ operations and generally speeds up the whole process of computing eigenvalues [TBI97]. Second, for the QR algorithm on tridiagonal matrices, Wilkinson shifts can be applied to the algorithm give global convergence of the bottom–right entry to an eigenvalue of the input matrix [Wat07], [Par80]. The QR algorithm can

be generalised to the infinite dimensional case of bounded operators on $\ell^2$ [Han08], [Han09], and so too can the related Toda flow [DLT85]. However, there is an issue with the use of shifts to accelerate convergence: there is no bottom-right entry! In finite dimensions, since the QL algorithm is equivalent to the QR algorithm after rotating the entries $\pi$ radians, we can force the *top-left* entry to converge rapidly to an eigenvalue using Wilkinson shifts. However, this logic does not follow for the infinite dimensional case as there is no infinite dimensional analogue of the matrix $E$.

Olver and Townsend proposed the following idea a footnote of [OT14]. In principal, if one could perform the QL algorithm to an infinite dimensional matrix, it could be possible to utilise shifts to yield rapid convergence of the top–left entry to an eigenvalue (if the matrix has any point spectrum). However, there were no known methods to compute the QL factorisation of a (non-compact) infinite dimensional matrix. What is the issue? To compute a QR factorisation we apply elementary orthogonal transformations to $A$ to create zeros in the first column, then the second column etc. until we have an upper triangular matrix. In contrast, to compute a QL factorisation we apply elementary orthogonal transformations to $A$ to create zeros in the *final* column, then the penultimate column etc. There is no obvious way to do this for an infinite dimensional matrix as there is no final column.

One of the main contributions of this chapter is Theorem 5.1.17, which goes partway to solving the problem posed by Olver and Townsend. Here is the gist. For some bounded operators $A$ on $\ell^2$ we can find an analytical solution to $A = QL$. For example, if $A = \Delta - \frac{5}{4}I$ where $\Delta$ is the free Jacobi operator from Chapter 4, then the QL factorisation of $A$ is

$$\Delta - \frac{5}{4}I = \begin{pmatrix} -\frac{\sqrt{3}}{2} & \frac{1}{2} & & & & \\ -\frac{\sqrt{3}}{4} & -\frac{3}{4} & \frac{1}{2} & & & \\ -\frac{\sqrt{3}}{8} & -\frac{3}{8} & -\frac{3}{4} & \frac{1}{2} & & \\ -\frac{\sqrt{3}}{16} & -\frac{3}{16} & -\frac{3}{8} & -\frac{3}{4} & \frac{1}{2} & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \frac{\sqrt{3}}{2} & & & & \\ -1 & 1 & & & \\ \frac{1}{4} & -1 & 1 & & \\ & -\frac{1}{4} & -1 & 1 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \quad (5.1)$$

A proof of this new result (for a general Toeplitz-plus-finite-rank Jacobi operator) is given in Theorem 5.2.6. Such QL factorisations can be used to compute the QL factorisation of banded matrices which contain said matrix as the bottom-right submatrix. Specifically, suppose that $A$ is a banded, bounded operator on $\ell^2$ with

bandwidth $b$ and block form

$$A = \begin{pmatrix} A_n & B \\ C & A_\infty \end{pmatrix}, \tag{5.2}$$

where $A_n \in \mathbb{R}^{n\times n}$ and $A_\infty$ has an *a priori* known QL factorisation $A_\infty = Q_\infty L_\infty$. Then the way to compute the QL factorisation of $A$ is seen by noting that

$$\begin{pmatrix} I_n & \\ & Q_\infty^T \end{pmatrix} \begin{pmatrix} A_n & B \\ C & A_\infty \end{pmatrix} = \begin{pmatrix} A_n & B \\ Q_\infty^T & L_\infty \end{pmatrix}. \tag{5.3}$$

The right hand side has finitely many nonzero entries above the diagonal (bandedness implies $B$ has finitely many nonzero entries). Therefore the standard approach of introducing zeros from the rightmost column applies and can be used to complete the QL factorisation in finitely many operations. Full details are given in Theorem 5.1.17.

There are three immediate questions to address.

First, on a finite computer how is the storage of these infinite-dimensional objects managed? If the matrices are sufficiently structured, and QL iterations respect that structure, then the operators can be stored with finitely many scalar values. Consider the example above in equation (5.1). The $Q$ and the $L$ in the QL factorisation are Toeplitz-plus-finite-rank, and in fact the product $LQ$ is also Toeplitz-plus-finite-rank. The orthogonal operator is not banded, so it would appear that we can only approximately store it in memory, but actually, an orthogonal lower Hessenberg operator is uniquely determined by the first upper-diagonal and a $\pm 1$ scaling factor for the first column, hence the $Q$ in equation (5.1) can be represented by simply the numbers $\frac{1}{2}$ and $+1$ to represent the recurring number on the upper diagonal and the scaling of the first column. This is the so-called Schur parametrisation of Hessenberg matrices [Gra86] (see Lemma 5.2.4).

It is future research to ascertain which structured operators will also have QL factorisations and QL iterates that have structure which can be stored in finite memory. In Chapter 6 we suggest plausible examples of operators which could come under this framework.

Second, is it reasonable to expect that an analytical QL factorisation for a given operator from a scientific application to be sufficiently structured to be computable? It does seem unreasonable at first, but recent research on spectral methods produced operators with banded-plus-finite-rank matrix structure, with simple asymptotics of

the entries, and led to the development of a practical framework for solving infinite dimensional linear systems on a computer [OT13], [OT14], [SO17] [Olvb]. The key to producing these highly structured matrices is an appropriate choice of basis, so it perhaps it is not unreasonable to suggest that many operators in applications can be represented by highly structured matrices with the right choice of bases.

Third, is this not going to a lot of trouble when there are already established methods such as the finite section method (see Subsection 1.1.6) [Bof10], [Tre00], [TE05]?

One of the most surprising results to come out of this work is that the existence of a QL factorisation is not always guaranteed, unlike the case for the QR factorisation [Han08], [Han09]. We prove that a Jacobi operator has a QL factorisation if and only if the essential spectrum does not contain zero (see Theorem 5.2.2). Notably, the free Jacobi operator $\Delta$ does not have a QL factorisation. In Theorem 5.1.3, Theorem 5.1.14 and Corollary 5.1.15 we prove generalisations of this for banded, selfadjoint operators, but their exact statements have some technical points we will explain there and not here.

In Section 5.1 we prove existence and nonexistence results for QL factorisations of bounded selfadjoint case and briefly indicate if there is an easy generalisation of a result to the non-selfadjoint case. In Section 5.2 we restrict these results to Jacobi operators and find that the statement of the results is simpler. Then in Section 5.2.2 we make practical considerations for running the QL algorithm for Jacobi operators on a computer, and derive a method to compute the QL factorisation of a Toeplitz-plus-finite-rank Jacobi operators, using only a finite amount of memory.

In Section 5.3 we consider the infinite dimensional QL algorithm which utilises these infinite dimensional QL factorisations. We prove that for a bounded Jacobi operator $J$ such that there is an eigenvalue $\lambda_0$ satisfying

$$0 < |\lambda_0| < \eta := \min_{\lambda \in \sigma(J) \setminus \lambda_0} |\lambda|, \tag{5.4}$$

the unshifted QL algorithm converges in the sense that the $(0,1)$ entry is $\mathcal{O}\left(\left|\frac{\lambda_0}{\eta}\right|^k\right)$. This implies (as is done in the finite dimensional case [Par80]) that if a shift is chosen sufficiently close to an isolated eigenvalue, then there will be rapid convergence of the top-left entry to that eigenvalue.

At the end of the chapter we demonstrate the fruits of Chapter 4 and the present chapter to compute functions of Jacobi operators. We use this to compute the solution

to a discrete Schrödinger equation with double potential wells (demonstrating discrete quantum tunnelling), and some diffusion equations, including fractional order ones. A brief comparison to the traditional finite section method approach is made (see Subsection 1.1.6).

## 5.1 Basic properties

First we make perfectly clear what we mean by a QL factorisation. We consider the real separable Hilbert space $\ell^2$.

**Definition 5.1.1.** A QL factorisation of an operator $A \in \mathcal{B}(\ell^2)$ is a factorisation

$$A = QL \tag{5.5}$$

where $Q$ is an orthogonal operator ($Q^TQ = QQ^T = I$) and $L$ is lower triangular.

### 5.1.1 Existence

The existence of a QR factorisation for a bounded invertible operator $A \in \mathcal{B}(\ell^2)$ was proven by Deift, Li and Tomei [DLT85]. Hansen later removed the invertibility restriction [Han08]. When a bounded operator is invertible, the existence of a QL factorisation follows directly from the existence of a QR factorisation proven by Hansen.

**Lemma 5.1.2.** *If a linear operator $A \in \mathcal{B}(\ell^2)$ is invertible, then it has a unique QL factorisation such that $L$ has positive diagonal elements.*

*Proof.* By the Open Mapping Theorem, if $A$ is bounded then its inverse is also bounded. Since transposes of bounded operators are also bounded, we have $A^{-T} \in \mathcal{B}(\ell^2)$. By [Han08, Thm 31] (see also [Han09, Thm 4.2]), there exists a unique QR factorisation $A^{-T} = QR$ such that the diagonal elements of $R$ are positive. Then we have

$$A = (A^{-T})^{-T} = (QR)^{-T} = QR^{-T}. \tag{5.6}$$

Since $R^{-T}$ is lower triangular with positive diagonal elements, this constitutes the desired QL factorisation of $A$. Working backwards from a QL factorisation to the unique QR factorisation shows uniqueness. □

What happens when $A \in \mathcal{B}(\ell^2)$ is singular? In finite dimensions a QL factorisation *always* exists. We have the following positive result. See Appendix D.2 for an overview of the basic definitions and theorems regarding Fredholm operators.

**Theorem 5.1.3.** *Let $A \in \mathcal{B}(\ell^2)$ be selfadjoint. Then the following are equivalent.*

*(i) $A$ is Fredholm with $\dim(\mathrm{Ker}(A)) = d$, and $\mathrm{Ker}(A) \cap \mathrm{Span}\{e_d, e_{d+1}, \ldots\} = \{0\}$*

*(ii) $A$ has a QL factorisation such that the first $d$ rows of $L$ are zero and $L_{d:\infty \times d:\infty}$ is an invertible operator on $\ell^2$. It is unique up to the space spanned by the first $d$ columns of $Q$ and the signs of the columns of $L$.*

*Proof.* For $(i) \implies (ii)$: First suppose that $d = 0$. Then by Lemma 5.1.2, $A$ has a unique QL factorisation.

Now let $d > 0$ and choose an orthonormal basis $q_0, q_1, \ldots q_{d-1}$ for $\mathrm{Ker}(A)$, and an orthonormal basis $q_d, q_{d+1}, \ldots$ for the orthogonal complement $\mathrm{Ker}(A)^\perp$. Define the matrix

$$Q_1 = \begin{pmatrix} q_0 & q_1 & \cdots & q_{d-1} & q_d & \cdots \end{pmatrix}. \tag{5.7}$$

Since this is merely a change of orthonormal basis for $\ell^2$, $Q_1$ is an orthogonal matrix. Write $A_1 = Q_1^T A$. Note that for $i = 0, 1, \ldots, d-1$, we have $e_i^T A_1 = q_i^T A = (Aq_i)^T = 0$. Hence the first $d$ rows of $A_1$ are zero. Hence we may write

$$A_1 = \begin{pmatrix} 0_{d\times d} & 0_{d\times\infty} \\ B_{\infty\times d} & C_{\infty\times\infty} \end{pmatrix} \tag{5.8}$$

We claim that $C = C_{\infty\times\infty}$ is invertible, so that by Lemma 5.1.2 we would have a unique QL factorisation $C = Q_2 L_2$ such that $L_2$ has positive diagonal elements. Then we would have

$$A = \left( Q_1 \begin{pmatrix} I_{d\times d} & 0 \\ 0 & Q_2 \end{pmatrix} \right) \begin{pmatrix} 0_{d\times d} & 0 \\ Q_2^T B & L_2 \end{pmatrix}, \tag{5.9}$$

as desired.

Let $S_{\text{left}}$ and $S_{\text{right}}$ be the left and right shift operators respectively. Then $C = S_{\text{left}}^d Q^T A S_{\text{right}}^d$. Since $S_{\text{left}}$ and $S_{\text{right}}$ are Fredholm with indices 1 and $-1$ respectively, by Theorem D.2.5 $C$ is Fredholm with index $d + 0 + 0 - d = 0$. By Theorem D.2.4, $C$ is invertible if and only if it is injective. Suppose $x \in \mathrm{Ker}(C)$. Then $AS_{\text{right}}^d x = 0$. Hence $S_{\text{right}}^d x \in \mathrm{Ker}(A) \cap \mathrm{Span}\{e_d, e_{d+1}, \ldots\} = \{0\}$. Since $S_{\text{right}}^d x = 0 \implies x = 0$, we have that $C$ is invertible, which completes the proof.

For $(ii) \implies (i)$: Let $A = QL$ where the first $d$ rows of $L$ are zero and $L_1 = L_{d:\infty \times d:\infty}$ is invertible. Then $L = S^d_{\text{right}} L_1 S^d_{\text{left}} + L_{0:\infty \times 0:d-1}$. Since the shift operators are Fredholm with index -1 and 1, $L_1$ is invertible, and $L_{0:\infty:d-1}$ is finite rank and hence compact, by Theorem D.2.3 we have that $L$ is a Fredholm operator with index $-d + 0 + d = 0$. Hence by Theorem D.2.5 $A = QL$ is a Fredholm operator with index 0.

For the kernel condition, note that $\text{Ker}(A) = \text{Ker}(L)$ and $\text{Ker}(L_1) = \{0\}$. Therefore $\text{Ker}(A) \cap \text{Span}\{e_d, e_{d-1}, \ldots\} = \text{Ker}(L) \cap \text{Span}\{e_d, e_{d-1}, \ldots\} = S^d_{\text{right}} \text{Ker}(L_1) = \{0\}$. □

*Remark* 5.1.4. To generalise this to non-selfadjoint operators we simply add the condition that the index of the operator is zero to $(i)$. In Section 5.2 we will see that the technical requirement on the kernel is irrelevant for Jacobi operators.

**Definition 5.1.5** (Upper and lower bandwidth)**.** An operator $A$ has *upper bandwidth* $u$ if $A_{ij} = 0$ for $j > i + u$, *lower bandwidth* $l$ if $A_{ij} = 0$ for $i > j + l$, and *bandwidth* $b$ if $b$ is at least the maximum of upper and lower bandwidths of $A$.

**Proposition 5.1.6.** *Let $A \in \mathcal{B}(\ell^2)$ have bandwidth $b$. Suppose that $A$ is Fredholm with $\dim(\text{Ker}(A)) = d \leq b$ and such that $\text{Ker}(A) \cap \text{Span}\{e_d, e_{d+1}, \ldots\} = \{0\}$. Suppose further that $A$ has bandwidth $b$. If $A = QL$ is a QL factorisation, then $Q$ has upper bandwidth $b$ and $L$ has lower bandwidth $2b$.*

*Proof.* When $d = 0$, for any $i, j \geq 0$ we have

$$e_i^T Q e_j = e_i^T A L^{-1} e_j = \sum_{k=0}^{\infty} a_{i,k} \left[L^{-1}\right]_{k,j} = \sum_{k=i-b}^{i+b} a_{i,k} \left[L^{-1}\right]_{k,j}$$
$$e_i^T L e_j = e_i^T Q^T A e_j = \sum_{k=0}^{\infty} q_{k,i} a_{k,j} = \sum_{k=j-b}^{j+b} q_{k,i} a_{k,j}$$

Since $L^{-1}$ is lower triangular, $e_i^T Q e_j = 0$ for $j > i + b$. Hence $Q$ has upper bandwidth $b$. From this it follows that $e_i^T L e_j = 0$ for $i > j + 2b$. Hence $L$ has lower bandwidth $2b$.

When $d > 0$, by Theorem 5.1.3, we can write $L$ in block form

$$L = \begin{pmatrix} 0_{d\times d} & 0 \\ b & L_1 \end{pmatrix},$$

in which $L_1$ is invertible because it has positive diagonal elements. If we define the matrix

$$L^+ = \begin{pmatrix} I_{d\times d} & 0 \\ -L_1^{-1} b & L_1^{-1} \end{pmatrix},$$

the $LL^+e_j = e_j$ for all $j \geq d$. Hence we can show just as above that $e_i^T Q e_j = e_i^T A L^+ e_j = 0$ for $j > i + b$ (which is within the valid range since $d \leq b$) and then that the lower bandwidth of $L$ is $2b$ in the same way. □

These results appear quite specialised at first. What about non-Fredholm operators? For example, what about the free Jacobi operator $\Delta$ from Chapter 4? Since 0 is in the essential spectrum of $\Delta$, by definition it does not have closed range and hence is not Fredholm. In the sequel we show that $\Delta$ and many other non-Fredholm operators do not have a QL factorisation.

### 5.1.2 Nonexistence

Somewhat surprisingly, a QL factorisation does not exist for all bounded operators $A \in \mathcal{B}(\ell^2)$. The approach we use to prove and understand this involves using what we will call *tempered distributional sequences*, which are sequences that grow polynomially.

**Definition 5.1.7** (Schwartz sequences)**.** Define the following subspace of $\ell^2$, the space of *Schwartz sequences*.

$$\ell_{\mathcal{S}} = \left\{ v \in \ell^2 : \sum_{k=0}^{\infty} (k+1)^m |v_k|^2 < \infty \text{ for } m = 0, 1, 2, \ldots \right\}. \tag{5.10}$$

*Remark* 5.1.8. With the countable family of inner products $\langle v, w \rangle_m = \sum_{k=0}^{\infty} (1+k)^{2m} v_k w_k$ for $m = 0, 1, 2, \ldots$, it is a *countably Hilbert space* [GV64, Ch. 1, Sec. 3]. The inner products induce norms $\|v\|_{2,m} = \langle v, v \rangle_m$. The notion of convergence on $\ell_{\mathcal{S}}$ is that of convergence in a single one of these norms.

**Lemma 5.1.9** (Tempered distributional sequences)**.** *The dual of $\ell_{\mathcal{S}}$ with respect to the bilinear form $\langle v, w \rangle = \sum_{k=1}^{\infty} v_k w_k$ is the space of* tempered distributional sequences*,*

$$\ell_{\mathcal{S}}^* = \left\{ w \in \mathbb{R}^{\infty} : \text{ there exists } m' \geq 0 \text{ such that } \sum_{k=0}^{\infty} \left| (1+k)^{-m'} w_k \right|^2 < \infty \right\}. \tag{5.11}$$

*Proof.* We must show that the linear functional $v \to \langle v, w \rangle$ is continuous on $\ell_{\mathcal{S}}$ if and only if $w$ is in the space defined in equation (5.11). By Remark 5.1.8 this is equivalent to the existence of $m', C \geq 0$ such $\langle v, w \rangle \leq C \|v\|_{2,m'}$ for all $v \in \ell_{\mathcal{S}}$.

Suppose that $w$ is a vector in the set described in equation (5.11). Then we may take $C = \left( \sum_{k=0}^{\infty} |(1+k)^{-m'} w_k|^2 \right)^{\frac{1}{2}}$ in the above condition. Conversely if there exists

$m', C \geq 0$ satisfying the above condition then

$$\begin{aligned}\left(\sum_{k=0}^{\infty}\left|(1+k)^{-m'}w_k\right|^2\right)^{\frac{1}{2}} &= \sup_{v\in\ell^0}\frac{\sum_{k=0}^{\infty}v_k(1+k)^{-m'}w_k}{\|v\|_2}\\ &= \sup_{v\in\ell^0}\frac{\langle v,w\rangle}{\|v\|_{2,m'}}\\ &\leq C,\end{aligned}$$

where $\ell^0$ is the space of finite sequences. Hence $w \in \ell^*_{\mathcal{S}}$. □

The triple of spaces $\ell_{\mathcal{S}} \subset \ell^2 \subset \ell^*_{\mathcal{S}}$ is the discrete analogue of the triple $\mathcal{S}(\mathbb{R}) \subset L^2(\mathbb{R}) \subset \mathcal{S}(\mathbb{R})^*$ of Schwartz functions and tempered distributions on the real line, well known in distribution theory. They are both examples of the more general concept of a *Gelfand triple*, or a *rigged Hilbert space* [GV64, Ch. 1,Sec. 4].

**Theorem 5.1.10.** *A banded operator $A \in \mathcal{B}(\ell^2)$ is a continuous operator mapping $\ell_{\mathcal{S}}$ into itself. Hence it can be extended to a continuous linear operator on $\ell^*_{\mathcal{S}}$.*

*Proof.* Let $A$ have bandwidth $b$. Then for any $m$ and any $v \in \ell_{\mathcal{S}}$,

$$\begin{aligned}\|Av\|^2_{2,m} &= \sum_{k=0}^{\infty}(1+k)^{2m}\left|\sum_{j=k-b}^{k+b}a_{kj}v_j\right|^2\\ &\leq \sum_{k=0}^{\infty}(1+k)^{2m}\left(\sum_{j=k-b}^{k+b}|a_{kj}|^2\right)\left(\sum_{j=k-b}^{k+b}|v_j|^2\right)\\ &= \sum_{j=0}^{\infty}\|Ae_j\|_2^2|v_j|^2\sum_{k=j-b}^{j+b}(1+k)^{2m}\\ &\leq \|A\|_2^2(2b+1)\sum_{j=0}^{\infty}(1+j+b)^{2m}|v_j|^2\\ &\leq \|A\|_2^2(2b+1)(1+b)^{2m}\|v\|^2_{2,m}.\end{aligned}$$

Hence $A$ is continuous on $\ell_{\mathcal{S}}$.

By the same proof, this is also true of $A^T$. For any $w \in \ell^*_{\mathcal{S}}$, $Aw$ is the vector in $\ell^*_{\mathcal{S}}$ such that

$$\langle v, Aw\rangle = \langle A^T v, w\rangle \text{ for all } v \in \ell_{\mathcal{S}},$$

which exists and is unique by the definition of $\ell^*_{\mathcal{S}}$ by duality. The operator $A$ is continuous on $\ell^*_{\mathcal{S}}$ because the map $(v, w) \mapsto \langle v, Aw\rangle$ is a composition of the continuous maps $(v, w) \mapsto (A^T v, w)$ and $(v, w) \to \langle v, w\rangle$. □

*Remark* 5.1.11. This theorem is analogous to the fact that a pseudodifferential operator is a continuous map from the space of Schwartz functions $\mathcal{S}(\mathbb{R})$ into itself, so can be extended to a continuous map on tempered distributions $\mathcal{S}(\mathbb{R})^*$ by duality using the $L^2(\mathbb{R})$ inner product.

**Definition 5.1.12.** Let $A \in \mathcal{B}(\ell^2)$ be banded. The *distributional kernel* of $A$ is the subspace of $\ell^*_{\mathcal{S}}$ whose elements are mapped to 0 by $A$.

**Lemma 5.1.13.** *Let $A \in \mathcal{B}(\ell^2)$ be banded and selfadjoint with a finite dimensional distributional kernel consisting only of $\ell^2$ vectors. Then $A$ is a compact perturbation of an invertible operator.*

*Proof.* Suppose the distributional kernel consists of only $\ell^2$ vectors, $v_0, v_1, \ldots, v_{d-1}$. Consider the compact perturbation $\tilde{A} = A + \sum_{k=0}^{d-1} v_k v_k^T$. Then $\tilde{A}$ has zero distributional kernel. If we can show that $\tilde{A}$ is invertible, then this will show that $A$ is a compact perturbation of an invertible operator.

To prove this we actually appeal to a very powerful theorem, the spectral theorem for rigged Hilbert spaces, following [GV64, Ch. 1, Sec. 3,4]. Theorem 5' in Chapter 1 Section 4.5 of [GV64] states that a selfadjoint operator on a rigged Hilbert space has a complete set of eigenvectors in $\ell^*_{\mathcal{S}}$. This implies that if the distributional kernel is $\{0\}$ then $0 \notin \sigma(A)$, so the operator is invertible. □

**Theorem 5.1.14.** *Let $A \in \mathcal{B}(\ell^2)$ be banded and selfadjoint with finite dimensional distributional kernel. If $A$ has a QL factorisation then $A$ is a compact perturbation of an invertible operator.*

*Proof.* Let $w_0, w_1, w_2, \ldots, w_{d-1}$ be a basis for the distributional kernel of $A$. If $A = QL$ is the QL factorisation, then $w_0, w_1, \ldots, w_{d-1}$ is also a basis for the distributional kernel of $L$. Considering simultaneous forward substitution on the system $L(w_0, w_1, \ldots, w_{d-1}) = (0, 0, \ldots, 0)$, we see that $L$ must have at least $d$ zeros on its diagonal.

By back substitution on the upper triangular system $L^T v = 0$, which has at least $d$ zeros on its diagonal, we see that $L^T$ has $d$ linearly independent vectors $v_0, v_1, \ldots, v_{d-1}$ in $\ell^0 \subset \ell^2$ such that $L^T v_i = 0$. Hence the vectors $Qv_0, Qv_1, \ldots, Qv_{d-1}$ (which are all in

$\ell^2$) are all in the kernel of $A^T$, which is $A$ since $A$ is selfadjoint. Since the distributional kernel of $A$ has dimension $d$, we must have that $Qv_0, Qv_1, \ldots, Qv_{d-1}$ is a basis for the distributional kernel. Hence the distributional kernel is contained in $\ell^2$.

By Lemma 5.1.13, $A$ is a compact perturbation of an invertible operator. □

**Corollary 5.1.15** (Nonexistence of QL factorisation)**.** *Let $A \in \mathcal{B}(\ell^2)$ be banded and selfadjoint with finite dimensional distributional kernel. If $0 \in \sigma_{ess}(A)$ then $A$ does not possess a QL factorisation.*

*Proof.* Since $0 \in \sigma_{ess}(A)$, $A$ cannot be a compact perturbation of an invertible operator. By Theorem 5.1.14, $A$ cannot have a QL factorisation. □

*Remark* 5.1.16. The condition that $A$ have a finite dimensional distributional kernel is satisfied by Jacobi operators. We therefore have some simpler statements of these theorems applied to Jacobi operators in Section 5.2.

### 5.1.3 Framework for computation of QL factorisations

How do we compute the QL factorisation of a banded and highly structured matrix? The following theorem shows that if an analytical QL decomposition can be computed for the "tail" of a banded matrix, then the rest can be finished by the finite dimensional QL algorithm.

**Theorem 5.1.17.** *Let $A$ be a banded, bounded operator on $\ell^2$ with bandwidth $b$ and block form*

$$A = \begin{pmatrix} A_n & B \\ C & A_\infty \end{pmatrix}, \tag{5.12}$$

*where $A_n \in \mathbb{R}^{n\times n}$ and $A_\infty$ has an* a priori *known QL factorisation $A_\infty = Q_\infty L_\infty$. Then the QL factorisation of $A$ can be computed in the following sense. The information can be stored that allows the computation of $Qv$ and $Lv$ for any $v \in \ell^0$ in a finite number of operations.*

*Proof.* Using the QL factorisation of $A_\infty$, we have

$$\begin{pmatrix} I_n & 0 \\ 0 & Q_\infty^T \end{pmatrix} \begin{pmatrix} A_n & B \\ C & A_\infty \end{pmatrix} = \begin{pmatrix} A_n & B \\ D & L_\infty \end{pmatrix} =: \tilde{A}, \tag{5.13}$$

where $D = Q_\infty^T C$. By Proposition 5.1.6, $Q$ has upper bandwidth $b$, and since $A$ has bandwidth $b$, $D$ is only nonzero in its first $b$ rows. Therefore, $D$ is only nonzero in its

first $2b$ rows. Also, because $A$ has bandwidth $b$, the matrix $B$ is only nonzero in its first $b$ columns. Hence we can rewrite $\tilde{A}$ in the form

$$\tilde{A} = \begin{pmatrix} A_{n+b,n+b} & 0 \\ \tilde{D} & M_\infty \end{pmatrix}, \tag{5.14}$$

where $A_{n+b} \in \mathbb{R}^{(n+b)\times(n+b)}$ and the matrices $\tilde{D}$ and $M_\infty$ are $D$ and $L_\infty$ with their first $b$ rows removed, respectively.

Now, let the QL factorisation of $A_{n+b}$ be $Q_{n+b}L_{n+b}$. Define

$$Q = \begin{pmatrix} I_n & 0 \\ 0 & Q_\infty \end{pmatrix} \begin{pmatrix} Q_{n+b} & 0 \\ 0 & I_\infty \end{pmatrix}. \tag{5.15}$$

Then $Q^T A$ is lower triangular. We can also define

$$L = \begin{pmatrix} L_{n+b} & 0 \\ 0 & I_\infty \end{pmatrix} \begin{pmatrix} I_n & 0 \\ \tilde{D} & M_\infty \end{pmatrix}. \tag{5.16}$$

This completes the proof. □

*Remark* 5.1.18. We have not proven that knowledge of the "tail" of the matrix is *required* in order to compute a QL factorisation, merely that it is sufficient. However, the author conjectures that it must be necessary because in order to create zeros on the upper triangular part of $A$ using Householder reflections or Givens rotations acting on the left, it appears that we must begin at the right and work leftwards or else the zeros created will be undone. It would be interesting to see a proof formalising this reasoning.

## 5.2 QL factorisation of Jacobi operators

In this section we take the above research on QL factorisations and specialise to Jacobi operators, usually obtaining more precise and succinct results.

### 5.2.1 Existence for Jacobi operators

**Lemma 5.2.1.** *Let $J$ be a Jacobi operator. Then the distributional kernel has dimension at most 1. Furthermore, the first element of the vector must be nonzero.*

*Proof.* Suppose that $v$ is in the distributional kernel of $J$ and $v_0 \neq 0$. Then since $\beta_k \neq 0$ for all $k$, each row determines a successive element of the vector $v$. Hence there can only be one such vector. Now suppose that $v_0 = 0$. Then $Jv = 0$ is equivalent to a lower triangular system $\tilde{J}\tilde{v} = 0$, where $\tilde{J}$ is $J$ with the first column removed and $\tilde{v}$ is $v$ with the first element removed. Since $\beta_k \neq 0$, by forward substitution we obtain $\tilde{v} = 0$, and hence $v = 0$. □

**Theorem 5.2.2.** *Let $J$ be a Jacobi operator. Then $J$ has a QL factorisation if and only if $0 \notin \sigma_{ess}(J)$. Furthermore, the diagonal elements of $L$ can be made positive, unless $J$ is singular, in which case the top-left entry of $L$ is zero.*

*Proof.* By Lemma 5.2.1, the distributional kernel of $J$ is finite dimensional. If $J$ has a QL factorisation then by Theorem 5.1.14 it is a compact perturbation of an invertible operator. Hence $0 \notin \sigma_{ess}(J)$.

Now suppose that $0 \notin \sigma_{ess}(J)$. Then $J$ is a compact perturbation of an invertible operator, so by Theorem D.2.3 it is Fredholm. By Lemma 5.2.1, $\operatorname{Ker}(J) \cap \{e_1, e_2, \ldots\} = \{0\}$. Hence by Theorem 5.1.3 there exists a QL factorisation such that $L$ has positive diagonal elements except if $J$ is singular, in which case the top row of $L$ is zero. □

### 5.2.2 Practical computation and storage for Jacobi operators

**Lemma 5.2.3.** *Let $J$ be a Jacobi operator. If $J = QL$ is a QL factorisation, then $Q$ is lower Hessenberg and $L$ has lower bandwidth 2.*

*Proof.* This follows from Proposition 5.1.6 and Lemma 5.2.1 with $b = 1$ and $d \leq 1$. □

By the above lemma we only need to consider $Q$ to be a lower Hessenberg matrix. The next lemma shows that such operators can be parametrised quite simply.

**Lemma 5.2.4** (Schur parametrisation [Gra86])**.** *Let $Q$ be a lower Hessenberg orthogonal operator. There exists a sequence of angles $\theta_1, \theta_2, \ldots \in (-\pi, \pi]$ such that*

$$Q = Q(\theta_0, \theta_1, \ldots) := \begin{pmatrix} c_{-1}c_0 & -s_0 & 0 & 0 & \\ c_{-1}s_0c_1 & c_0c_1 & -s_1 & 0 & \ddots \\ c_{-1}s_0s_1c_2 & c_0s_1c_2 & c_1c_2 & -s_2 & \ddots \\ c_{-1}s_0s_1s_2c_3 & c_0s_1s_2c_3 & c_1s_2c_3 & c_2c_3 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{5.17}$$

*where* $s_i = \sin(\theta_i)$, $c_i = \cos(\theta_i)$ *and* $c_{-1} = \pm 1$. *Written explicitly, for* $i, j = 0, 1, \ldots,$

$$q_{ij} = \begin{cases} -s_i & \text{if } j = i+1 \\ c_{j-1} s_j s_{j+1} \cdots s_{i-1} c_i & \text{if } j \leq i \\ 0 & \text{otherwise} \end{cases} \tag{5.18}$$

*Note that when* $\theta_i = 0$ *for* $i \neq k$ *for a given* $k$, *then we get a Givens rotation,*

$$Q(\theta_k e_k) = \operatorname{diag}\left(I_k, \begin{pmatrix} c_k & -s_k \\ s_k & c_k \end{pmatrix}, I_\infty\right), \tag{5.19}$$

*and the full matrix is an infinite product of these givens rotations,*

$$Q(\theta_0, \theta_1, \ldots) = \lim_{N \to \infty} Q(\theta_N e_N) \cdots Q(\theta_0 e_0) \operatorname{diag}(c_{-1}, 1, 1, \ldots). \tag{5.20}$$

Consider Theorem 5.1.17. During the proof for the computation of a QL factorisation, we first use an analytical form for the tail $A_\infty = Q_\infty L_\infty$. We first apply $\operatorname{diag}(I_n, Q_\infty^T)$ and then apply finitely many Givens rotations to finish the computation. Let us be more specific in the case of a Jacobi operator and use the parametrisation of orthogonal lower Hessenberg matrices. What remains after performing the first $\operatorname{diag}(I_n, Q_\infty^T)$ must be of the following form.

$$Q(0, 0, \ldots, 0, \theta_k, \theta_{k+1}, \ldots)^T J = \begin{pmatrix} \alpha_0 & \beta_0 & & & & & & \\ \beta_0 & \alpha_1 & \beta_1 & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & \beta_{k-2} & \alpha_{k-1} & \beta_{k-1} & & & \\ & & & \gamma_k^1 & \gamma_k^0 & & & \\ & & & \ell_{k+1}^2 & \ell_{k+1}^1 & \ell_{k+1}^0 & & \\ & & & & \ell_{k+2}^2 & \ell_{k+2}^1 & \ell_{k+2}^0 & \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix} \tag{5.21}$$

Considering what the next step in the algorithm should be in order to compute a QL decomposition allows us to derive a backward recurrence relating the entries of $J$, $Q$, $L$ and the sequences $\gamma^0$ and $\gamma^1$ generated by this hypothetical procedure. The purpose of this is that in some specific cases of highly structured Jacobi operators, the tail of the

backward recurrence can be solved analytically, thus allowing us to "work backwards from infinity".

**Lemma 5.2.5.** *Suppose* $c_0, c_1, \ldots$, $s_0, s_1, \ldots$, $\gamma_0^1, \gamma_1^1, \ldots$, $\gamma_1^1, \gamma_2^1, \ldots$ *are sequences satisfying the following backward recurrence relationship,*

$$c_k = \frac{\operatorname{sign}(\alpha_{k+1})\gamma_{k+1}^0}{\sqrt{\left(\gamma_{k+1}^0\right)^2 + \left(\beta_k\right)^2}}, \quad s_k = -\frac{\operatorname{sign}(\alpha_{k+1})\beta_k}{\sqrt{\left(\gamma_{k+1}^0\right)^2 + \left(\beta_k\right)^2}}, \tag{5.22}$$

$$\gamma_k^1 = c_k\beta_{k-1}, \quad \gamma_k^0 = c_k\alpha_k + s_k\gamma_{k+1}^1, \tag{5.23}$$

*and* $c_{-1} = \operatorname{sign}(\gamma_0^0\alpha_0)$. *Then if we define the operator* $Q$ *to be the orthogonal lower Hessenberg matrix in Lemma 5.2.4 with these* $c_k$*'s and* $s_k$*'s, and we define the operator* $L$ *to be*

$$L = \begin{pmatrix} \ell_0^0 & & & \\ \ell_1^1 & \ell_1^0 & & \\ \ell_2^2 & \ell_2^1 & \ell_2^0 & \\ & \ddots & \ddots & \ddots \end{pmatrix}, \tag{5.24}$$

*where*

$$\ell_k^0 = -\frac{\beta_{k-1}}{s_{k-1}}, \quad \ell_k^1 = c_{k-1}\gamma_k^1 - s_{k-1}\alpha_{k-1}, \quad \ell_k^2 = -s_{k-1}\beta_{k-2},$$

*then* $J = QL$ *where* $J$ *is the Jacobi operator with diagonal entries* $\alpha_0, \alpha_1, \ldots$ *and off-diagonal entries* $\beta_0, \beta_1, \ldots$.

*Proof.* Simply by virtue of $Q$ being lower Hessenberg and $L$ having lower bandwidth 2, we have

$$e_k^T QLe_j = \begin{cases} 0 & j > k+1 \\ q_{k,k+1}\ell_{k+1}^0 & j = k+1 \\ q_{k,k}\ell_k^0 + q_{k,k+1}\ell_{k+1}^1 & j = k \\ q_{k,k-1}\ell_{k-1}^0 + q_{k,k}\ell_k^1 + q_{k,k+1}\ell_{k+1}^2 & j = k-1 \\ q_{k,j}\ell_j^0 + q_{k,j+1}\ell_{j+1}^1 + q_{k,j+2}\ell_{j+2}^2 & j < k-1 \end{cases}.$$

We show that each case gives the appropriate entry of $J$. The first case has 0 as it should. The second case is the following.

$$q_{k,k+1}\ell_{k+1}^0 = -s_k\left(-\frac{\beta_k}{s_k}\right) = \beta_k.$$

For the remaining cases first note that

$$c_{k-1}\ell_k^0 = c_{k-1}\left(-\frac{\beta_{k-1}}{s_{k-1}}\right) = \gamma_k^0 = c_k\alpha_k + s_k\gamma_{k+1}^1.$$

The third case is

$$\begin{aligned} q_{k,k}\ell_k^0 + q_{k,k+1}\ell_{k+1}^1 &= c_{k-1}c_k\ell_k^0 - s_k\ell_{k+1}^1 \\ &= c_k\left(c_k\alpha_k + s_k\gamma_{k+1}^1\right) - s_k\left(c_k\gamma_{k+1}^1 - s_k\alpha_k\right) \\ &= \alpha_k. \end{aligned}$$

The fourth case is

$$\begin{aligned} q_{k,k-1}\ell_{k-1}^0 + q_{k,k}\ell_k^1 + q_{k,k+1}\ell_{k+1}^2 &= c_{k-2}s_{k-1}c_k\ell_{k-1}^0 + c_{k-1}c_k\ell_k^1 - s_k\ell_{k+1}^2 \\ &= s_{k-1}c_k\left(c_{k-1}\alpha_{k-1} + s_{k-1}\gamma_k^1\right) \\ &\quad + c_{k-1}c_k\left(c_{k-1}\gamma_k^1 - s_{k-1}\alpha_{k-1}\right) + s_ks_k\beta_{k-1} \\ &= c_k\gamma_k^1 + s_k^2\beta_{k-1} \\ &= \beta_{k-1}. \end{aligned}$$

This fifth and final case is

$$q_{k,j}\ell_j^0 + q_{k,j+1}\ell_{j+1}^1 + q_{k,j+2}\ell_{j+2}^2 = s_{j+2}s_{j+3}\cdots s_{k-1}c_k\left(c_{j-1}s_js_{j+1}\ell_j^0 + c_js_{j+1}\ell_{j+1}^1 + c_{j+1}\ell_{j+2}^2\right).$$

The bracketed term is equal to

$$\begin{aligned} &c_{j-1}s_js_{j+1}\ell_j^0 + c_js_{j+1}\ell_{j+1}^1 + c_{j+1}\ell_{j+2}^2 \\ &\quad = s_js_{j+1}(c_j\alpha_j + s_j\gamma_{j+1}^1) + c_js_{j+1}(c_j\gamma_{j+1}^1 - s_j\alpha_j) - s_{j+1}c_{j+1}\beta_j \\ &\quad = s_{j+1}\gamma_{j+1}^1 - s_{j+1}c_{j+1}\beta_j \\ &\quad = 0. \end{aligned}$$

This completes the proof. □

The above lemma can be used to provide an analytical solution for the tail of the sequences in question for $Q$ and $L$. The following is exactly that for Toeplitz-plus-finite-rank Jacobi operators.

**Theorem 5.2.6.** *Let $J$ be a Toeplitz-plus-finite-rank Jacobi operator in with diagonal entries $\alpha_0, \alpha_1, \ldots, \alpha_n, t_0, t_0, \ldots$ and offdiagonal entries $\beta_0, \beta_1, \ldots, \beta_{n-1}, t_1, t_1, \ldots$. A QL*

*decomposition exists if and only if* $|t_0| > 2t_1$*, in which case,* $Q$ *is an orthogonal lower Hessenberg matrix as in Lemma 5.2.4 with*

$$s_k = s := \frac{-t_0 + \operatorname{sign}(t_0)\sqrt{t_0^2 - 4t_1^2}}{2t_1}, \quad c_k = c := \sqrt{1 - s^2}, \text{ for all } k > n, \tag{5.25}$$

*and* $s_k, c_k$ *for* $k \leq n$ *computed by the backward recurrence in Lemma 5.2.5. The lower triangular matrix* $L$ *is as stated in Lemma 5.2.5.*

*Proof.* The essential spectrum of $J$ is the interval $[t_0 - 2t_1, t_0 + 2t_1]$ by Theorem 4.3.21. Hence by Theorem 5.2.2 there exists a QL decomposition if and only if $|t_0| > 2t_1$.

All we need to check is that the sequences $s_0, s_1, \ldots$ and $c_{-1}, c_0, c_1 \ldots$ defined in the statement of this theorem generate an orthogonal matrix $Q$ and a lower triangular matrix $L$ such that $J = QL$.

When $|t_0| > 2t_1$, this makes $|s| < 1$ and $|c| < 1$, so that the $Q$ that is defined here is orthogonal. $L$ is constructed so that it is lower triangular and $J = QL$ as long as the recurrence for $s_0, s_1, \ldots, c_{-1}, c_0, c_1 \ldots, \gamma_0^0, \gamma_1^0, \ldots, \gamma_1^1, \gamma_2^1, \ldots$ satisfy the equations

$$c_k = \frac{\operatorname{sign}(\alpha_{k+1})\gamma_{k+1}^0}{\sqrt{\left(\gamma_{k+1}^0\right)^2 + \left(\beta_k\right)^2}}, \quad s_k = -\frac{\operatorname{sign}(\alpha_{k+1})\beta_k}{\sqrt{\left(\gamma_{k+1}^0\right)^2 + \left(\beta_k\right)^2}}, \tag{5.26}$$

$$\gamma_k^1 = c_k\beta_{k-1}, \quad \gamma_k^0 = c_k\alpha_k + s_k\gamma_{k+1}^1, \tag{5.27}$$

Therefore, we just need to check that these equations are satisfied if we have the $s_k$'s and $c_k$'s as defined in the statement of the theorem.

For $k \leq n$ define the $\gamma_k^0$'s and $\gamma_k^1$'s to follow from the definitions of $s_k$'s and $c_k$'s in the statement of the theorem: $\gamma_k^1 = c_k\beta_{k-1} = ct_1$ and $\gamma_k^0 = c_k\alpha_k + s_k\gamma_{k+1}^1 = ct_0 + s\gamma^1 = ct_0 + cst_1$. Then the definition of $c_k$ from $\gamma_k^0$ is consistent:

$$\begin{aligned}
\frac{\operatorname{sign}(t_0)\gamma_{k+1}^0}{\sqrt{\left(\gamma_{k+1}^0\right)^2 + \left(\beta_k\right)^2}} &= \operatorname{sign}(t_0)\frac{ct_0 + cst_1}{\sqrt{\left(ct_0 + cst_1\right)^2 + \left(t_1\right)^2}} \\
&= c\operatorname{sign}(t_0)\frac{t_0 + st_1}{\sqrt{\left(t_0 + st_1\right)^2 - s^2\left(t_0 + st_1\right)^2 + \left(t_1\right)^2}} \\
&= c\operatorname{sign}(t_0)\frac{t_0 + st_1}{\sqrt{\left(t_0 + st_1\right)^2 - s^2\left(t_0 + st_1\right)^2 + \left(t_1\right)^2}} \\
&= c\operatorname{sign}(t_0)\frac{t_0 + st_1}{\sqrt{\left(t_0 + st_1\right)^2 + (t_1 + t_0 s + t_1 s^2)(t_1 - t_0 s - t_1 s^2)}}
\end{aligned}$$

$$= c\text{sign}(t_0)\text{sign}(t_0 + st_1),$$

because $s$ satisfies $t_1 s^2 + t_0 s + t_1 = 0$. Also, since $|t_0| > 2t_1$, we have that $\text{sign}(t_0 + st_1) = \text{sign}(t_0+)$. Hence we have consistency for $c$. It is also tedious but possible to check the same is true for $s$ to complete the proof.

□

### 5.2.3 Example QL factorisations of Jacobi operators

Now let us consider some examples of infinite dimensional QL factorisations. The first example is computed by hand using the Theorem 5.2.6. The second example is computed numerically using an open source software package *SpectralMeasures*. This package is written in Julia by the author and collaborator Sheehan Olver (University of Sydney), using the open source package *ApproxFun*. All of the code is available to download at http://www.github.com/JuliaApproximation/SpectralMeasures.jl. See Appendix A for more information and parts of the code.

By Theorem 5.2.2, the free Jacobi operator $\Delta$ does not have a QL factorisation. However, if we shift it by a sufficiently large multiple of the identity then it will. Consider $J = \Delta - \frac{5}{4}I$. Then the QL decomposition is

$$\Delta - \frac{5}{4}I = \begin{pmatrix} -\frac{\sqrt{3}}{2} & \frac{1}{2} & & & & \\ -\frac{\sqrt{3}}{4} & -\frac{3}{4} & \frac{1}{2} & & & \\ -\frac{\sqrt{3}}{8} & -\frac{3}{8} & -\frac{3}{4} & \frac{1}{2} & & \\ -\frac{\sqrt{3}}{16} & -\frac{3}{16} & -\frac{3}{8} & -\frac{3}{4} & \frac{1}{2} & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \frac{\sqrt{3}}{2} & & & & \\ -1 & 1 & & & \\ \frac{1}{4} & -1 & 1 & & \\ & -\frac{1}{4} & -1 & 1 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}$$

When the shift $\lambda$ in $\Delta - \lambda I$ gets closer to $\pm 1$, which are the edges of the essential spectrum of $\Delta$, the value of the elements on the upper diagonal tends to $\pm 1$, so that in the limit $Q$ becomes the backward shift operator. In the limit the QL facorisation becomes

$$\Delta - I = \begin{pmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ & & 0 & 1 & & \\ & & & 0 & 1 & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \frac{1}{2} & & & & \\ -1 & \frac{1}{2} & & & \\ \frac{1}{2} & -1 & \frac{1}{2} & & \\ & \frac{1}{2} & -1 & \frac{1}{2} & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

which is not a QL factorisation because this $Q$ is not orthogonal. Note that any tridiagonal matrix has such a factorisation as a backward shift of a lower triangular matrix.

Now for a nontrivial example. Let us use an example from Chapter 4, a $3 \times 3$ perturbation of $\Delta$,

$$J = \begin{pmatrix} \frac{3}{4} & 1 & & & & \\ 1 & -\frac{1}{4} & \frac{3}{4} & & & \\ & \frac{3}{4} & \frac{1}{2} & \frac{1}{2} & & \\ & & \frac{1}{2} & 0 & \frac{1}{2} & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}$$

Note that since $0 \in \sigma_{ess}(J)$, by Theorem 5.2.2, $J$ does not have a QL factorisation. It will if we use a shift. Using the techniques of Chapter 4 we can compute eigenvalues at approxiately $-1.173$ and $1.580$. If we use $-1.173$ as a shift, the QL factorisation is (to 3 decimal places),

$$\begin{pmatrix} 0.400 & 0.917 & & & & \\ -0.769 & 0.335 & 0.544 & & & \\ 0.414 & -0.180 & 0.695 & 0.559 & & \\ -0.231 & 0.101 & -0.389 & 0.687 & 0.559 & \\ 0.129 & -0.057 & 0.218 & -0.384 & 0.687 & \ddots \\ \vdots & & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} 0.000 & & & & & \\ 2.098 & 1.091 & & & & \\ 0.544 & 1.024 & 1.377 & & & \\ & 0.420 & 1.280 & 0.894 & & \\ & & 0.280 & 1 & 0.894 & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}.$$

For this choice of shift, which is not exactly an eigenvalue of $J$, the top-left entry of $L$ is not exactly zero. For this example it is $2.429 \times 10^{-15}$. Here for both $Q$ and $L$ the fifth, sixth, etc. columns are each a shift of the previous column.

## 5.3 The shifted QL algorithm

An issue discussed at the start of the chapter is that for the infinite dimensional QR algorithm, shifts cannot be employed to generate rapid convergence of the bottom-right entry to an eigenvalue (if one exists), as is done in the finite dimensional case, because there is no bottom-right entry. In this section we show that shifts can be employed for the QL algorithm to generate rapid convergence of the top-left entry to an eigenvalue of the operator (if one exists).

**Lemma 5.3.1** (Perfect shift)**.** *Suppose $\lambda$ is an eigenvalue of $J$ and there exists a QL factorisation*

$$QL = J - \lambda I.$$

*Then*

$$Q^T J Q = LQ + \lambda I = \begin{pmatrix} \lambda & \\ & \tilde{J} \end{pmatrix},$$

*where $\tilde{J}$ is a Jacobi operator in which $\tilde{J} - \lambda I$ is nonsingular.*

*Proof.* By Theorem 5.2.2, since $J - \lambda I$ is singular, $L$ has a zero in the top-left entry. Therefore the first row of $LQ$ is zero. By symmetry the first column of $LQ$ is zero. Adding $\lambda I$ will give the block form above.

Suppose for a contradiction that $\tilde{J} - \lambda I$ is singular. Then $\lambda$ must be an eigenvalue of $\tilde{J}$, or else the QL factorisation would not have existed by Theorem 5.2.2. By undoing the similarity transformation $Q^T J Q$, we see that this would imply that $\lambda$ is a multiple eigenvalue of $J$, which is not possible as Jacobi operators have simple spectrum. □

What happens if we use a shift which is *close* to an eigenvalue of $J$? We follow the proof of [Par80, Thm 8.6.1] very closely, which deals with finite dimensional Jacobi matries. We merely demonstrate that each step is still valid for any Jacobi operator satisfying the condition in equation (5.28), with no further assumptions on the structure of the Jacobi operator necessary.

**Lemma 5.3.2** (Inverse iteration lemma)**.** *Let $J$ be a Jacobi operator with an eigenpair $(\lambda_0, v_0)$ satisfying*

$$0 < |\lambda_0| < \eta := \min_{\lambda \in \sigma(J) \setminus \lambda_0} |\lambda|, \tag{5.28}$$

*Define for $k = 1, 2, \ldots,$ the* inverse iteration*,*

$$u_0 = e_0, \qquad u_k = \frac{J^{-1} u_{k-1}}{\|J^{-1} u_{k-1}\|_2}. \tag{5.29}$$

*Then*

$$u_k = v_0 + \mathcal{O}\left( \left| \frac{\lambda_0}{\eta} \right|^k \right). \tag{5.30}$$

*Remark* 5.3.3. Note that no structure beyond that in equation (5.28) is assumed for $J$. It could even be unbounded.

*Proof.* Without loss of generality, assume $\|v_0\|_2 = 1$. Then we may write

$$u_k = \cos(\theta_k)v_0 + \sin(\theta_k)w_k, \tag{5.31}$$

where $\theta_k \in (-\pi/2, \pi/2)$, $\|w_k\|_2 = 1$ and $\langle v_0, w_k\rangle = 0$. Multiplying $J^{-1}$ on this equality for $u_{k-1}$, and using $u_k = J^{-1}u_{k-1}/\|J^{-1}u_{k-1}\|_2$ gives

$$u_k = \frac{\cos(\theta_{k-1})}{\lambda_0\|J^{-1}u_{k-1}\|_2}v_0 + \frac{\sin(\theta_{k-1})}{\|J^{-1}u_{k-1}\|_2}J^{-1}w_{k-1}. \tag{5.32}$$

Now, since $J$ and hence $J^{-1}$ is selfadjoint, we have

$$\langle v_0, J^{-1}w_{k-1}\rangle = \langle J^{-1}v_0, w_{k-1}\rangle = \lambda_0^{-1}\langle v_0, w_{k-1}\rangle = 0. \tag{5.33}$$

Therefore, equation (5.31) and equation (5.32) both give the same orthogonal decomposition onto $\mathrm{Span}\{v_0\}$ and $\mathrm{Span}\{v_0\}^\perp$. Equating the two parts gives

$$\cos(\theta_k) = \frac{\cos(\theta_{k-1})}{\lambda_0\|J^{-1}u_{k-1}\|_2}, \qquad \sin(\theta_k)w_k = \frac{\sin(\theta_{k-1})}{\|J^{-1}u_{k-1}\|_2}J^{-1}w_{k-1}. \tag{5.34}$$

These combine to give

$$\frac{\tan(\theta_k)}{\tan(\theta_{k-1})} = \lambda_0\|J^{-1}w_{k-1}\|_2. \tag{5.35}$$

Since $w_{k-1}$ is perpendicular to $v_0$, we can estimate the norm using equation (5.28) and the spectral theorem for Jacobi operators [Dei00],

$$\|J^{-1}w_{k-1}\|_2 \leq \sup_{w\perp v_0, \|w\|_2=1} \|J^{-1}w\|_2 = \eta^{-1}. \tag{5.36}$$

Therefore $|\tan(\theta_k)| \leq \left|\frac{\lambda_0}{\eta}\right|^k \tan(\theta_0)$. Since for $\theta \in (-\pi/2, \pi/2)$ we have $|\theta| \leq |\tan(\theta)|$, this implies $\theta_k = \mathcal{O}\left(\left|\frac{\lambda_0}{\eta}\right|^k\right)$.

To complete the proof, we calculate,

$$\begin{aligned}\|u_k - v_0\|_2^2 &= \|u_k\|_2^2 + \|v_0\|_2^2 - 2\langle v_0, u_k\rangle \\ &= 2(1 - \cos(\theta_k)) \\ &= \mathcal{O}(\theta_k^2).\end{aligned}$$

Taking the square root completes the proof. □

**Theorem 5.3.4** (Convergence of QL algorithm)**.** *Let $J$ be a bounded Jacobi operator with an eigenpair $(\lambda_0, v_0)$ satisfying the conditions of Lemma 5.3.2. Let $J^{(0)}, J^{(1)}, J^{(2)}, \ldots$ be the Jacobi operators generated by the (unshifted) QL algorithm starting at $J$. Then the $(1,0)$ entry of the iterates satisfies*

$$\beta_0^{(k)} = \mathcal{O}\left(\left|\frac{\lambda_1}{\eta}\right|^k\right). \tag{5.37}$$

*Remark* 5.3.5. Note that there is no requirement on the structure of $J$, only on the eigenstructure.

*Proof.* Define the orthogonal operators $P_0, P_1, P_2, \ldots$ by

$$P_k = Q^{(0)}Q^{(1)}\cdots Q^{(k)}, \tag{5.38}$$

where $J^{(k)} = Q^{(k)}L^{(k)}$ is the QL factorisation. Then $J^{(k)} = P_{k-1}^T J P_{k-1}$ and $u_k = P_{k-1}e_0$ where $u_0, u_1, u_2, \ldots$ is the inverse iteration in Lemma 5.3.2 (by convention $P_{-1} = I$). Then, combining this with the result of Lemma 5.3.2, we have

$$\begin{aligned} J^{(k)}e_0 &= P_{k-1}^T J u_k \\ &= P_{k-1}^T J\left(v_0 + \mathcal{O}\left(\left|\frac{\lambda_1}{\eta}\right|^k\right)\right) \\ &= \lambda_0 P_{k-1}^T v_0 + \mathcal{O}\left(\left|\frac{\lambda_1}{\eta}\right|^k\right). \end{aligned}$$

For this final line we have used the fact that $J$ is bounded. Now writing $v_0 = u_k + \mathcal{O}\left(\left|\frac{\lambda_1}{\eta}\right|^k\right)$, we have

$$\begin{aligned} J^{(k)}e_0 &= \lambda_0 P_{k-1}^T u_k + \mathcal{O}\left(\left|\frac{\lambda_1}{\eta}\right|^k\right) \\ &= \lambda_0 e_0 + \mathcal{O}\left(\left|\frac{\lambda_1}{\eta}\right|^k\right). \end{aligned}$$

To complete the proof, note that $\beta_0^{(k)} \leq \sqrt{\alpha_0^{(k)} + (\beta_0^{(k)})^2} = \|J^{(k)}e_0\|_2$. □

**Corollary 5.3.6.** *Let $J$ be a bounded Jacobi operator with an eigenvalue $\lambda$ such that $\operatorname{dist}(\pm\lambda, \sigma(J) \setminus \{\lambda\}) = \delta > 0$ and let $\sigma \in \mathbb{R}$ be such that $|\lambda - \sigma| = \varepsilon < \delta$. Then the*

*iterates of the shifted QL algorithm starting from $J$ with shift $\sigma$ satisfy*

$$\alpha_0^{(k)} = \lambda + \mathcal{O}\left(\left|\frac{\varepsilon}{\delta - \varepsilon}\right|^k\right), \qquad \beta_0^{(k)} = \mathcal{O}\left(\left|\frac{\varepsilon}{\delta - \varepsilon}\right|^k\right). \tag{5.39}$$

*Proof.* Apply Theorem 5.3.4 to $J - \sigma I$. Then $|\lambda_0| = |\lambda - \sigma|$ and $\eta > \delta - \varepsilon$. □

*Remark* 5.3.7. If Lemma 5.3.2 could be modified to conclude that

$$(J - \lambda_0)u_k = \mathcal{O}\left(\left|\frac{\lambda_1}{\eta}\right|^k\right), \tag{5.40}$$

then Theorem 5.3.4 could be modified to hold for unbounded Jacobi operators.

For the tridiagonal finite dimensional QL algorithm, Wilkinson shifts lead to global convergence of $\beta_0^{(0)}$ to zero, so that $\alpha_0^{(k)}$ converges to an eigenvalue [Par80]. Further, the convergence is cubically exponential in the limit, and globally quadratically exponential, which is much better than the convergence in Theorem 5.3.4. Hence it would be advantageous to use Wilkinson shifts.

The Wilkinson shift is the eigenvalue of the $2 \times 2$ principal submatrix closest to $\alpha_0$. Can we use Wilkinson shifts in an infinite dimensional QL algorithm for Jacobi operators? The answer, surprisingly, is: not always. Consider the operator

$$J = \begin{pmatrix} 0 & 1 & & & \\ 1 & 0 & \frac{1}{2} & & \\ & \frac{1}{2} & 0 & \frac{1}{2} & \\ & & \frac{1}{2} & 0 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}. \tag{5.41}$$

It is a Basic Perturbation 2 example from Chapter 4. We showed that it has discrete eigenvalues at $\pm 2/\sqrt{3}$, which we would like to see approximated in the top-left corner after some Wilkinson-shifted QL iterations. However, a Wilkinson shift for this matrix is $\sigma = 1$. This is problematic, because $J - I$ does not have a QL factorisation by Theorem 5.2.2.

### 5.3.1 Example QL iterations for Jacobi operators

Since it appears there will be difficulties for implementing Wilkinson shifts and further research is required, for now, we can use the techniques of Chapter 4 to compute

approximations to the eigenvalues, which are then used as shifts in the QL algorithm. See the following example, which we computed using the *SpectralMeasures* package (see Appendix A for code).

Let us continue the example from the end of the previous section involving a $3 \times 3$ perturbation of Toeplitz Jacobi operator. The eigenvalues computed using the techniques of Chapter 4 are $-1.173$ and $1.580$. Using $-1.173$ as a shift and performing one step of the QL algorithm gives (to 3 decimal places),

$$J^{(1)} = \begin{pmatrix} -1.173 & 0.000 & & & & \\ 0.000 & 1.116 & 0.594 & & & \\ & 0.594 & 0.342 & 0.770 & & \\ & & 0.770 & 0.156 & 0.500 & \\ & & & 0.500 & 0 & \ddots \\ & & & & \ddots & \ddots \end{pmatrix}.$$

The off-diagonal entry on the first row is not exactly zero; it is $9.159 \times 10^{16}$. We have converged to a satisfactory precision after one step. Notice that one step of the QL algorithm increases the size of the perturbation of Toeplitz by one. If we remove the first row and column and perform a shifted QL algorithm step with shift 1.580, then we find (to 3 decimal places),

$$J^{(2)} = \begin{pmatrix} 1.580 & 0.000 & & & & \\ 0.000 & 0.587 & 0.480 & & & \\ & 0.480 & -0.216 & 0.498 & & \\ & & 0.498 & 0.020 & 0.500 & \\ & & & 0.500 & 0 & \ddots \\ & & & & \ddots & \ddots \end{pmatrix}.$$

Now, we used two orthogonal lower Hessenberg matrices to perform this change, each requiring finite storage because we simply store $s_0, s_1, s_2, s_3, s$ and $c_{-1}, c_0, c_1, c_2, c_3, c$ for each Hessenberg matrix. Hence we have found an orthogonal matrix with upper

bandwidth 2 such that (to 3 decimal places),

$$Q^T J Q = \begin{pmatrix} -1.173 & 0.000 & & & & \\ 0.000 & 1.580 & 0.000 & & & \\ & 0.000 & 0.587 & 0.488 & & \\ & & 0.488 & -0.216 & 0.498 & \\ & & & 0.498 & 0 & \ddots \\ & & & & \ddots & \ddots \end{pmatrix}.$$

If we denote the operator we are left with, after removing the first two rows and columns, by $\tilde{J}$, then since $J$ had two eigenvalues, $\tilde{J}$ must have purely continuous spectrum. By the results of Chapter 4, we can compute the connection coefficients operator $C$, which is upper triangular, Toeplitz-plus-finite-rank, and invertible, such that

$$C^{-1}\tilde{J}C = \Delta.$$

Combining these into the operator

$$U = \begin{pmatrix} I_{2\times 2} & 0 \\ 0 & C \end{pmatrix} Q^T, \tag{5.42}$$

we have

$$UJU^{-1} = \begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \Delta \end{pmatrix}, \tag{5.43}$$

where $\Delta$ is the free Jacobi operator. This constitutes a canonical form for Toeplitz-plus-finite-rank Jacobi operators. In the next section we use this to describe a functional calculus, made possible by the fact that functions of diagonal matrices and Toeplitz operators are easily computed.

## 5.4 Computing functions of operators

Using the normal form derived in the previous section, a function $g : \sigma(J) \to \mathbb{R}$ of a Toeplitz-plus-finite-rank Jacobi operator $J = \Delta + F$ can be written in the form,

$$g(J) = U^{-1} \begin{pmatrix} g(\lambda_1) & & & & \\ & \ddots & & & \\ & & g(\lambda_d) & & \\ & & & g(\Delta) & \end{pmatrix} U, \tag{5.44}$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $J$.

Hence computation of $g(J)$ is reduced to that of computing $g$ for the discrete spectra, and $g(\Delta)$. The following proposition shows that if we can find a Chebyshev polynomial approximation for $g$ then $g(\Delta)$ has a very simple form.

**Proposition 5.4.1.** *Let* $g(s) = \sum_{k=0}^{m} a_k T_k(s)$. *Then*

$$g(\Delta) = \frac{1}{2} \begin{pmatrix} 2a_0 & a_1 & a_2 & a_3 & \cdots \\ a_1 & 2a_0 & a_1 & a_2 & \ddots \\ a_2 & a_1 & 2a_0 & a_1 & \ddots \\ a_3 & a_2 & a_1 & 2a_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} - \frac{1}{2} \begin{pmatrix} a_2 & a_3 & a_4 & a_5 & \cdots \\ a_3 & a_4 & a_5 & a_6 & \cdots \\ a_4 & a_5 & a_6 & a_7 & \cdots \\ a_5 & a_6 & a_7 & a_8 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{5.45}$$

*Remark* 5.4.2. Compare this formula to the multiplication formula in Section 2.2 of [OT13].

*Proof.* We prove this for all polynomials of the form $g(s) = T_k(s)$ by induction on $k$. The full result follows by linearity. Note that the Chebyshev polynomials $T_k$ satisfy

$$\begin{aligned} T_0(s) = 1, \quad T_1(s) = s \\ T_{k+1}(s) = 2sT_k(s) - T_{k-1}(s) \text{ for } k = 1, 2, \ldots. \end{aligned}$$

The result is clearly true for $T_0$ and $T_1$. Now assume $k \geq 1$ and the inductive hypothesis that the result holds for $T_k$ and $T_{k-1}$.

Recall the notation from Subsection 1.1.2 the notation for a Toeplitz and Hankel operators in terms of the symbol $f(z) = \sum_{k=-\infty}^{\infty} t_k z^k$. Then our inductive hypothesis is $T_k(\Delta) = T(\frac{1}{2}(z^k + z^{-k})) - H(\frac{1}{2}z^{1-k})$. Then we have, using the multiplicative formula

for Toeplitz and Hankel operators (see equation (1.24)),

$$\begin{aligned}
T_{k+1}(\Delta) &= 2\Delta T_k(\Delta) - T_{k-1}(\Delta) \\
&= \frac{1}{2}T(z+z^{-1})T(z^k+z^{-k}) - \frac{1}{2}T(z+z^{-1})H(z^{1-k}) \\
&\quad - \frac{1}{2}T(z^{k-1}+z^{1-k}) + \frac{1}{2}H(z^{2-k}) \\
&= \frac{1}{2}T(z^{k+1}+z^{k-1}+z^{1-k}+z^{-k-1}) - \frac{1}{2}H(z+z^{-1})H(z^k+z^{-k}) \\
&\quad - \frac{1}{2}H(z^{2-k}+z^{-k}) + \frac{1}{2}H(z+z^{-1})T(z^{k-1}) \\
&\quad - \frac{1}{2}T(z^{k-1}+z^{1-k}) + \frac{1}{2}H(z^{2-k}) \\
&= \frac{1}{2}T(z^{k+1}+z^{-k-1}) - \frac{1}{2}H(z^{-k}) + \frac{1}{2}H(z^{-1})(T(z^{k-1}) - H(z^{-k})) \\
&= \frac{1}{2}T(z^{k+1}+z^{-k-1}) - \frac{1}{2}H(z^{-k}),
\end{aligned}$$

where the last line follows because top row of $T(z^{k-1}) - H(z^{-k})$ is equal to zero. This completes the proof. □

An explanation of how the functional calculus is implemented in *SpectralMeasures* is contained at the end of Appendix A.

### 5.4.1 Discrete Schrödinger equation

As an example application, consider a discrete Schrödinger equation on the half line,

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} + iJu(t) = 0, \quad u(0) \in \ell^2, \tag{5.46}$$

where $J = -\Delta + I + \mathrm{diag}(V)$ for $V \in \ell^*_S$, represents the potential in the quantum system. The solution is

$$u(t) = \exp(-iJt)u(0). \tag{5.47}$$

Therefore the problem amounts to computing the function $g(z) = \exp(izt)$ of the operator $J$ applied to the vector $u(0)$. We consider two example Schrödinger equations designed to demonstrate the quantum tunnelling effect. Let us explain them by showing the *SpectralMeasures* code that will generate them (see Appendix A).

```
# Initial condition is skinny Gaussian
u0 = pad!(exp(-(-28:.75:28).^2),100)
```

```
# discrete Laplacian
D = free\_jacobi\_operator() - I

# small barrier potential
Vs = [zeros(40),ones(3)]
# discrete Schrodinger with potential Vs
Js = -D + SymTriToeplitz(Vs,[0.])
xs,Us = eig(Js)

# An example computation for t = 10:
t = 10
us =  Us \ (exp(-im*xs*10)*(Us*u0))

# big barrier potential
Vb = [zeros(40),2*ones(5)]
# discrete Schr\"odinger with potential Vb
Jb = -D + SymTriToeplitz(Vb,[0.])
xb,Ub = eig(Jb)

# An example computation for t = 10:
t = 10
ub =  Ub \ (exp(-im*xb*t)*(Ub*u0))
```

In Figure 5.1 we demonstrate the discrete Schrödinger equation with potential `Vs` from the above Julia code. The physical interpretation is that at time $t = 0$ there is a quantum particle to the left of a barrier. The wave function of the particle then spreads out according to the Schrödinger equation, and although there is a large barrier which a classical particle would not be able to pass, some of the wave function of the quantum particle passes through, indicating that there is some probability of measuring the particle to the right of the barrier at a later time, say $t = 30$.

In Figure 5.2 we demonstrate the discrete Schrödinger equation with potential `Vb` from the above Julia code. This potential is taller and wider than `Vs`, and as such demonstrates that although quantum tunnelling is a possibility, if the barrier is sufficiently wide and tall, then the particle is blocked. The impedance on the wave function entering the barrier is exponential, as seen in both figures, but more pronounced in Figure 5.2, because of the thicker and taller barrier.

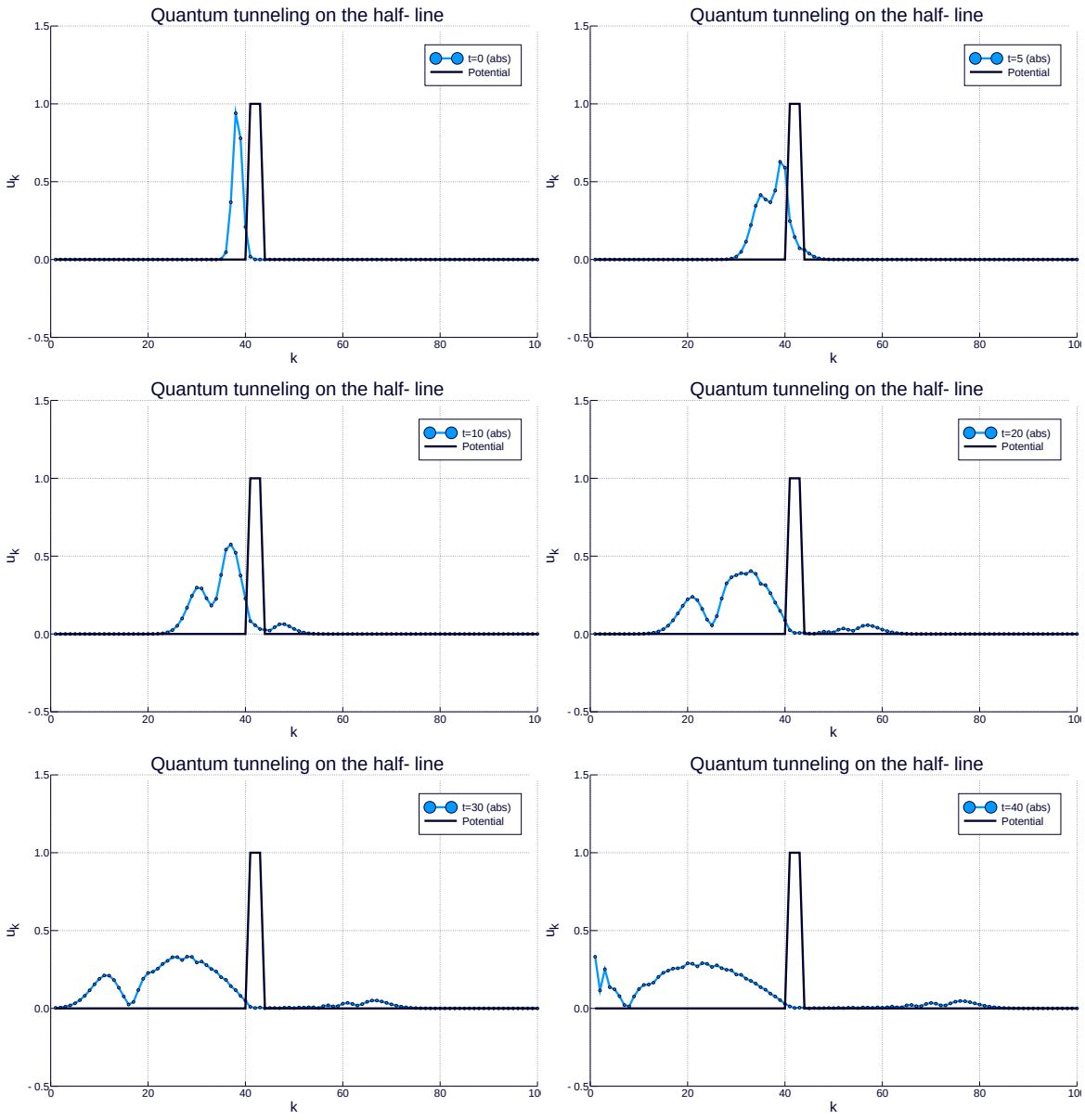



Fig. 5.1 Plotted here is a numerical solution to the discrete Schrödinger equation (5.46), with a potential function (the black line) expressible in Julia code as `Vs = [zeros(40),ones(3)]`. The functional calculus in *SpectralMeasures* was used. Plotted as blue lines is the absolute value of the solution at times $t = 0, 5, 10, 20, 30, 40$ in the range $k = 1, 2, \ldots, 100$. Hence going from left to right then top to bottom steps through a chronological sequence of times for the solution. The physical effect on display here is quantum tunnelling, where a quantum particle starts to the left of a barrier and after colliding with the barrier, some of its wave function is reflected and some transmitted through, a phenomenon which cannot happen in classical mechanics.

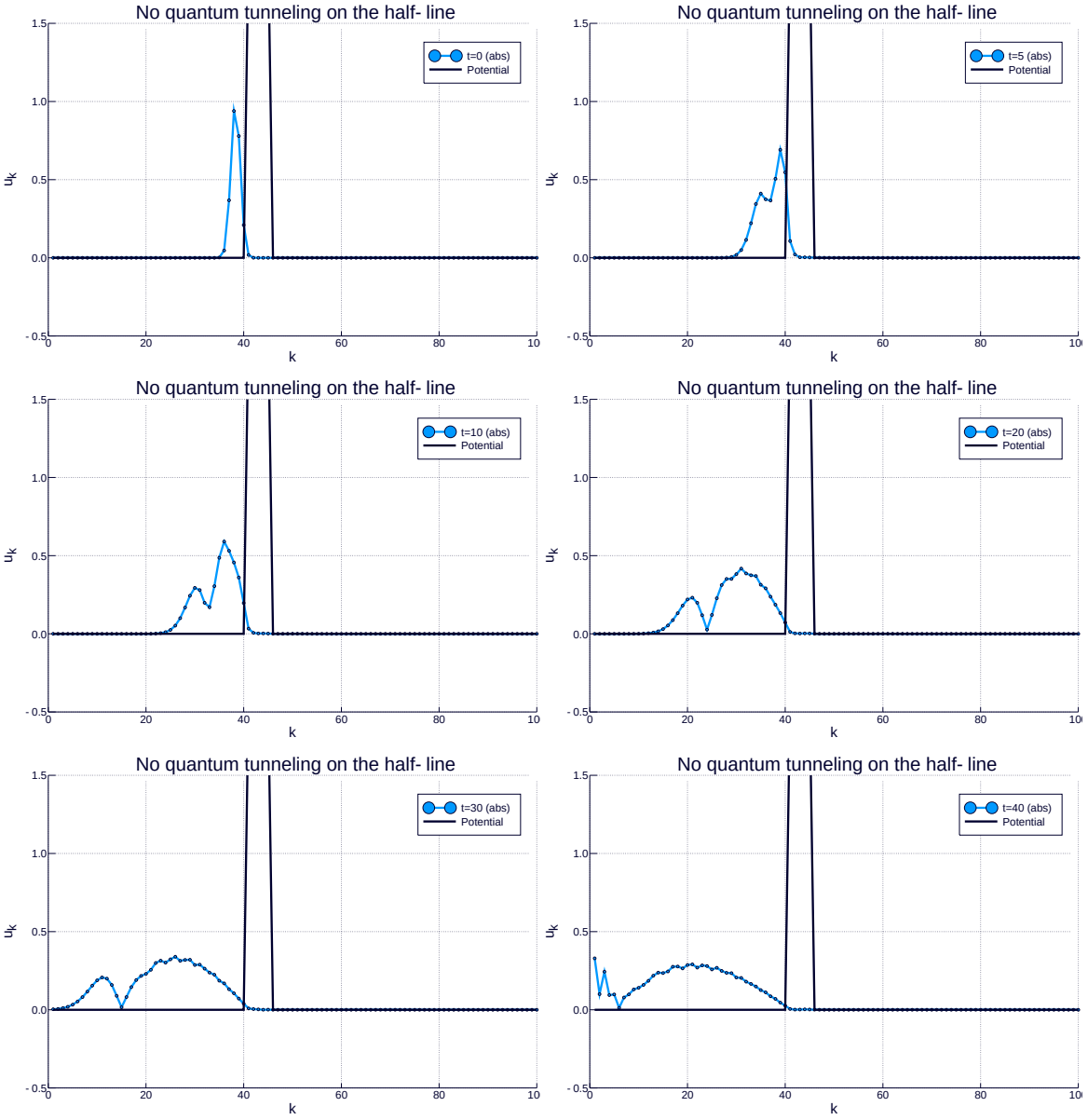



Fig. 5.2 Plotted here is a numerical solution to the discrete Schrödinger equation (5.46), with a potential function (the black line) expressible in Julia code as `Vb = [zeros(40),2*ones(5)]`. The functional calculus in *SpectralMeasures* was used. Plotted as blue lines is the absolute value of the solution at times $t = 0, 5, 10, 20, 30, 40$ in the range $k = 1, 2, \ldots, 100$. Hence going from left to right then top to bottom steps through a chronological sequence of times for the solution. The physical effect on display here is quantum tunnelling. Comparing this situation to that in Figure 5.1, the potential function is much taller and wider, so the quantum particle is completely reflected upon collision.

### 5.4.2 Discrete diffusion equation

Now consider the discrete diffusion equation on the half line,

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} + Ju(t) = 0, \quad u(0) \in \ell^2, \tag{5.48}$$

where $J = I - \Delta + F$, a Toeplitz-plus-finite-rank Jacobi operator. The perturbation $F$ represents a reaction term, in which the diffusing quantity is either used up or created at a rate proportional to its value. The solution is

$$u(t) = \exp(-Jt)u(0). \tag{5.49}$$

Hence, just as with the discrete Schrödinger equation, we can compute the solution by computing the exponential of an operator.

In this Subsection we will demonstrate two discrete diffusion equations. The first is the case where $F = 0$, solutions for which appear in Figure 5.3 in the top two plots. We begin with the initial datum $u(0) \in \ell^2$ which is zero for all indices except at $k = 21, 22, \ldots, 25$ where it is equal to 1 and at $k = 46, 47, 48$ where it is equal to 1.2. This is the blue line in each of the figures.

In the top-left plot of Figure 5.3 we solve the discrete diffusion equation using the functional calculus for Toeplitz-plus-finite-rank Jacobi operators in *SpectralMeasures*, although for this first case the finite rank part is equal to zero. In the top-right plot we solve the discrete diffusion equation using a finite section method with dimension 70. Notice that the for small $t$ the actual solution (given in the top-left) is mainly supported within $k = 1, \ldots 70$, but for sufficiently large $t$, the finite section method will lose some of the solution as it spreads out farther than this range. This is one advantage of working directly with the Jacobi operator itself, and not with truncations.

In the bottom plot of Figure 5.3, we demonstrate the solution with the same initial condition, but with a different discrete diffusion operator. This time the operator has a reaction term which exponentially depletes parts of the solution. Explicitly,

$$J = \begin{pmatrix} 1+f_0 & -\frac{1}{2} & & & \\ -\frac{1}{2} & 1+f_1 & -\frac{1}{2} & & \\ & -\frac{1}{2} & 1+f_2 & -\frac{1}{2} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{5.50}$$

where $f_k = 0$ for all $k$ except $k = 36, 37, 38$ where it is equal to 1. The solution loses significant mass around the points $k = 36, 37, 38$ compared to the pure diffusion.

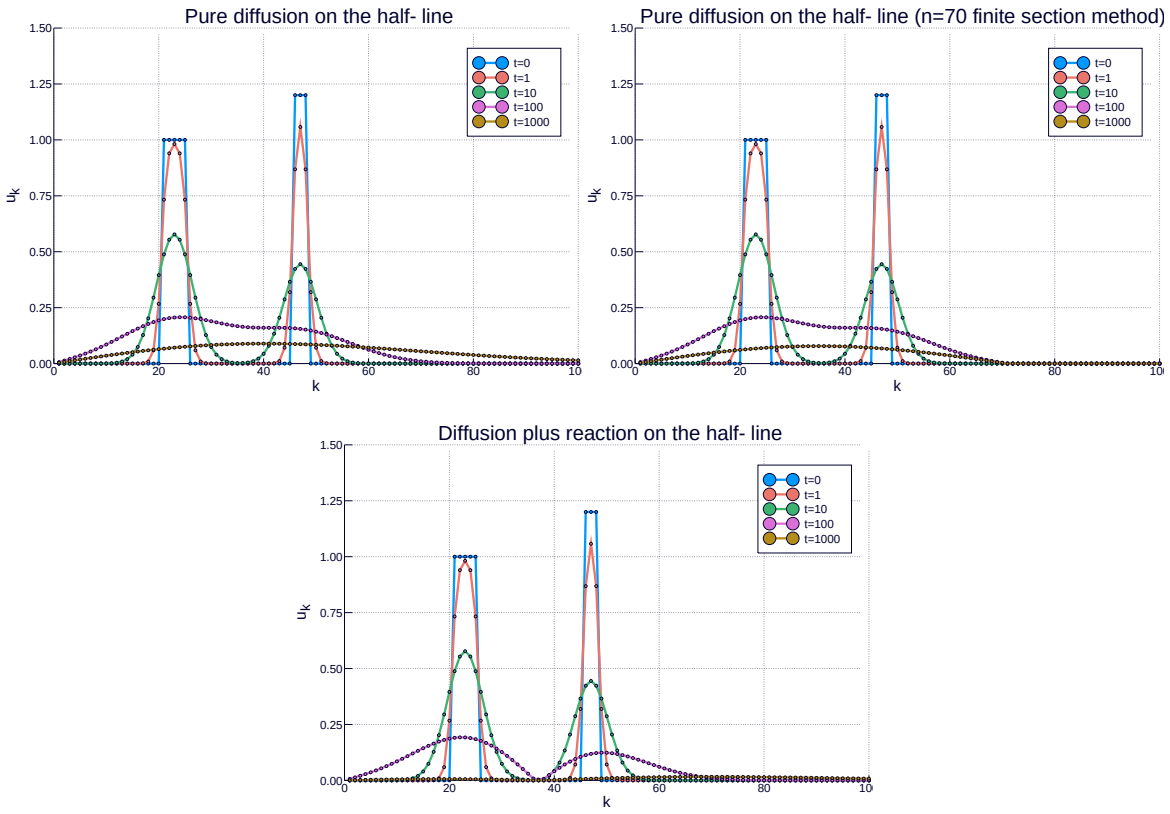



Fig. 5.3 Plotted here are numerical solutions to discrete diffusion equations of the form in equation (5.48). We use the same initial data for each: $u(0) \in \ell^2$ which is zero for all indices except at $k = 21, 22, \ldots, 25$ where it is equal to 1 and at $k = 46, 47, 48$ where it is equal to 1.2. This is the blue line. Then for times $t = 1, 10, 100, 1000$ we plot the solution in red, green, purple and yellow, as in the legends. For the top two plots, the Jacobi operator, $J$ is the discrete Laplacian, and in the bottom plot it is that in (5.50), which causes the solution to be depleted around $k = 37$. The top-left and bottom plots are solved using functional calculus in *SpectralMeasures*, and the top-right plot is solved using a finite section method with dimension $n = 70$. We can see that for large times the truncation is eventually not big enough to be accurate compared to the infinite dimensional approach.

### 5.4.3 Discrete fractional diffusion equations

We can also consider *fractional* diffusion equations. Fractional diffusion equations come from diffusive processes in which the underlying stochastic process is a heavy-tailed Lévy flight, as opposed to a Brownian motion, whose second moments are finite. Let $\alpha \in (0, 1)$. Then the discrete (space-)fractional diffusion equation on the half line is

$$\frac{\mathrm{d}u(t)}{\mathrm{d}t} + |J|^{\alpha} u(t) = 0, \quad u(0) \in \ell^2, \tag{5.51}$$

where $J = I - \Delta + F$, a Toeplitz-plus-finite-rank Jacobi operator. The solution is

$$u(t) = \exp(-|J|^\alpha t)u(0). \tag{5.52}$$

Hence the solution can be computed using functional calculus on $J$. In Figure 5.4, in the top and left plots we demonstrate the purely diffusive fractional diffusion equation (i.e. $F = 0$) for $\alpha = 0.25, 0.65, 0.85$. As before, we use the initial datum $u(0) \in \ell^2$ which is zero for all indices except at $k = 21, 22, \ldots, 25$ where it is equal to 1 and at $k = 46, 47, 48$ where it is equal to 1.2. This is the blue line in each of the figures. We see that despite having the same initial conditions, the shape of the resulting diffusion is affected severely by the choice of $\alpha$. In the bottom-right plot of Figure 5.4, $\alpha = 0.85$ and there is a reaction term added, so that the operator is of the form

$$J = \begin{pmatrix} 1+f_0 & -\frac{1}{2} & & & \\ -\frac{1}{2} & 1+f_1 & -\frac{1}{2} & & \\ & -\frac{1}{2} & 1+f_2 & -\frac{1}{2} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{5.53}$$

where $f_k$ is equal to zero everywhere, except at $k = 21, 22, 23$, where it is equal to 0.1. The effect is quite small, but this positive perturbation slows down the diffusion at the points $k = 21, 22, 23$, so you can see that in the fractional diffusion with this reaction term, the value of the solution at these points is slightly higher than that of the pure diffusion term. If we were to have taken a larger perturbation that this (which has $\|f\|_\infty = 0.1$), then they could have caused discrete spectrum. In this case, rather than merely slow down the diffusion locally, it would cause blowup.

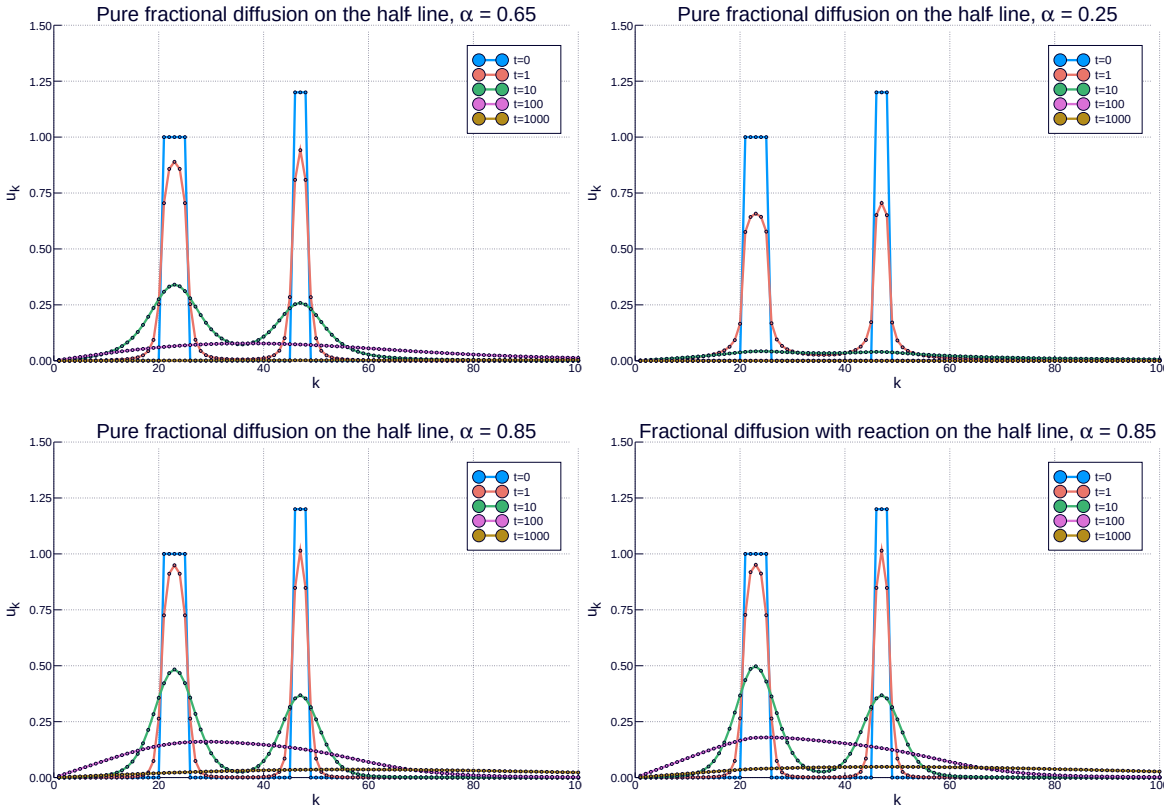



Fig. 5.4 Plotted here are numerical solutions to discrete fractional diffusion equations of the form in equation (5.51) for various values of $\alpha$ which are stated in the titles of each plot. We use the same initial data for each: $u(0) \in \ell^2$ which is zero for all indices except at $k = 21, 22, \ldots, 25$ where it is equal to 1 and at $k = 46, 47, 48$ where it is equal to 1.2. This is the blue line. Then for times $t = 1, 10, 100, 1000$ we plot the solution in red, green, purple and yellow, as in the legends. For the top two plots, the Jacobi operator, $J$ is the discrete Laplacian, and in the bottom plot it is that in (5.53). The perturbation in the bottom-right plot causes the diffusion around the points $k = 21, 22, 23$ to happen slower than in the unperturbed bottom-left plot. All plots are solved using functional calculus in *SpectralMeasures*.

画龙点睛，破壁飞去

(Huà lóng diǎn jīng, pò bì fēi qù)

Lit. Paint a dragon, dot the eyes. Breaks free and flies away.
Fig. To add the finishing touch

# Chapter 6

# Conclusion

In this chapter we discuss the accomplishments of the thesis, looking towards what can be done in the future.

## 6.1 Isospectral flows

In Chapter 2 we gave an exposition of isospectral flows, which included a discussion of isospectral gradient flows in an arbitrary metric, which was a straightforward extension of work already in the literature. We discussed the relationship between the Toda flow, the double bracket flow, the QR flow and the QR algorithm under the common framework of gradient flows with different metrics. We briefly discuss the Bloch-Iserles system, and give the new result that the KdV equation can be parametrised by an infinite dimensional Bloch-Iserles system. We introduced a new isospectral flow called the gradient QR flow, and proved that for normal matrices it coincides with the QR flow, allowing a new proof of some known convergence properties of the QR algorithm, but that for nonnormal matrix there is a perturbation term between the gradient QR flow and the QR flow, giving some insight into the lack of convergence results in the nonnormal case.

There are two directions for future research that come to mind.

First, the implications of the relationship between the KdV equation should be investigated further. Either, properties of the KdV equation that were not known before could be derived from known properties of the Bloch-Iserles system, or vice versa. It is also worth exploring how the KdV hierarchy might correspond to a Bloch-Iserles hierarchy. Alternatively, new numerical methods for the KdV equation could be derived based on numerical methods for the Bloch-Iserles equation, such as the

quadratic Magnus expansion given by Kaur [Kau16] following the work of Iserles on the double bracket equation [Ise02].

The second direction is to further investigate the gradient QR flow. One can start simply by conducting some numerical experiments comparing the gradient QR flow with the QR flow. How does the gradient QR flow behave for the examples of nonconvergence of the QR algorithm given in [Bat90], [Day96]? Is there a way to implement the gradient QR flow in a practical way for nonnormal matrices?

## 6.2 The symmetric Toeplitz inverse eigenvalue problem

In Chapter 3 we studied two isospectral flow approaches to the numerical solution of the symmetric Toeplitz inverse eigenvalue problem, a gradient flow and Chu's flow. To facilitate their study, we fully formalise how the restriction to bisymmetric matrices affects the isospectral manifold, showing that the manifold splits into a certain number of connected components, each acted upon by the Lie group of centrosymmetric orthogonal matrices $\mathrm{SCO}(n)$. The dimension of $\mathrm{SCO}(n)$ is less than $n/4$, so the number of free dimensions is reduced greatly. We utilise this to give an analytical solution for $3 \times 3$ bisymmetric isospectral flows, and conduct the first numerical study of $4 \times 4$ isospectral flows for the symmetric Toeplitz inverse eigenvalue problem. At the end of the chapter we prove a theoretical result, proving that the Solvability Complexity index of the symmetric Toeplitz inverse eigenvalue problem is equal to 1.

In Theorem 3.3.4 we showed that there exists a $3 \times 3$ real symmetric Toeplitz matrix with even eigenvalues $\lambda_1, \lambda_2$ and odd eigenvalue $\mu_1$ if and only if

$$(\lambda_1 - \mu_1)(\mu_1 - \lambda_2) + 2(\lambda_1 - \lambda_2)^2 \geq 0.$$

One could conjecture that for higher dimensions, nonnegativity of certain translation invariant, homogeneous polynomials in odd and even eigenvalues (satisfying appropriate symmetries) will be equivalent to the existence of a symmetric Toeplitz matrix with those prescribed even and odd eigenvalues, but what these polynomials would be is a mystery right now.

The following is the main conjecture in the field:

**Conjecture 6.2.1.** *Given $n$ distinct real eigenvalues, Chu's flow is convergent to a Toeplitz matrix for all initial data which are bisymmetric and whose eigenvalues alternate in parity.*

In Chapter 3 we proved this to be the case for $n = 1, 2, 3$, and gave compelling numerical evidence for the case $n = 4$. Chu's flow contains periodic orbits, so some restrictions are necessary. The claim is that bisymmetric and alternating eigenvalue parity (both of which are easily prescribed) is a sufficient restriction on the flow to guarantee convergence. This is a difficult problem because it involves analysing the behaviour of a very high-dimensional flow which evolves on a multiply-connected domain (a torus, because Lie groups of orthogonal matrices are isomorphic to tori).

We showed that the Solvability Complexity Index of the problem is equal to 1. There are two things that can be considered to take this further. First, the proof appears to generalise to other inverse eigenvalues problems, by changing the function $\Psi$ and using a different Lie algebra (instead of $\mathfrak{sco}(n)$. Second, it would be good to show that the solution could be computed with error control.

## 6.3 Spectra of Jacobi operators via connection coefficients

In Chapter 4 we proved new results about the relationship between the connection coefficients matrix between two different families of orthonormal polynomials, and the spectral theory of their associated Jacobi operators. We specialised the discussion to finite-rank perturbations of the free Jacobi operator and demonstrated explicit formulas for the principal resolvent and the spectral measure in terms of entries of the connection coefficients matrix. We showed that the results extend to trace class perturbations. Finally, we discussed computability aspects of the spectra of Toeplitz-plus-compact Jacobi operators. We showed that the spectrum of a Toeplitz-plus-compact Jacobi operator can be computed with error control, as long as the tail of the coefficients can be suitably estimated.

Regarding regularity properties of the Radon-Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ between the spectral measures $\nu$ and $\mu$ of Jacobi operators $D$ and $J$ respectively given in Propositions 4.2.5 and 4.2.8 and Corollary 4.2.9. Can much weaker regularity of $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ be shown to be equivalent to certain weak properties of $C = C_{J\to D}$? For example, we conjecture that

the Kullbeck-Leibler divergence,

$$K(\mu|\nu) = \begin{cases} \int \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(s) \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(s)\, \mathrm{d}\nu(s) & \text{if } \nu \text{ is absolutely continuous w.r.t. } \mu \\ \infty & \text{otherwise,} \end{cases} \tag{6.1}$$

is finite if and only if the function of operators, $C^T C \log(C^T C)$ is well-defined as an operator mapping $\ell_{\mathcal{F}} \to \ell^{\star}_{\mathcal{F}}$. The reasoning comes from Lemma 4.2.7. Making such statements more precise for the case where $D = \Gamma$ or $D = \Delta$ (see equation (4.59)) could give greater insight into Szégo and quasi-Szégo asymptotics (respectively) for orthogonal polynomials [GNR16], [DS06a], [KS03].

Regarding computability, is there a theorem covering ground between that covered by Theorem 4.5.7 (for Toeplitz-plus-finite-rank Jacobi operators) and Theorem 4.5.9 (for Toeplitz-plus-compact Jacobi operators)? What can be said about the convergence of the continuous part of the spectral measure of a Toeplitz-plus-finite-rank truncation $J^{[m]}$ of a Toeplitz-plus-trace-class Jacobi operator $J$? Proposition 4.4.2 implies that this convergence is at least weak in the probabilists' sense.

The computability theorems in Section 4.5 all assume real arithmetic. What can be said about floating point arithmetic? Under what situations can the computation fail to give an unambiguously accurate solution?

Chapter 4 opens some broader avenues for future research.

Lower Hessenberg operators define polynomials orthogonal with respect to Sobolev inner products [Gau04, pp. 40–43]. Therefore, we have two families of (Sobolev) orthogonal polynomials with which we may define connection coefficient matrices, as discussed in [GM09, p. 77]. Whether the connection coefficient matrices (which are still upper triangular) have structure which can be exploited for studying and computing the spectra of lower Hessenberg operators is yet to be studied.

Besides spectra of *discrete* operators defined on $\ell^2$, we conjecture that the results of this paper will also be applicable to continuous Schrödinger operators on $L^2(\mathbb{R})$, which are of the form $L_V[\phi](x) = -\phi''(x) + V(x)\phi(x)$ for a potential function $V : \mathbb{R} \to \mathbb{R}$. The reference operator is the negative Laplacian $L_0$ (which is the "free" Schrödinger operator). In this scenario, whereas the entries of a discrete connection coefficient matrix satisfy a discrete second order recurrence relation on $\mathbb{N}_0^2$ (see Lemma 4.2.1), the continuous analogue of the connection coefficient operator $C_{L_V \to L_0}$ is an integral operator whose (distributional) kernel satisfies a second order PDE on $\mathbb{R}^2$. This conjecture will be explored by the present authors in the future.

Spectra of banded self-adjoint operators may be accessible with these types of techniques too. Either using connection coefficient matrices between matrix orthogonal polynomials [DPS08], or developing tridiagonalisation techniques are possible approaches (see [Han08]), but the authors also consider this nothing more than conjecture at present. The multiplicity of the spectrum for operators with bandwidth greater than 1 appears to be a major challenge here.

## 6.4 Infinite dimensional QL algorithm

In Chapter 5 we laid the foundations for the infinite dimensional QL algorithm. We proved the basic existence and nonexistence theorems for infinite dimensional QL factorisations of bounded self-adjoint operators. We showed that for Jacobi operators the existence and uniqueness is completely determined by whether 0 was in the essential spectrum. We implemented a shifted QL algorithm for Toeplitz-plus-finite-rank Jacobi operators in Julia, proved a convergence result for shifts which approximate an eigenvalue sufficiently well, and concluded by computing functions of these operators to solve some infinite dimensional ODEs.

The nonexistence of QL factorisations of selfadjoint operators whose essential spectrum contains 0 was a completely surprising outcome, because the QR factorisation had been proven to exist for all bounded operators by Hansen [Han08], [Han09]. It would be interesting to see what can be said in the non-selfadjoint case. Also, what can be said for unbounded operators? All of this should then be done in the complex domain, rather than the real as we have done here.

This nonexistence of QL factorisation also causes a small problem. It is possible that when a shift is taken, that the resulting operator now has 0 in its essential spectrum. We gave an example of this occurring in equation (5.41) for the Wilkinson shift. From this follows the question: Can a new shift strategy be formulated which guarantees that the resulting operator has a QL factorisation, whilst also accelerating the convergence of the top-left entry to an eigenvalue?

We partially solve the problem posed by Olver and Townsend in [OT14] about how to implement an infinite dimensional QL algorithm, and provide a framework for computing the QL factorisation of bounded, banded operators of the form

$$A = \begin{pmatrix} A_n & B \\ C & A_\infty \end{pmatrix}, \tag{6.2}$$

where $A_n \in \mathbb{R}^{n\times n}$ and $A_\infty$ has an *a priori* known QL factorisation $A_\infty = Q_\infty L_\infty$. This begs the question, which operators do we have *a priori* known QL factorisations? This is a completely new question. At present we have only given an analytical QL factorisation for Toeplitz Jacobi operator, $2t_1\Delta + t_0 I$. Using this in the block form it is straightforward to write algorithms to compute the QL factorisation for any banded finite rank perturbation of this Toeplitz Jacobi operator.

For a Jacobi operator $J$, Lemma 5.2.5 gives a nonlinear recurrence that the orthogonal Hessenberg operator $Q$, such that $Q^T J$ is lower triangular, must satisfy. The $c_k$ and $s_k$ parameters in the orthogonal Hessenberg operator (see Lemma 5.2.4) must satisfy

$$c_k = \operatorname{sign}(\alpha_{k+1}) \frac{c_{k+1}\alpha_{k+1} + s_{k+1}c_{k+2}\beta_{k+1}}{\sqrt{\left(c_{k+1}\alpha_{k+1} + s_{k+1}c_{k+2}\beta_{k+1}\right)^2 + \left(\beta_k\right)^2}}$$
$$s_k = -\operatorname{sign}(\alpha_{k+1}) \frac{\beta_k}{\sqrt{\left(c_{k+1}\alpha_{k+1} + s_{k+1}c_{k+2}\beta_{k+1}\right)^2 + \left(\beta_k\right)^2}},$$

where $\alpha_0, \alpha_1, \ldots,$ and $\beta_0, \beta_1, \ldots$ are the three term recurrence parameters in $J$. Future research will involve trying to solve this backward recurrence for differently structured Jacobi operators. One example structure to consider is Jacobi operators with periodic structure i.e. there exists an integer $p$ such that $\alpha_{k+p} = \alpha_k$,
$\beta_{k+p} = \beta_k$ for all $k$. When $p = 0$, we have a Toeplitz Jacobi operator, which we have already considered. If $p = 2$ we have

$$J_p = \begin{pmatrix} a & x & & & \\ x & b & y & & \\ & y & a & x & \\ & & x & b & \ddots \\ & & & \ddots & \ddots \end{pmatrix}. \tag{6.3}$$

See [DKS10] and [Tes00] discussions of periodic Jacobi operators. These operators have essential spectrum consisting of at most $p$ disjoint intervals. Spectral pollution can be a real problem if one uses the finite section method to compute the spectrum of such operators (with so-called spectral gaps) [LS04], [DP04], so infinite dimensional techniques could be very useful here.

Another type of structured Jacobi operator to consider is those whose entries grow at a defined polynomial rate. For example, the Jacobi operator for the Hermite polynomials on the real line (whose spectral measure is the Gaussian), and the Laguerre polynomials on the half-line is (whose spectral measure is $e^{-x}\,\mathrm{d}x$) are

$$J_H = \begin{pmatrix} 0 & 1 & & & \\ 1 & 0 & \sqrt{2} & & \\ & \sqrt{2} & 0 & \sqrt{3} & \\ & & \sqrt{3} & 0 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}, \qquad J_L = \begin{pmatrix} 1 & 1 & & & \\ 1 & 3 & 2 & & \\ & 2 & 5 & 3 & \\ & & 3 & 7 & \ddots \\ & & & \ddots & \ddots \end{pmatrix} \tag{6.4}$$

It would be very interesting to be able to work with these operators as they are unbounded.

Tridiagonalisation techniques might be possible to reduce a banded self-adjoint operator to a tridiagonal one, hopefully with a structured tail like one of those in the table above [Han08].

Computing an analytical QL factorisation of a general Toeplitz operator $T(f)$ where $f \in L^2(\mathbb{T})$ could be performed directly without tridiagonalisation in the following fashion. Assume $f$ is is nonzero *inside* the unit disc. Then we may take the logarithm of the symbol to find a function $g(z) = \sum_{k=-b}^{b} g_k z^k$ such that $\exp(g(z))) = f(z)$. Then we can define $g(z) = a(z) + b(z)$ where the coefficients of $a$ and $b$ are

$$a_k = \begin{cases} g_k & \text{if } k > 0 \\ -\overline{g_{-k}} & \text{if } k \leq 0 \end{cases} \qquad b_k = \begin{cases} 0 & \text{if } k > 0 \\ g_k + \overline{g_{-k}} & \text{if } k \leq 0 \end{cases} \tag{6.5}$$

Then define $q(z) = \exp(a(z)))$ and $l(z) = \exp(b(z))$, so that $f(z) = q(z)l(z)$. We have constructed $q$ so that $q(z)\overline{q(\overline{z^{-1}})} = 1$ and $l$ so that $T(l)$ is lower triangular. The tail of $T(q)$ is the tail of an orthogonal operator because of this relation, but it is not an orthogonal operator itself, because

$$\begin{aligned} T(q)T(q)^H &= T(q)T(\overline{q(\overline{z^{-1}})}) && (6.6)\\ &= T(1) - H(q)H(E\overline{q(\overline{z^{-1}})}) && (6.7)\\ &= I - H(q)H(E\overline{q(\overline{z^{-1}})}), && (6.8) \end{aligned}$$

using the multiplication rules for Toeplitz matrices (see equation (1.24)). If we can characterise a compact operator $K_q$ such that $Q = T(q)(I + K_q)$ is orthogonal, then we can compute a lower triangular and compact operator $K_l$ such that if we set $L = T(l) + K_l$ then $T(f) = QL$ by taking $T_l = Q^H T(f) - T(l)$. This approach appears to be related to the Wiener-Hopf factorisation of $f$, which is the LU factorisation of the symbol, instead of the QL factorisation we are describing here. Further research in this direction is to work out how to compute $K_q$ for general bandwidths $b > 1$. This appears a little challenging, but a possible avenue for the computation.

The results in Section 5.3 on the convergence of the QL algorithm for Jacobi operators are very encouraging. It deserves a little investigation into whether the result can be extended to unbounded Jacobi operators, as indicated in Remark 5.3.7. The main result only concerned a constant shift which was sufficiently close to an isolated eigenvalue, which yielded linear convergence of the top-left entry to the eigenvalue. However, in finite dimensions, Rayleigh shifts and Wilkinson shifts can yield quadratic and locally cubic convergence rates, and a mixed strategy can yield globally cubic rates [Wan01]. It would be good to be able to implement such strategies, but we gave an example of a Jacobi operator in which a Wilkinson shift would give an operator which doesn't give a QL factorisation. Can a good shifting strategy be formulated?

A proper numerical comparison between established approaches to computing functions of operators by finite sections, and new approach would be interesting to see. In fact, deeper study into the numerical properties of the algorithms created, such as numerical stability and efficient implementation, should be explored.

# Appendix A
# *SpectralMeasures* Julia package

Here we list the source code for the software package *SpectralMeasures*, written in Julia in a collaboration between the author and Sheehan Olver at the University of Sydney. The software implements many of the ideas from Chapters 4 and 5. Throughout the package, features from the ApproxFun package are used extensively [Olvb].

Although the source code is freely available at https://github.com/JuliaApproximation/SpectralMeasures.jl, the code is subject to constant updates and improvements. Not only that, but the Julia language itself is not yet at version 1.0, so its syntax can change between versions. Hence this appendix is merely a snapshot of some of the code as it is right now.

## A.1 Connection coefficient matrices

To compute the connection coefficients matrix $C_{J\to D}$ between Jacobi operators $J$ and $D$ (see Definition 4.0.2), we use the recurrence relation defined by the 5-point formula in Lemma 4.2.1.

`connection_coeffs_matrix` takes as input, vectors `a` and `b`, which are vectors representing the initial parts of $\alpha_0, \alpha_1, \ldots$ and $\beta_0, \beta_1, \ldots$ in $J$, and vectors `c` and `d`, which are vectors representing the initial parts of $\gamma_0, \gamma_1, \ldots$ and $\delta_0, \delta_1, \ldots$ in $D$, and an integer `N`. The output is the principal `N` $\times$ `N` submatrix of $C_{J\to D}$, assuming that the vectors `a` and `c` extend by zeros, and `b` and `d` extend by halves if they are not of length at least $N$.

```
function connection_coeffs_matrix(a,b,c,d,N)
  if N>max(length(a),length(b)+1,length(c),length(d)+1)
  a = [a;zeros(N-length(a))]; b = [b;.5+zeros(N-length(b))]
  c = [c;zeros(N-length(c))]; d = [d;.5+zeros(N-length(d))]
```

```
  end

  C = zeros(eltype(a),N,N)
  C[1,1] = 1
  C[1,2] = (c[1]-a[1])/b[1]
  C[2,2] = d[1]/b[1]
  for j = 3:N
    C[1,j] = ((c[1]-a[j-1])*C[1,j-1] + d[1]*C[2,j-1] - b[j-2]*C[1,j
        -2])/b[j-1]
    for i = 2:j-1
      C[i,j] = (d[i-1]*C[i-1,j-1] + (c[i]-a[j-1])*C[i,j-1] + d[i]*C[i
          +1,j-1] - b[j-2]*C[i,j-2])/b[j-1]
    end
    C[j,j] = d[j-1]*C[j-1,j-1]/b[j-1]
  end
  C
end
```

`connection_coeffs_operator` takes as input, vectors `a` and `b`, to be interpreted as above. The output is the connection coefficients operator $C_{J\to\Delta}$ where $\Delta$ is the free Jacobi operator. The output is an operator given by the type `PertToeplitz`, explained in the next Section. The following functions are used from the ApproxFun package:

- `bzeros` is an ApproxFun function returning a `BandedMatrix` (an ApproxFun type) containing only zeros
- `ToeplitzOperator` is an ApproxFun type representing a banded Toeplitz operator
- `FiniteOperator` is an ApproxFun type representing an operator with only finitely many nonzero entries, stored as a `BandedMatrix`

```
function connection_coeffs_operator(a,b)
  n = max(2,length(a),length(b)+1)
  N = 2*n #This is sufficient only because we go from Delta
  a = [a;zeros(N-length(a))]; b = [b;.5+zeros(N-length(b))]

  elType = eltype(a)
  ToeplitzVec = zeros(elType,N)
  K = bzeros(elType,n,N,0,N+1) # banded matrix of zeros
  K[1,1] = 1
  K[1,2] = -a[1]/b[1]
```

```
11    K[2,2] = .5/b[1]
12
13    # The recurrence for the first n+1 cols depend on a and b
14    for j = 3:n+1
15      K[1,j] = (-a[j-1]*K[1,j-1] + .5*K[2,j-1] - b[j-2]*K[1,j-2])/b[j
          -1]
16      for i = 2:j-2
17        K[i,j] = (.5*K[i-1,j-1] -a[j-1]*K[i,j-1] + .5*K[i+1,j-1] - b[j
            -2]*K[i,j-2])/b[j-1]
18      end
19      K[j-1,j] = (.5*K[j-2,j-1] -a[j-1]*K[j-1,j-1] - b[j-2]*K[j-1,j-2])
          /b[j-1]
20      if j<n+1
21        K[j,j] = .5*K[j-1,j-1]/b[j-1]
22      end
23    end
24    ToeplitzVec[1] = K[n,n]
25    ToeplitzVec[2] = K[n,n+1]
26
27    # The recurrence for rows n+2 to 2n do not depend on a and b
28    for j = n+2:N
29      K[1,j] = K[2,j-1] - K[1,j-2]
30      for i = 2:N-j
31        K[i,j] = K[i-1,j-1] + K[i+1,j-1] - K[i,j-2]
32      end
33      if j < N
34        K[N+1-j,j] = K[N-j,j-1] + K[N+2-j,j-1] - K[N+1-j,j-2]
35      end
36      ToeplitzVec[2*(j-n)-1] = K[N+1-j,j-1]
37      ToeplitzVec[2*(j-n)] = K[N+1-j,j]
38    end
39    T = ToeplitzOperator(elType[],chop!(ToeplitzVec))
40    for j = 1:N
41      for i = 1:min(j,N+1-j)
42        K[i,j]-=T[i,j]
43      end
44    end
45    T+FiniteOperator(K)
46  end
```

## A.2 Types for Toeplitz-plus-finite-rank operators

The `PertToeplitz` type simply stores a `ToeplitzOperator`, `T` and a `FiniteOperator`, `K`, and behaves as you would expect the operator `T + K` to behave. The following are used from the ApproxFun package:

- `ToeplitzOperator` is an ApproxFun type representing a banded Toeplitz operator
- `FiniteOperator` is an ApproxFun type representing an operator with only finitely many nonzero entries, stored as a `BandedMatrix`
- `bandinds` is a function in ApproxFun which returns the upper and lower bandwidths of a banded operator
- `SequenceSpace` is a `Space` of functions in ApproxFun in which functions are simply sequences indexed by the natural numbers

```
## represents T + K where T is Toeplitz and K is finite-dimensional
immutable PertToeplitz{S} <: Operator{S}
  T::ToeplitzOperator{S}
  K::FiniteOperator{BandedMatrix{S},S}
end

# Returns what the domain and range of the operator is
for OP in (:domainspace,:rangespace)
  @eval $OP(::PertToeplitz) = SequenceSpace()
end

# Returns lower and upper bandwidths of the operator
bandinds(P::PertToeplitz) = min(bandinds(P.T,1),bandinds(P.K,1)),max(
    bandinds(P.T,2),bandinds(P.K,2))

# Accessing entries of the operator
getindex(P::PertToeplitz,k::Integer,j::Integer) = P.T[k,j]+P.K[k,j]
getindex(P::PertToeplitz,k::AbstractCount,j::AbstractCount) = P.T[k,j
    ]+P.K[k,j]
```

`SymTriOperator` is implemented as a subtype of the abstract type `TridiagonalOperator` (in ApproxFun). It stores two finite vectors: `dv` is the diagonal entries and `ev` is the offdiagonal entries.

```
immutable SymTriOperator{T} <: TridiagonalOperator{T}
  dv::Vector{T}
  ev::Vector{T}
end

SymTriOperator(A::Vector,B::Vector) = SymTriOperator{promote_type(
    eltype(A),eltype(B))}(A,B)

# Returns what the domain and range of the operator is
for OP in (:domainspace,:rangespace)
  @eval $OP(::SymTriOperator) = SequenceSpace()
end

# Accessing entries of the operator
function getindex(S::SymTriOperator,k::Integer,j::Integer)
  if k <= length(S.dv) && k == j
    S.dv[k]
   elseif k <= length(S.ev) && j==k+1
    S.ev[k]
  elseif 2 <= k <= length(S.ev)+1 && j==k-1
    S.ev[k-1]
  else
    zero(eltype(S))
  end
end

# Multiplying by a constant
*(c::Number,A::SymTriOperator) = SymTriOperator(c*A.dv,c*A.ev)
*(A::SymTriOperator,c::Number) = SymTriOperator(c*A.dv,c*A.ev)

# Creates a SymTridiagonal from UnitRange's
function Base.SymTridiagonal(S::SymTriOperator,kr::UnitRange,jr::
    UnitRange)
  n=last(kr)
  @assert n==last(jr)
  SymTridiagonal(pad(S.dv,n),pad(S.ev,n-1))
end
```

`SymTriToeplitz` is implemented as a subtype of `TridiagonalOperator` (in ApproxFun). It stores a finite number of diagonal entries and offdiagonal entries in `dv` and `ev` respectively, and it also stores `a`, which is the diagonal entry for indices larger than the length of `dv`, and `b`, which is the offdiagonal entry for indices larger than the length of `ev`.

```
# Represents a SymTriOperator + Symmetric ToeplitzOperator
immutable SymTriToeplitz{T} <: TridiagonalOperator{T}
  dv::Vector{T}
  ev::Vector{T}
  a::T
  b::T

  SymTriToeplitz(dv::Vector{T},ev::Vector{T},a::T,b::T) = new(dv,ev,a
      ,b)
  SymTriToeplitz(dv::Vector,ev::Vector,a,b) = new(Vector{T}(dv),
      Vector{T}(ev),T(a),T(b))
end

# Deals with number type issues
SymTriToeplitz(dv::Vector,ev::Vector,a,b) = SymTriToeplitz{
    promote_type(eltype(dv),eltype(dv),typeof(a),typeof(b))}(dv,ev,a,b
    )

# Constructs as the sum T + K
function SymTriToeplitz(T::ToeplitzOperator,K::SymTriOperator)
  @assert bandinds(T)==(-1,1) && issym(T)
  SymTriToeplitz(K.dv+T.nonnegative[1],K.ev+T.nonnegative[2],T.
      nonnegative...)
end

# Constructs with zero limiting values
function SymTriToeplitz{TT}(K::SymTriOperator{TT})
  SymTriToeplitz(K.dv,K.ev,zero(TT),zero(TT))
end

# Converts a tridiagonal ToeplitzOperator into a SymTriToeplitz
function SymTriToeplitz(T::ToeplitzOperator)
  @assert issym(T)

  if isdiag(T)
    SymTriToeplitz(eltype(T)[],eltype(T)[],T.nonnegative[1],zero(
        eltype(T)))
  elseif bandinds(T)==(-1,1)
    SymTriToeplitz(eltype(T)[],eltype(T)[],T.nonnegative...)
  else
    error("Not a tridiagonal operator")
  end
```

```
end

# Returns what the domain and range of the operator is
for OP in (:domainspace,:rangespace)
  @eval $OP(::SymTriToeplitz) = SequenceSpace()
end

# Accessing entries of the operator
function Base.getindex(S::SymTriToeplitz,kr::UnitCount{Int},jr::
    UnitCount{Int})
  k=first(kr)
  @assert k==first(jr)

  SymTriToeplitz(S.dv[k:end],S.ev[k:end],S.a,S.b)
end

function getindex(S::SymTriToeplitz,k::Integer,j::Integer)
  if 2 <= k && j ==k-1
    k<=length(S.ev)+1?S.ev[k-1]:S.b
  elseif j==k+1
    k<=length(S.ev)?S.ev[k]:S.b
  elseif j==k
    k<=length(S.dv)?S.dv[k]:S.a
  else
    zero(eltype(S))
  end
end
```

The following implements `connection_coeffs_operator` for an arbitrary Toeplitz-plus-finite-rank Jacobi operator by scaling and shifting the entries and using the main definition of `connection_coeffs_operator`.

```
connection_coeffs_operator(J::SymTriToeplitz) =
    connection_coeffs_operator(.5*(J.dv-J.a)/J.b,.5*J.ev/J.b)
```

## A.3 Spectral Measure

To compute the spectral measure of a Toeplitz-plus-finite-rank Jacobi operator we use Theorem 4.3.14, which says that the measure has the form

$$\mathrm{d}\mu(s) = \frac{1}{p_C(s)}\frac{2}{\pi}\sqrt{1-s^2}\mathrm{d}s + \sum_{k=1}^{r} w_k \delta_{\lambda_k}(s), \tag{A.9}$$

where $p_C$ is the polynomial given by the computable formula $p_C(s) = \sum_{k=0}^{2n-1}\langle e_k, CC^T e_0\rangle U_k(s)$ and $r \leq n$. By Theorem 4.3.21, the numbers $\lambda_k$ are found by finding the distinct real roots $z_k$ of $c$ (the Toeplitz symbol of the Toeplitz part of $C$, which here is a polynomial of degree $2n-1$) in the interval $(-1,1)$. Also by Theorem 4.3.21, the weights $w_k$ can be computed using the formula

$$w_k = \frac{1}{2} z_k^{-1}(z_k - z_k^{-1})\frac{c_\mu(z_k)}{c'(z_k)}.$$

`spectralmeasure` takes two vectors `a` and `b`, which are vectors representing the initial parts of $\alpha_0, \alpha_1, \ldots$ and $\beta_0, \beta_1, \ldots$ in $J$. The output is the spectral measure of $J$ in the form of a `RatFun` type object, which is explained in the next Section. We use the following features from ApproxFun.

- `Fun` is a type used to represent a function. It stores a vector of coefficients and a space for how to interpret those coefficients
- `Taylor` is the space spanned by monomials
- `Ultraspherical(1)` is the space spanned by Chebyshev polynomials of the second kind $U_k$
- `JacobiWeight(0.5,0.5,Ultraspherical(1))` is the space of `Ultraspherical(1)` each multiplied by the Jacobi weight $\sqrt{1-x^2}$.
- `DiracSpace` is the space spanned by Dirac delta measures based at a finite set of points
- `PointSpace` is the dual space of `DiracSpace` consisting of functions only defined at a finite number of real points

```
function spectralmeasure(a,b)
  TT = promote_type(eltype(a),eltype(b))
  # Chop the a and b down
  a = chop!(a); b = 0.5+chop!(b-0.5)
  n = max(2,length(a),length(b)+1)
  a = [a;zeros(TT,n-length(a))]; b = [b;0.5+zeros(TT,n-length(b))]

  # Finds C such that J*C = C*Toeplitz([0,1/2])
  C = connection_coeffs_operator(a,b)
  c = Fun(Taylor,C.T.nonnegative)
  pC = Fun(C*(C'*[1]),Ultraspherical(1))

  # Check for discrete eigenvalues
  z = sort(real(filter!(z->abs(z)<1 && abs(imag(z)) <= 10eps(TT)  &&
     !isapprox(abs(z),1),complexroots(c))))
  if length(z) > 0
    Cmu = connection_coeffs_operator(a[2:end],b[2:end]) # Technically
        not Cmu from the paper
    cmu = Fun(Taylor,[0;Cmu.T.nonnegative]/b[1]) # this is cmu from
       the paper
    cprime = differentiate(c)
    eigs=real(map(joukowsky,z))
    weights = 0.5*(1-1./z.^2).*(real(cmu(z))./real(cprime(z)))
    p = Fun(DiracSpace(eigs),weights) + Fun(JacobiWeight(0.5,0.5,
       Ultraspherical(1)),[2/TT(pi)])
    q = Fun(PointSpace(eigs),ones(TT,length(eigs))) + pC
    mu = RatFun(p,q)
  else
    mu = RatFun(Fun(JacobiWeight(0.5,0.5,Ultraspherical(1)),[2/TT(pi)
       ]),pC)
  end
  mu
end

joukowsky(z) = .5*(z+1./z)
```

The following implements `spectralmeasure` for an arbitrary Toeplitz-plus-finite-rank Jacobi operator by scaling and shifting the entries and using the main definition above.

```
function spectralmeasure(J::SymTriToeplitz)
  mu = spectralmeasure(.5*(J.dv-J.a)/J.b,.5*J.ev/J.b)
```

```
  2*J.b*setdomain(mu,domain(mu) + J.a)
end
```

## A.4 A type for rational functions with Dirac weights

The formula for the spectral measure in equation (A.9) is represented using a type called `RatFun`. This type stores a numerator `p` and a denominator `q` which are both of type `Fun` (represents functions in ApproxFun). For the spectral measure, `p` is the weight function $\sqrt{1-x^2}$ plus a sum of Dirac delta functions, and `q` is the polynomial $p_C$ plus a sum of `PointSpace` functions with the same points as those in the Dirac deltas in the numerator and with weights all equal to 1. There are also commands for plotting a RatFun using the Plots.jl interface (https://github.com/JuliaPlots/Plots.jl), but it is not worth putting them here.

```
immutable RatFun{S1,T1,S2,T2}
  p::Fun{S1,T1}
  q::Fun{S2,T2}
end

domain(r::RatFun) = domain(r.p)

function evaluate(r::RatFun,x)
  (r.p)(x)./(r.q)(x)
end

@compat (r::RatFun)(x) = evaluate(r,x)

# Basic operations on RatFuns
for op = (:*,:.*)
  @eval $op(r1::RatFun,r2::RatFun)=RatFun($op(r1.p,r2.p),$op(r1.q,r2.
      q))
  @eval $op(r::RatFun,a::Union{Number,Fun}) = RatFun($op(r.p,a),r.q)
  @eval $op(a::Union{Number,Fun},r::RatFun) = RatFun($op(a,r.p),r.q)
end
Base.inv(r::RatFun) = RatFun(r.q,r.p)
Base.vec(r::RatFun) = RatFun.(vec(r.p),vec(r.q))
(./)(r1::RatFun,r2::RatFun)=r1.*inv(r2)
(./)(a,r::RatFun)=a.*inv(r)
(/)(r1::RatFun,r2::RatFun)=r1*inv(r2)
```

```
(/)(a,r::RatFun)=a*inv(r)
(./)(r::RatFun,a)=inv(r)*a

for op = (:+,:.+,:-,:.-)
  @eval $op(r1::RatFun,r2::RatFun) = RatFun($op((r1.p.*r2.q),(r2.p.*
      r1.q)),r1.q.*r2.q)
end

Base.convert(::Type{Fun},r::RatFun) = r.p/r.q
```

## A.5 Principal resolvent

The principal resolvent $G(\lambda)$, of a Toeplitz-plus-finite-rank Jacobi operator $J$, by Theorem 4.3.12 can be evaluated for any $\lambda \in \mathbb{C} \setminus \sigma(J)$ by the formula

$$G(\lambda) = \frac{G_\Delta(\lambda) - p_C^\mu(\lambda)}{p_C(\lambda)},$$

where $p_C$ is as above and $p_C^\mu(\lambda) = \sum_{k=0}^{2n-1} \langle e_k, C^\mu C^T e_0 \rangle U_k(\lambda)$. Both `principal_resolvent` and `disc_resolvent` take two vectors `a` and `b`, which are vectors representing the initial parts of $\alpha_0, \alpha_1, \ldots$ and $\beta_0, \beta_1, \ldots$ in $J$.

```
function principal_resolvent(a,b)
  # Compute the necessary polynomials
  C = connection_coeffs_operator(a,b)
  Cmu = connection_coeffs_operator(a[2:end],b[2:end]) # Technically
      not Cmu from the paper
  f = Fun((C*(C'*[1])),Ultraspherical(1))
  fmu = Fun(Ultraspherical(1),coefficients(Cmu*((C'*[1]).coefficients
      [2:end])/b[1]))

  # Return the resolvent
  x->(2*sqrt(complex(x-1)).*sqrt(complex(x+1))-2*x-extrapolate(fmu,x)
      )./extrapolate(f,x)
end
```

The mapped principal resolvent $G(\lambda(z))$, of a Toeplitz-plus-finite-rank Jacobi operator $J$, is the principal resolvent mapped to $z$ in the unit disc by the Joukowski map $\lambda : z \to \frac{1}{2}(z + z^{-1})$. This is computed using the simple formula from Theorem

4.3.20,

$$G(\lambda(z)) = -\frac{c_\mu(z)}{c(z)},$$

where $c$ and $c_\mu$ are the Toeplitz symbols of the Toeplitz parts of $C$ and $C^\mu$ respectively (these are polynomials of degree $2n-1$ and $2n-2$ respectively).

```
function disc_resolvent(a,b)
  # Compute the necessary polynomials
  C = SpectralMeasures.connection_coeffs_operator(a,b)
  Cmu = SpectralMeasures.connection_coeffs_operator(a[2:end],b[2:end
     ]) # Technically not Cmu from the paper
  c = Fun(Taylor,C.T.nonnegative)
  cmu = Fun(Taylor,[0;Cmu.T.nonnegative]/b[1]) # this is the cmu from
       the paper

  # Return the rational function
  x->-cmu(x)./c(x)
end
```

## A.6 Eigenvalues and spectrum

`discreteeigs` returns the discrete spectrum of a Toeplitz-plus-finite-rank Jacobi operator. This uses the ApproxFun complexroots command to compute the roots of the Toeplitz symbol of the connection coefficients operator. `validated_spectrum` uses the ValidatedNumerics package (https://github.com/dpsanders/ValidatedNumerics.jl) to compute the spectrum of $J$ with a guaranteed error estimate (completely rigorous if we assume that the connection coefficients have been computed exactly without rounding error). Both functions take two vectors `a` and `b`, which are vectors representing the initial parts of $\alpha_0, \alpha_1, \ldots$ and $\beta_0, \beta_1, \ldots$ in $J$.

```
function discreteeigs(a,b)
  a = chop!(a); b = .5+chop!(b-.5)
  n = max(2,length(a),length(b)+1)
  a = [a;zeros(n-length(a))]; b = [b;.5+zeros(n-length(b))]
  # Finds C such that C*J = Toeplitz([0,1/2])*C
  C = connection_coeffs_operator(a,b)
  Tfun = Fun(Taylor,C.T.nonnegative)
  sort(real(map(joukowsky,filter!(z->abs(z)<1 && isreal(z) && !
     isapprox(abs(z),1),complexroots(Tfun)))))
end
```

```
10
11 discreteeigs(J::SymTriToeplitz) = 2*J.b*discreteeigs(.5*(J.dv-J.a)/J.
       b,.5*J.ev/J.b) + J.a
12
13 using ValidatedNumerics, ValidatedNumerics.RootFinding
14
15 function validated_spectrum(a,b)
16   # Finds C such that J*C = C*Toeplitz([0,1/2])
17   C = SpectralMeasures.connection_coeffs_operator(a,b)
18   c = Fun(Taylor,C.T.nonnegative)
19
20   rts = find_roots(x->c(x),-1,1)
21   if length(rts) > 0
22     eigs=real(map(x->joukowsky(x.interval),rts))
23     eigserrs = map(midpoint_radius,eigs)
24     spectrum = ApproxFun.Interval(-1,1)
25     maxerr = 0
26     for (eig,err) in eigserrs
27       maxerr = max(maxerr,err)
28       spectrum = union(ApproxFun.Point(eig), spectrum)
29     end
30   else
31     spectrum = ApproxFun.Interval(-1,1)
32     maxerr = 0
33   end
34   spectrum, maxerr
35 end
```

## A.7 QL factorisation

The function `givenstail` takes as input the tail entries of a Toeplitz-plus-finite-rank Jacobi operator, `t0` and `t1`. The output is the tail parameters of $Q$ (in the Schur parametrisation of a unitary Hessenberg operator in Lemma 5.2.4), the Toeplitz part of $L$ and two entries `alph` and `beta` for the partial QL factorisation of $J$ used below.

```
1 # returns the parameters for the limiting Toeplitz
2 function givenstail(t0::Real,t1::Real)
3   @assert t0^2-4t1^2>=0
4   sinf = (t0 - sqrt(t0^2-4t1^2))/(2t1)
5   l0 = (t0 + sqrt(t0^2-4t1^2))/2
6   if sinf^2 > 1
```

```
    sinf = (t0 + sqrt(t0^2-4t1^2))/(2t1)
    l0 = (t0 - sqrt(t0^2-4t1^2))/2
  end
  cinf = -sqrt(1-sinf^2)
  alph = t1*cinf
  beta = cinf*t0 - sinf*alph
  l1 = 2t1
  l2 = t1*sinf
  cinf,sinf,ToeplitzOperator([l1,l2],[l0]),alph,beta
end
```

The function `ql` takes as input two vectors `a` and `b`, which are vectors representing the initial parts of $\alpha_0, \alpha_1, \ldots$ and $\beta_0, \beta_1, \ldots$ in $J$, and two tail entries `t0` and `t1` for $J$. The function returns an orthogonal operator `Q` which is of type `UnitaryHessenberg` (explained in the next Section) and a lower triangular operator `L` which is of type `PertToeplitz` (explained in a previous section), such that `Q` and `L` for the QL factorisation of $J$.

```
function ql(a,b,t0,t1)
  if t0^2<4t1^2
    error("A QL decomposition only exists outside the essential
        spectrum")
  end
  # The Givens rotations coming from infinity (with parameters cinf
     and sinf) leave us with the almost triangular
  # a[n-1]  b[n-1]    0    0    0
  # b[n-1]   a[n]    t1    0    0
  #   0      alph   beta    0    0
  #   0       l2     l1    l0    0
  #   0        0     l2    l1   l0

  cinf,sinf,TL,alph,beta=givenstail(t0,t1)

  if TL[1,1] < 0
    # we want positive on L diagonals
    Q,L=ql(-a,-b,-t0,-t1)
    return -Q,L
  end

  # Here we construct this matrix as L
  n = max(length(a),length(b)+1)
  J = jacobimatrix(a,b,t0,t1,n+1)
```

```
23    J[n,n+1] = t1
24    #     L[n+1,n+2] = 0
25    J[n+1,n+1]=beta
26    J[n+1,n]=alph
27    c,s,L=tridql!(J)
28
29    Q=HessenbergUnitary('L',true,c,s,cinf,-sinf)
30
31    for j=1:n+1
32      L[j,j]-=TL.nonnegative[1]
33      if j <= n
34        L[j+1,j]-=TL.negative[1]
35        if j <= n-1
36          L[j+2,j]-=TL.negative[2]
37        end
38      end
39    end
40    Q,TL+FiniteOperator(L)
41  end
42
43  # finite dimensional Jacobi matrix helper function
44  function jacobimatrix(a,b,t0,t1,N)
45    J = BandedMatrix(Float64,N,N,1,1)
46    for i = 1:min(length(a),N)
47      J[i,i] = a[i]
48    end
49    for i=length(a)+1:N
50      J[i,i] = t0
51    end
52    for i = 1:min(length(b),N-1)
53      J[i,i+1] = J[i+1,i] = b[i]
54    end
55    for i=length(b)+1:N-1
56      J[i,i+1] = J[i+1,i] = t1
57    end
58    J
59  end
60
61  ql(A::SymTriToeplitz) = ql(A.dv,A.ev,A.a,A.b)
```

## A.8 Types for banded-above unitary operators

The orthogonal operator in a QL factorisation of a Toeplitz-plus-finite-rank Jacobi operator can be stored in finite memory using the Schur parametrisation in Lemma 5.2.4. We first have an abstract type `UnitaryOperator`, and implement the basic functions that encode that its inverse is simply the transpose.

```
abstract UnitaryOperator{T} <: Operator{T}

# Basic operations for the abstract type, UnitaryOperator
Base.inv(Q::UnitaryOperator) = Q'
Base.transpose{T<:Real}(Q::UnitaryOperator{T}) = Q'
\(Q::UnitaryOperator,v::Number;opts...) = Q'*v
\(Q::UnitaryOperator,v::Array;opts...) = Q'*v
\{S,T,DD,Q}(A::UnitaryOperator,b::Fun{MatrixSpace{S,T,DD,1},Q};kwds
    ...) = Q'*b
\(Q::UnitaryOperator,v::Fun;opts...) = Fun(space(v),Q'*v.coefficients
    )
```

The type `HessenbergUnitary` implements the `UnitaryOperator` type. It stores the parameters for the Schur parametrisation in Lemma 5.2.4 in `sign`, `c`, `s`, `cinf` and `sinf`. For a Hessenberg operator, one of the bandwidths is always 1, but the other bandwidth is always infinity. However, the entries decay exponentially, so it has an approximate lower bandwidth. Being able to return upper and lower bandwidths is a requirement for operators in ApproxFun. Fast operator-vector multiplication schemes are included, which use the fact that `HessenbergUnitary` is a product of Givens rotations.

```
immutable HessenbergUnitary{uplo,T} <: UnitaryOperator{T}
  sign::Bool
  c::Vector{T}
  s::Vector{T}
  cinf::T
  sinf::T
  band::Int

  function HessenbergUnitary(sgn::Bool,c::Vector{T},s::Vector{T},cinf
      ::T,sinf::T,bnd::Int)
    @assert isapprox(sinf^2+cinf^2,1)
    @assert length(c)==length(s)+1
    @assert isapprox(abs(first(c)),1)

```

```
14      for (cc,ss) in zip(c[2:end],s)
15        @assert isapprox(cc^2+ss^2,1)
16      end
17
18      new(sgn,c,s,cinf,sinf,bnd)
19    end
20  end
21
22  function HessenbergUnitary(uplo::Char,sign,c,s,cinf,sinf,band)
23    @assert uplo=='L' || uplo=='U'
24    HessenbergUnitary{uplo,promote_type(eltype(c),eltype(s),
25    typeof(cinf),typeof(sinf))}(sign,c,s,cinf,sinf,band)
26  end
27
28  function HessenbergUnitary(uplo::Char,sign,c,s,cinf,sinf)
29    @assert isapprox(sinf^2+cinf^2,1)
30    @assert length(c)==length(s)+1
31    @assert isapprox(abs(first(c)),1)
32
33    for (cc,ss) in zip(c[2:end],s)
34      @assert isapprox(cc^2+ss^2,1)
35    end
36
37    band=0
38    n=length(s)
39
40    cur=c[1]*c[2]
41    tol=eps()
42
43    # Compute the bandwidth of the matrix
44    k=1
45    for j=1:n+2
46      while abs(cur) > tol
47        cur*=k<=n?s[k]:sinf
48        k+=1
49        band+=1
50      end
51      # increase column and row by one
52      if band==0
53        # we don't need to divide or multiply by s
54        if j<=n-1
55          cur=c[j+1]*c[j+2]
```

```
56        elseif j==n
57          cur=c[j+1]*cinf
58        else
59          cur=cinf^2
60        end
61      else
62        if j<=n-1
63          cur*=(k<=n?s[k]:sinf)*c[j+2]/(c[j]*s[j])
64        elseif j==n
65          cur*=(k<=n?s[k]:sinf)*cinf/(c[j]*s[j])
66        elseif j==n+1
67          cur*=(k<=n?s[k]:sinf)*cinf/(c[j]*sinf)
68        else
69          cur*=sinf*cinf/(cinf*sinf)
70        end
71      end
72      k+=1
73    end
74
75    HessenbergUnitary(uplo,sign,c,s,cinf,sinf,band)
76  end
77
78  # Returns conjugate transpose of the operator
79  Base.ctranspose{T<:Real}(Q::HessenbergUnitary{'L',T}) =
        HessenbergUnitary('U',Q.sign,Q.c,Q.s,Q.cinf,Q.sinf,Q.band)
80  Base.ctranspose{T<:Real}(Q::HessenbergUnitary{'U',T}) =
        HessenbergUnitary('L',Q.sign,Q.c,Q.s,Q.cinf,Q.sinf,Q.band)
81
82  # Returns the upper and lower bandwidths
83  bandinds(Q::HessenbergUnitary{'L'}) = -Q.band,1
84  bandinds(Q::HessenbergUnitary{'U'}) = -1,Q.band
85
86  # Returns what the domain and range of the operator is
87  domainspace(::HessenbergUnitary) = SequenceSpace()
88  rangespace(::HessenbergUnitary) = SequenceSpace()
89
90  # Helper functions for getindex
91  hc(c,cinf,k) = k<=length(c)?c[k]:cinf
92  hs(s,sinf,k) = k<=length(s)?s[k]:sinf
93  hc(Q::HessenbergUnitary,k) = hc(Q.c,Q.cinf,k)
94  hs(Q::HessenbergUnitary,k) = hs(Q.s,Q.sinf,k)
95
```

```
96  # Accessing entries of the operator
97  getindex(Q::HessenbergUnitary{'L'},k::Integer,j::Integer) =
        hessuni_getindex(Q.sign,Q.c,Q.s,Q.cinf,Q.sinf,j,k)
98  getindex(Q::HessenbergUnitary{'U'},k::Integer,j::Integer) =
        hessuni_getindex(Q.sign,Q.c,Q.s,Q.cinf,Q.sinf,k,j)
99  function hessuni_getindex{T}(sgn::Bool,c::AbstractVector{T},s::
        AbstractVector{T},
100   cinf::T,sinf::T,
101   k::Integer,j::Integer)
102   si=sgn?1:-1
103
104   if k>j+1
105     zero(T)
106   elseif k>=2 && j ==k-1
107     -si*hs(s,sinf,k-1)
108   else
109     col0=hc(c,cinf,k)*hc(c,cinf,k+1)
110     for p=k+1:j
111       col0*=hs(s,sinf,p-1)*hc(c,cinf,p+1)/hc(c,cinf,p)
112     end
113     si*col0
114   end
115 end
116
117 # Fast multiplication
118 function *(Q::HessenbergUnitary{'U'},v::Vector)  si=Q.sign?1:-1
119
120   ret = pad!(si*v,length(v)+1)
121   # Compute each Givens rotation starting from the right
122
123   for i = length(v):-1:1
124     ret[i],ret[i+1] = hc(Q,i+1)*ret[i] + hs(Q,i)*ret[i+1], -hs(Q,i)*
          ret[i] + hc(Q,i+1)*ret[i+1]
125   end
126   ret[1]*=hc(Q,1)
127   ret
128 end
129
130 # Fast multiplication
131 function *(Q::HessenbergUnitary{'L'},v::Vector)
132   N =  max(length(v),length(Q.s))+1
133   si=Q.sign?1:-1
```

```
134   ret = pad!(si*v,N)
135
136   # This part does the computation we are certain we have to do
137   ret[1] *= hc(Q,1)
138   for i = 1:N-1
139     ret[i],ret[i+1] = hc(Q,i+1)*ret[i] -hs(Q,i)*ret[i+1],
140     hs(Q,i)*ret[i] + hc(Q,i+1)*ret[i+1]
141   end
142
143   # After this point, ret is monotonically decreasing to zero
144   i = N
145   while abs(ret[i]) > eps()
146     push!(ret,(Q.sinf)*ret[i])
147     ret[i] *= Q.cinf
148     i += 1
149   end
150   ret
151 end
152
153 # Computes the negative of the operator
154 -{uplo}(Q::HessenbergUnitary{uplo})= HessenbergUnitary{uplo,eltype(Q)
      }(!Q.sign,Q.c,Q.s,Q.cinf,Q.sinf,Q.band)
155
156 # Removes the first row and column (deflation)
157 deflate{uplo}(Q::HessenbergUnitary{uplo})=HessenbergUnitary(uplo,Q.
      sign,[(Q.sign?1:(-1))*sign(Q.c[1]);Q.c],[0;Q.s],Q.cinf,Q.sinf,Q.
      band)
158 deflate(Q::HessenbergUnitary,k::Integer)=k==0?Q:deflate(deflate(Q),k
      -1)
```

The type `BandedUnitary` implements the `UnitaryOperator` type. It stores a vector of `HessenbergUnitary` operators and it behaves as though it were the product of these operators by lazy evaluation.

```
1 immutable BandedUnitary{uplo,T} <: UnitaryOperator{T}
2   ops::Vector{HessenbergUnitary{uplo,T}}
3 end
4
5 Base.ctranspose(Q::BandedUnitary)=BandedUnitary(reverse!(map(
      ctranspose,Q.ops)))
6 getindex(Q::BandedUnitary,k::Integer,j::Integer)=TimesOperator(Q.ops)
      [k,j]
```

```
bandinds(Q::BandedUnitary)=bandinds(TimesOperator(Q.ops))
domainspace(::BandedUnitary) = SequenceSpace()
rangespace(::BandedUnitary) = SequenceSpace()

function *(Q::BandedUnitary,v::Vector)
  ret=v
  for k=length(Q.ops):-1:1
    ret=Q.ops[k]*ret
  end
  Fun(rangespace(Q),ret)
end
```

## A.9 QL iterations

All that is required to compute a QL iteration from a QL factorisation is to implement the multiplication `L*Q`. The following implements this for the QL factorisation of a Toeplitz-plus-finite-rank Jacobi operator.

Most of the code is checking that the inputs will actually produce a symmetric tridiagonal operator. If these tests fail then we simply use a `TimesOperator` type, which is a type in ApproxFun that stores two or more banded operators and behaves as if it were the product of those two operators by lazy evaluation.

```
function *(L::PertToeplitz,Q::HessenbergUnitary{'L'})
  n=max(size(L.K.matrix,1),length(Q.s)+3)

  if bandinds(L)==(-2,0)
  # We check if L*Q is tridiagin al
    tol=1E-14*(maximum(L.T)+maximum(L.K))
    istri=true
    for k=3:n
      if abs(L[k,k-2]*hc(Q,k-1)+L[k,k-1]*hs(Q,k-2)*hc(Q,k)+L[k,k]*hs(
          Q,k-2)*hs(Q,k-1)*hc(Q,k+1))>tol
        istri=false
        break
      end
    end
    if istri
      issym=true
      if !isapprox(-L[1,1]*hs(Q,1),L[2,1]*hc(Q,1)*hc(Q,2)+L[2,2]*hc(Q
          ,1)*hc(Q,3)*hs(Q,1);atol=tol)
```

```
17          issym=false
18        end
19
20        if issym
21          for k=2:n+1  # kth row
22            if !isapprox(-L[k+1,k-1]*hs(Q,k-1)+L[k+1,k]*hc(Q,k)*hc(Q,k
                  +1)+L[k+1,k+1]*hc(Q,k)*hc(Q,k+2)*hs(Q,k),-L[k,k]*hs(Q,k)
                  ;atol=tol)
23              issym=false
24              break
25            end
26          end
27        end
28
29        if issym
30          # result is SymTriToeplitxz
31          ev=Array(Float64,max(min(size(L.K.matrix,1),size(L.K.matrix
                ,2)),
32          length(Q.s)))
33          for k=1:length(ev)
34            ev[k]=-L[k,k]*hs(Q,k)
35          end
36
37          dv=Array(Float64,max(length(Q.s)+1,size(L.K.matrix,1)))
38          dv[1]=hc(Q,1)*hc(Q,2)*L[1,1]
39          for k=2:length(dv)
40            dv[k]=-hs(Q,k-1)*L[k,k-1]+hc(Q,k)*hc(Q,k+1)*L[k,k]
41          end
42
43          t1=-L.T[1,1]*Q.sinf
44          t0=-Q.sinf*L.T[2,1]+Q.cinf^2*L.T[1,1]
45
46          si=Q.sign?1:-1
47          return SymTriToeplitz(si*dv,si*ev,si*t0,si*t1)
48        end
49      end
50    end
51
52    # default constructor
53    TimesOperator(L,Q)
54  end
```

## A.10 Functions of operators

The *SpectralMeasures* function `eig` is intended to act in as similar a way as possible to the `eig` command in Matlab, but instead of taking a matrix, we now take a Toeplitz-plus-finite-rank Jacobi operator. The syntax in Matlab is demonstrated the following example.

```
>> A = rand(3); A = A+A'

A =
    1.9298    1.1148    1.1125
    1.1148    0.9708    1.2220
    1.1125    1.2220    1.8315

>> [V,D] = eig(A)

V =
   -0.2571   -0.7428    0.6182
    0.8658    0.1072    0.4889
   -0.4294    0.6608    0.6156

D =
    0.0336         0         0
         0    0.7792         0
         0         0    3.9192
```

The following is the analogous example for a Toeplitz-plus-finite-rank Jacobi operator in *SpectralMeasures*, with `A` becoming `J`, `D` becoming `x`, and `V` becoming `U`.

```
julia> using ApproxFun, SpectralMeasures

julia> J = SymTriToeplitz(5*rand(3),3*rand(2),0.0,0.5)
SymTriToeplitz:ApproxFun.SequenceSpace()->ApproxFun.SequenceSpace()
  0.9631  0.2743
  0.2743  3.4060  0.5426
          0.5426  1.6881  0.5
                  0.5     0.0  0.5
                          0.5  0.0  0.5
                               0.5  0.0  0.5   ...
                                         ...   ...

```

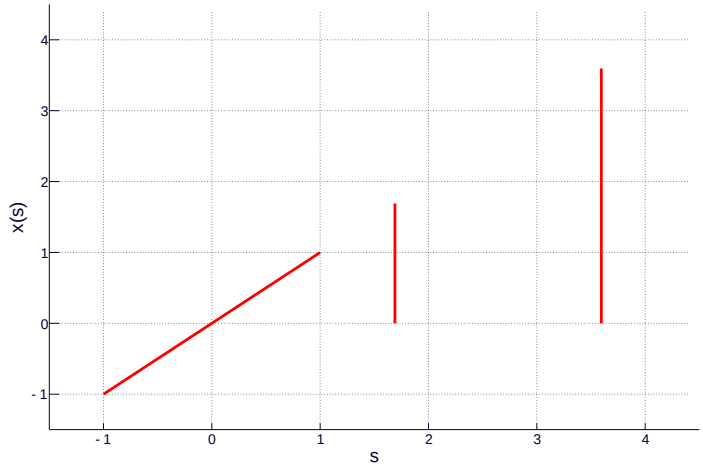
```
julia> x,U = eig(J);

julia> x
Fun(ApproxFun.PointSpace{Float64}([1.69011])+ApproxFun.PointSpace{
    Float64}([3.59504])+Ultraspherical(1,[-1.0,1.0])
    ,[1.69011,3.59504,0.0,0.5])

julia> U
SpaceOperator:ApproxFun.SequenceSpace()->ApproxFun.PointSpace{Float64
    }([1.69011])+ApproxFun.PointSpace{Float64}([3.59504])+
    Ultraspherical(1,[-1.0,1.0])
  0.1018     0.2698    -0.9046   ...     -0.001118     -0.0003662    ...
  0.0993     0.9534     0.2818            2.2988e-6      3.2616e-7    ...
  3.0867   -13.1866    90.7631            4.8174e-12    -4.2064e-12   ...
 -1.2894    10.3748   -79.174            -0.6916         8.3062e-12   ...
  0.1219    -2.7782    27.1397           13.3176        -0.691612     ...
             0.2223    -3.9551   ...    -93.2645        13.3176       ...
                        0.2048          296.516        -93.2645       ...
                                       -464.574        296.516        ...
                                        387.279       -464.574        ...
                                       -172.439        387.279        ...
                                  ...      ...            ...          ...
```

The variable `J` is of type `SymTriToeplitz`, here a $3 \times 3$ perturbation of $\Delta$.

The variable `x` is of type `Fun`, and is the identity function on the space

$$S = \mathrm{Span}(U_0, U_1, U_2, \ldots) \oplus \mathbb{1}_{1.69011} \oplus \mathbb{1}_{3.59504}, \tag{A.10}$$

which are functions supported on the spectrum of $J$. Below is a plot of the function `x` on its domain $\sigma(J) = [-1, 1] \cup \{1.69011, 3.5904\}$.



The variable `U` is of type `SpaceOperator` and is an operator mapping sequences to functions in the space $S$. The first and second rows of `U` are the eigenvectors of `J` corresponding to the eigenvalues 1.69011 and 3.59504.

The reasoning behind these choices of data structure is the following equality for any vector `v`: `Jv = U\(x*(U*x))`, just like in Matlab we have `A = V\(D*(V*v))`. This means that for a given function `f`, we can apply the operator `f(J)` to a vector `v` using the syntax, `U\(f(x)*(U*v))`.

The backslash here is the ApproxFun version of Matlab's backslash; it uses infinite dimensional adaptive QR approach to its solution (see [OT14],[OT13]).

For an example, consider the following Julia output. The operator `U` maps the vector to a function in $S$, we can take a function (exp here) of the function `x` which lies in $S$ and multiply it. Then we can invert the operator $U$ to obtain a final vector. At the bottom we compare the result to that from a finite section method.

```
julia> U*[32,4,1,.5,.25,.125,.0675]
Fun(ApproxFun.PointSpace{Float64}([1.69011])+ApproxFun.PointSpace{
    Float64}([3.59504])+Ultraspherical(1,[-1.0,1.0])
    ,[3.25603,7.2975,11.265,22.8424,-7.18042,-8.89572,12.5823,...])

julia> exp(x)*(U*[32,4,1,.5,.25,.125,.0675])
Fun(ApproxFun.PointSpace{Float64}([1.69011])+ApproxFun.PointSpace{
    Float64}([3.59504])+Chebyshev([-1.0,1.0])
    ,[17.6479,265.755,32.1056,43.1205,16.1741,-16.8182,12.5179,...])

julia> U\(exp(x)*(U*[32,4,1,.5,.25,.125,.0675]))
Fun(ApproxFun.SequenceSpace()
    ,[106.57,248.457,63.5795,8.62155,1.5722,0.497249,0.209471,...])

julia> expm(full(J[1:10,1:10]))*[32;4;1;.5;.25;.125;.0675;zeros(3)]
10-element Array{Float64,1}:
106.57
248.457
63.5795
8.62155
1.5722
0.497249
0.209471
0.0626988
0.0128002
```

```
0.00189908
```

The following types and functions from ApproxFun are used in the code below.

- `Interval` is an ApproxFun type which stores the endpoints of a real interval
- `SpaceOperator` creates an operator with prescribed domain and range
- `BlockOperator` is a type representing a matrix whose entries are operators.
- `SequenceSpace` is a `Space` of functions in ApproxFun in which functions are simply sequences indexed by the natural numbers
- `SumSpace` is a `Space` which represents to sum of two spaces of functions. Below it is used to combine some instances of `DiracSpace` and a continuous function space into a combined space.

```
function Base.eig(Jin::SymTriToeplitz)
  Qret=Array(HessenbergUnitary{'U',Float64},0)
  eigapprox=sort(discreteeigs(Jin))

  # The continuous spectrum of Jin
  ctsspec = ApproxFun.Interval(Jin.a-2*abs(Jin.b),Jin.a+2*abs(Jin.b))

  J=Jin

  # If there are no discrete eigenvalues
  if length(eigapprox) == 0
    C=connection_coeffs_operator(J)
    x=Fun(identity,Ultraspherical(1,ctsspec)) # Domain Fun is scaled
       and shifted to ctsspec
    U=SpaceOperator(C,SequenceSpace(),space(x))
    return x,U
  end

  # If there are discrete eigenvalues then perform a shifted QL
     iteration
  # with those eigenvalues as shifts
  eig=Array(Float64,0)
  tol=1E-14
  for k=1:length(eigapprox)
    mu=eigapprox[k]
```

```
24      Q,L=ql(J-mu*I)
25      push!(Qret,deflate(Q',k-1))
26      J=(L*Q+mu*I)
27
28      while abs(J[1,2]) > tol
29        # mu=J[1,1] DO NOT DO THIS. IF MU IS NOT ACCURATE, J[1,1] CAN
              BE AN INVALID SHIFT (MW)
30        Q,L=ql(J-mu*I)
31        J=L*Q+mu*I
32        push!(Qret,deflate(Q',k-1))
33      end
34
35      push!(eig,J[1,1])
36      J=J[2:end,2:end]
37    end
38
39    if length(eig) == 1
40      Q=Qret[1]
41      C=BlockOperator(eye(length(eig)),connection_coeffs_operator(J))
42      x=Fun(identity,SumSpace(PointSpace(eig[1]),Ultraspherical(1,
            ctsspec)))
43      U=SpaceOperator(C*Q,SequenceSpace(),space(x))
44      return x,U
45    else
46      Q=BandedUnitary(reverse!(Qret))
47      C=SpaceOperator(BlockOperator(eye(length(eig)),
            connection_coeffs_operator(J)),SequenceSpace(),SequenceSpace()
            )
48      x=Fun(identity,SumSpace(mapreduce(PointSpace,SumSpace,eig),
            Ultraspherical(1,ctsspec)))
49      U=SpaceOperator(C*Q,SequenceSpace(),space(x))
50      return x,U
51    end
52  end
```

# Appendix B
# Riemannian geometry and Lie theory

## B.1 Manifolds, Lie groups and Lie algebras

Without being precise, a manifold is a smooth set which looks locally like $\mathbb{R}^D$. The easiest examples that come to mind are smooth surfaces in $\mathbb{R}^D$, but a manifold needn't be defined in that way. There are coordinate-free, "intrinsic" definitions of manifolds which do not rely on such an embedding, but we will not use them because all the manifolds we are interested in are manifolds of matrices, which *do* reside naturally in an ambiant euclidean space, $\mathbb{C}^{n^2}$. The following definition will suffice, and we are sure that the geometrically initiated can translate the exposition into the more general abstract framework. We mainly follow [IMKNZ00].

**Definition B.1.1** (Manifold)**.** A $d$-dimensional manifold $\mathcal{M}$ is a $d$-dimensional smooth surface $\mathcal{M} \subseteq \mathbb{R}^D$ for some integer $D \geq d$.

Many manifolds of interest can be described as the zero set of a smooth function $g : \mathbb{R}^D \to \mathbb{R}^m$. For example, the unit sphere is the zero set of $g(x) = \|x\|_2 - 1$; the group of orthogonal matrices $\mathrm{O}(n)$ is the zero set of $g(X) = \|XX^T - I\|_F^2$; and some one-sheet hyperboloid is the zero set of $g(x, y, z) = x^2 + y^2 - z^2 - 1$.

The notion of a plane that is tangent to a surface is fundamental to calculus. In the same way, we want to define tangents at each point in the manifold.

**Definition B.1.2** (Tangent and cotangent space)**.** Let $\mathcal{M}$ be a $d$-dimensional manifold. The *tangent space* at $X \in \mathcal{M}$, denoted $T_X\mathcal{M}$, is vector space of vectors $V \in \mathbb{R}^D$ such that

$$V = \frac{\mathrm{d}\mu(s)}{\mathrm{d}s}\bigg|_{s=0} \tag{B.11}$$

for some smooth path $\mu$ in $\mathcal{M}$ such that $\mu(0) = X$. The *cotangent space* at $X$, denoted $T_X^*\mathcal{M}$, is the dual vector space to $T_X\mathcal{M}$. The *tangent bundle* and the *cotangent bundle* are $T\mathcal{M} = \bigcup_{X\in\mathcal{M}} T_X\mathcal{M}$ and $T^*\mathcal{M} = \bigcup_{X\in\mathcal{M}} T_X^*\mathcal{M}$ respectively.

**Definition B.1.3** (Vector and covector field)**.** A *vector field* on a manifold $\mathcal{M}$ is a smooth function $F : \mathcal{M} \to T\mathcal{M}$ such that $F(X) \in T_X\mathcal{M}$ for all $X \in \mathcal{M}$. The vector space of all vector fields on $\mathcal{M}$ is denoted $\mathfrak{X}(\mathcal{M})$. A *covector field* is defined in the obvious way. The vector space of all covector fields is denoted $\mathfrak{X}^*(\mathcal{M})$.

From these definitions, we see that differential equation

$$\dot{X}(t) = F(X(t)) \tag{B.12}$$

evolves on the manifold $\mathcal{M}$ for any initial datum $X_0$ if and only if $F \in \mathfrak{X}(\mathcal{M})$.

**Definition B.1.4** (Gradient covector field)**.** Let $\Psi : \mathcal{M} \to \mathbb{R}$ be a $C^1$ function. Then the *gradient covector field* of $\Psi$, denoted $\mathrm{grad}\Psi \in \mathfrak{X}^*(\mathcal{M})$, is the covector field defined by

$$\mathrm{grad}\Psi(X)(V) = \frac{\mathrm{d}}{\mathrm{d}s}\Psi(\mu(s))\bigg|_{s=0}, \tag{B.13}$$

where $\mu$ is a smooth path in $\mathcal{M}$ such that $\mu(0) = X \in \mathcal{M}$ and $\dot{\mu}(0) = V \in T_X\mathcal{M}$.

Note that a gradient vector field cannot be uniquely defined, because the isomorphism between a tangent space and cotangent space is dependent on the choice of inner product. This inner product is called a metric, and it allows us to define lengths and angles on a manifold. For our interests here though, it matters only because it changes what vectors the gradient covectors are mapped to.

**Definition B.1.5** (Metric)**.** For a manifold $\mathcal{M}$, a *metric* $g$ is a smooth assignment of a bilinear map $\langle\cdot,\cdot\rangle_g : T_X\mathcal{M} \times T_X\mathcal{M} \to \mathbb{R}$ for each $X \in \mathcal{M}$. The metric is said to be *Riemannian* if the bilinear map is positive definite for every $X \in \mathcal{M}$.

**Definition B.1.6** (Gradient vector field)**.** Let $\Psi : (\mathcal{M}, g) \to \mathbb{R}$ be a $C^1$ function. Then the *gradient vector field* of $\Psi$ is the unique vector field $\nabla_g\Psi \in \mathfrak{X}(\mathcal{M})$ such that

$$\langle\nabla_g\Psi(X), V\rangle_g = \mathrm{grad}\Psi(X)(V) \tag{B.14}$$

for all $X \in \mathcal{M}$ and $V \in T_X\mathcal{M}$.

## B.2 Lie groups and Lie algebras

Some manifolds are special[1]. Some manifolds are also groups. If the group operation is smooth, then we call the manifold a Lie group.

**Definition B.2.1** (Lie group)**.** A *Lie group* $\mathcal{G}$ is a group that is also a manifold. Additionally, the group operation must be a smooth map from $\mathcal{G} \times \mathcal{G}$ to $\mathcal{G}$. A *matrix Lie group* is a Lie group whose elements are matrices, with matrix multiplication as the group operation.

**Definition B.2.2** (Lie algebra)**.** A Lie algebra $\mathfrak{g}$ is a vector space endowed with a Lie bracket; that is a bilinear, skew-symmetric map $[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ satisfying the Jacobi identity

$$[A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0. \tag{B.15}$$

**Definition B.2.3** (Lie algebra associated to a Lie group)**.** Let $\mathcal{G}$ be a Lie group, then its associated Lie algebra is the tangent space at the identity, $T_I\mathcal{G}$. The induced Lie bracket is

$$[A, B] = \frac{\partial^2}{\partial s \partial t} \rho(s)\sigma(t)\rho(-s)\bigg|_{s=t=0}, \tag{B.16}$$

where $\rho$ and $\sigma$ are smooth curves in $\mathcal{G}$ such that $\rho(0) = \sigma(0) = I$ and $\dot{\rho}(0) = A$, $\dot{\sigma}(0) = B$.

*Remark* B.2.4. If $\mathcal{G}$ is a matrix Lie group, then the induced Lie bracket is the standard matrix commutator.

**Proposition B.2.5.** *Let $\mathcal{G}$ be a matrix Lie group with associated Lie algebra $\mathfrak{g}$. Then for each $P \in \mathcal{G}$, the tangent space is*

$$T_P\mathcal{G} = \{AP : A \in \mathfrak{g}\} = \{PA : A \in \mathfrak{g}\} \tag{B.17}$$

*Proof.* Let $P \in \mathcal{G}$ and let $\rho$ be a path in $\mathcal{G}$ such that $\rho(0) = P$. Then $\sigma_1 = \rho P^{-1}$ and $\sigma_2 = P^{-1}\rho$ are paths in $\mathcal{G}$ such that $\sigma_i(0) = I$. By definition, $\dot{\sigma}_i(0) =\in \mathfrak{g}$, which implies that $\dot{\rho}(0) = \dot{\sigma}_1(0)P = P\dot{\sigma}_2(0) \in T_P\mathcal{G}$. For the opposite inclusion, if $A \in \mathfrak{g}$ then there exists a path $\sigma$ in $\mathcal{G}$ such that $\sigma(0) = I$, $\dot{\sigma}(0) = A$. Then $\rho_1 = \sigma P$ and $\rho_2 = P\sigma$ are paths in $\mathcal{G}$ such that $\rho_i(0) = P$ and $\dot{\rho}_1(0) = AP$, $\dot{\rho}_2(0) = PA$. Hence $AP$ and $PA$ are members of $T_P\mathcal{G}$. □

[1] Manifolds like $\mathrm{SL}(n)$ and $\mathrm{SO}(n)$ are doubly "special".

This characterisation of the tangent spaces of a matrix Lie group generalises if we define multiplication between an element of $\mathfrak{g}$ and an element of $\mathcal{G}$ to be

$$AP = \frac{\mathrm{d}}{\mathrm{d}s}\left(\rho(s)P\right)\Big|_{s=0}, \quad PA = \frac{\mathrm{d}}{\mathrm{d}s}\left(P\rho(s)\right)\Big|_{s=0} \tag{B.18}$$

where $\rho$ is a path in $\mathcal{G}$ such that $\rho(0) = I$, $\dot{\rho}(0) = A$.

Unless otherwise stated, if we have both a generic Lie group $\mathcal{G}$ and a generic Lie algebra $\mathfrak{g}$ "in play", then the Lie algebra is that which is associated to the Lie group.

**Definition B.2.6** (Exponential map)**.** The exponential map $\exp : \mathfrak{g} \to \mathcal{G}$ is defined to be $\exp(A) = \rho(1)$, where $\rho$ is the unique smooth path in $\mathcal{G}$ such that

$$\dot{\rho}(t) = A\rho(t), \quad \rho(0) = I. \tag{B.19}$$

**Proposition B.2.7.** *For a matrix Lie algebra, the exponential map is precisely the matrix exponential*

$$\operatorname{expm}(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}.$$

*Proof.* Appeal to the uniqueness of solution to the IVP (B.19). □

The adjoint maps are some other maps on Lie groups and their Lie algebras that are of fundamental importance, especially for isospectral flows.

**Definition B.2.8** (AD, Ad and ad)**.** The ADjoint map, Adjoint map, and adjoint map are defined as follows:

$$\mathrm{AD} : \mathcal{G} \times \mathcal{G} \to \mathcal{G}, \quad \mathrm{AD}_P R = PRP^{-1}, \tag{B.20}$$

$$\mathrm{Ad} : \mathcal{G} \times \mathfrak{g} \to \mathfrak{g}, \quad \mathrm{Ad}_P B = PBP^{-1}, \tag{B.21}$$

$$\mathrm{ad} : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}, \quad \mathrm{ad}_A B = [A, B]\,. \tag{B.22}$$

For Ad, we have used the definition of multiplication as in (B.18).

**Lemma B.2.9** (Hadamard's lemma)**.** *Let* Ad *and* ad *be as in Definition B.2.8. Then for all* $A, B \in \mathfrak{g}$*,*

$$\mathrm{Ad}_{\exp(A)} B = \exp(A) B \exp(-A) = \sum_{k=0}^{\infty} \frac{1}{k!} \mathrm{ad}_A^k B = \exp(\mathrm{ad}_A) B. \tag{B.23}$$

*Proof.* The first and last equalities are by definition. For the main equality, consider the analytic function

$$f(s) = \exp(sA)B\exp(-sA),$$

which has the appropriate Taylor series for the desired result. □

## B.3 Differential equations, Lie groups and manifolds

We can endow a manifold with some of the features of a Lie group by finding what is known as a group action.

**Definition B.3.1** (Lie group action)**.** Let $\mathcal{G}$ be a Lie group and $\mathcal{M}$ be a manifold. Then a (left) *Lie group action* $\Lambda$ of $\mathcal{G}$ on $\mathcal{M}$ is a smooth map $\Lambda : \mathcal{G} \times \mathcal{M} \to \mathcal{M}$ such that the following two properties hold:

1. $\Lambda(I, X) = X$
2. $\Lambda(P, \Lambda(Q, X)) = \Lambda(PQ, X)$

*Remark* B.3.2. For general groups acting on sets, the definition is as above but without the smoothness requirement.

An example of a Lie group action is that of $\mathrm{SO}(n)$ on the unit sphere in $\mathbb{R}^n$ by multiplication on the left: $\Lambda(Q, x) = Qx$. Multiplication on the right *not* a left Lie group action, because $(xQ)P = x(QP) \neq x(PQ)$; it is a *right* group action.

A Lie group can act on itself by multiplication on the left. The ADjoint map is a less obvious Lie group action on itself. The Adjoint map is a Lie group action on its Lie algebra.

**Definition B.3.3.** An *orbit* of a group action on $\mathcal{M}$ is an equivalence class of equivalence relation

$$X \sim Y \iff \text{there exists } P \in \mathcal{G} \text{ such that } Y = \Lambda(P, X). \tag{B.24}$$

A group action is *transitive* if $\mathcal{M}$ consists of only one orbit.

For isospectral flows, the action we care about it that of the Adjoint map of $\mathrm{GL}(n)$ on $\mathfrak{gl}(n)$. Or is it? This does not describe the geometry and structure of an isospectral flow. We care more about the action on the *orbits*, because once a flow has begun

it cannot leave the orbit. This principle applies in general to flows that evolve on manifolds. We can restrict the manifold to orbits of a Lie group action to ensure that the action is transitive.

**Definition B.3.4.** A manifold $\mathcal{M}$ is said to be *homogeneous* if there exists a a *transitive* Lie group action $\Lambda : \mathcal{G} \times \mathcal{M} \to \mathcal{M}$ for some Lie group $\mathcal{G}$.

The manifold $\mathcal{M} = \{PX_0P^{-1} : P \in \mathrm{GL}(n)\}$ for some $X_0 \in \mathfrak{gl}(n)$ is homogeneous, because the action of $\mathrm{GL}(n)$ by the Adjoint map is transitive. However, there are some redundancies because $PX_0P^{-1} = (\alpha P X_0 (\alpha P)^{-1}$ for any nonzero constant $\alpha$. We would prefer to eliminate these redundancies by restricting in the Lie group to $\mathrm{SL}(n)$.

**Definition B.3.5.** A group action is *faithful* if $\Lambda(P, X) = \Lambda(Q, X)$ for all $X \in \mathcal{M}$ implies $P = Q$.

If a Lie group acts on a manifold, then there is a natural map between the Lie algebra and the vector fields on the manifold. This is the main motivation for introducing a Lie group action to a manifold, because it allows us to associate flows on the manifold with flows on the Lie group.

**Definition B.3.6.** Let $\Lambda : \mathcal{G} \times \mathcal{M} \to \mathcal{M}$ be a Lie group action. Then the action induces a Lie algebra homomorphism $\lambda : \mathfrak{g} \to \mathfrak{X}(\mathcal{M})$ by

$$\lambda(A)(X) = \frac{\mathrm{d}}{\mathrm{d}s}\Lambda(\rho(s), X)\bigg|_{s=0}, \tag{B.25}$$

where $\rho$ is a smooth path in $\mathcal{G}$ such that $\rho(0) = I$ and $\dot{\rho}(0) = A$.

Let $\mathcal{M} = \mathbb{R}^n$ and $\mathcal{G} = \mathrm{SO}(n)$ with group action $\Lambda(Q, x) = Qx$. Then for $\Theta \in \mathfrak{so}(n)$, $x \in \mathbb{R}^n$, $\lambda(\Theta)(x) = \Theta x$. For a less trivial example, let $\mathcal{M} = \mathfrak{gl}(n)$ and $\mathcal{G} = \mathrm{SL}(n)$ with group action $\Lambda(P, X) = PXP^{-1}$. Then $\lambda(A)(X) = [A, X]$.

**Lemma B.3.7** (Lie group equation induced by a group action)**.** *Let $\Lambda : \mathcal{G} \times \mathcal{M} \to \mathcal{M}$ be a Lie group action and let $A : [0, \infty) \times \mathcal{M} \to \mathfrak{g}$ be $C^1$. Then the flow of $\lambda(A)$ starting from $X_0 \in \mathcal{M}$,*

$$\dot{X}(t) = \lambda(A(t, X(t)))(X(t)), \quad X(0) = X_0 \in \mathcal{M}, \tag{B.26}$$

*can be written in the form*

$$X(t) = \Lambda(P(t), X_0), \tag{B.27}$$

*where $P : [0, \infty) \to \mathcal{G}$ satisfies the differential equation*

$$\dot{P}(t) = A(t, \Lambda(P(t), X_0))P(t), \quad P(0) = I, \tag{B.28}$$

*where multiplication is defined as in* (B.18).

*Proof.* Let $P$ be the unique solution to (B.28) and define $X(t) = \Lambda(P(t), X_0)$. For a fixed $t > 0$ let $\rho$ be a path in $\mathcal{G}$ such that $\rho(0) = I$ and $\dot{\rho}(0) = A(t, X(t))$. Then

$$\begin{aligned} \lambda(A(t, X(t)))(X(t))) &= \frac{\mathrm{d}}{\mathrm{d}s}\Lambda(\rho(s), X(t))\Big|_{s=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}s}\Lambda(\rho(s), \Lambda(P(t), X_0))\Big|_{s=0}. \end{aligned}$$

The second property of group actions gives

$$\begin{aligned} \lambda(A(t, X(t)))(X(t))) &= \frac{\mathrm{d}}{\mathrm{d}s}\Lambda(\rho(s)P(t), X_0)\Big|_{s=0} \\ &= \mathrm{grad}\Lambda(\rho(s)P(t), X_0)(\dot{\rho}(s)P(t))\Big|_{s=0} \\ &= \mathrm{grad}\Lambda(P(t), X_0)(A(t, X(t))P(t)) \\ &= \mathrm{grad}\Lambda(P(t), X_0)(\dot{P}(t)) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\Lambda(P(t), X_0) \\ &= \dot{X}(t), \end{aligned}$$

where the gradient is taken with respect to the Lie group variable. Since $X(0) = \Lambda(P(0), X_0) = X_0$, the $X$ we defined must be the desired flow by uniqueness. □

As we have already mentioned, this lemma allows us to take a differential equation which evolves on an homogeneous manifold and transform it into an equation on a Lie group. For isospectral flows, such abstraction is not necesssary, as the Lie group equation was elementary and intuitive to compute. However, here is where the abstract approach becomes very useful indeed. Using it we will be able to transform the Lie group equation into an equation on its Lie algebra, something which is not obviously possible for isospectral flows.

**Definition B.3.8.** Let $\varphi : \mathfrak{g} \to \mathcal{G}$ be a smooth function. Its differential is the *right trivialised tangent*, $\mathrm{d}\varphi : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$, defined such that for any path $\Omega$ in $\mathfrak{g}$,

$$\frac{\mathrm{d}}{\mathrm{d}s}\varphi(\Omega(s)) = \mathrm{d}\varphi_{\Omega(s)}(\dot{\Omega}(s))\varphi(\Omega(s)). \tag{B.29}$$

Here multiplication is as defined in (B.18).

The primary example of such a smooth function is $\varphi = \exp$. By differentiating term by term, we have

$$\frac{\mathrm{d}}{\mathrm{d}s}\exp(\Omega) = \sum_{k=0}^{\infty}\frac{1}{k!}\sum_{i=1}^{k}\Omega^{i-1}\dot{\Omega}\Omega^{k-i}.$$

We can induce cancellation by applying $\mathrm{ad}_\Omega$:

$$\begin{aligned}\mathrm{ad}_\Omega\left(\frac{\mathrm{d}}{\mathrm{d}s}\exp(\Omega)\right) &= \sum_{k=0}^{\infty}\frac{1}{k!}\left(\Omega^k\dot{\Omega} - \dot{\Omega}\Omega^k\right)\\ &= \exp(\Omega)\dot{\Omega} - \dot{\Omega}\exp(\Omega)\\ &= \left(\mathrm{Ad}_{\exp(\Omega)}\dot{\Omega} - \dot{\Omega}\right)\exp(\Omega).\end{aligned}$$

Now, multiplying both sides on the right by $\exp(-\Omega)$ and using Hadamard's Lemma B.2.9, we get

$$\mathrm{ad}_\Omega\left(\left(\frac{\mathrm{d}}{\mathrm{d}s}\exp(\Omega)\right)\exp(-\Omega)\right) = \exp(\mathrm{ad}_\Omega)(\dot{\Omega}) - \dot{\Omega}. \tag{B.30}$$

Hence we have the formal expression

$$\mathrm{d}\exp_\Omega = \frac{\exp(\mathrm{ad}_\Omega) - I}{\mathrm{ad}_\Omega} = \sum_{k=0}^{\infty}\frac{1}{(k+1)!}\mathrm{ad}_\Omega^k. \tag{B.31}$$

**Lemma B.3.9** (Lie algebra equation induced by $\varphi$)**.** *Let $\varphi : \mathfrak{g} \to \mathcal{G}$ be a smooth function such that $\varphi(O) = I$, and let $P : [0,\infty) \to \mathcal{G}$ satisfy the differential equation*

$$\dot{P}(t) = C(t, P(t))P(t), \quad P(0) = I, \tag{B.32}$$

*for a given $C^1$ function $C : [0,\infty) \times \mathcal{G} \to \mathfrak{g}$ (multiplication is as defined in (B.18)). Then $P$ can be written in the form*

$$P(t) = \varphi(\Omega(t)),$$

*where $\Omega : [0,\infty) \to \mathfrak{g}$ satisfies the differential equation*

$$\dot{\Omega}(t) = \mathrm{d}\varphi^{-1}_{\Omega(t)}(C(t, \varphi(\Omega(t)))), \quad \Omega(0) = 0. \tag{B.33}$$

*Proof.* Suppose that $\Omega : [0, \infty) \to \mathfrak{g}$ is the unique solution to (B.33). Then

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\varphi(\Omega(t)) &= \mathrm{d}\varphi_{\Omega(t)}(\dot{\Omega}(t))\varphi(\Omega(t)) \\ &= C(t, \varphi(\Omega(t)))\varphi(\Omega(t)), \end{aligned}$$

and $\varphi(\Omega(0)) = I$. By uniqueness of solution to (B.32), $P(t) = \varphi(\Omega(t))$. □

Since $\mathrm{d}\exp_\Omega$ is an analytic function in $\mathrm{ad}_\Omega$, it is straightforward to invert:

$$\mathrm{d}\exp_\Omega^{-1} = \frac{\mathrm{ad}_\Omega}{\exp(\mathrm{ad}_\Omega) - I} = \sum_{k=0}^{\infty} \frac{B_k}{k!}\mathrm{ad}_\Omega^k, \tag{B.34}$$

where $B_0, B_1, B_2, \ldots$ are the Bernoulli numbers, $1, -\frac{1}{2}, \frac{1}{6}, 0, -\frac{1}{30}, 0, \ldots$.

We can combine Lemmas B.3.7 and B.3.9 results to obtain a Lie algebra equation induced by a Lie group action on a manifold. For a Lie group action $\Lambda : \mathcal{G} \times \mathcal{M} \to \mathcal{M}$, $C^1$ function $A : [0, \infty) \times \mathcal{M} \to \mathfrak{g}$, smooth map $\varphi : \mathfrak{g} \to \mathcal{G}$ such that $\varphi(0) = I$, and $X_0 \in \mathcal{M}$, the solution $X$ of

$$\dot{X}(t) = \lambda(A(t, X(t)))(X(t)), \quad X(0) = X_0, \tag{B.35}$$

which evolves in $\mathcal{M}$, can be expressed as

$$X(t) = \Lambda(\varphi(\Omega(t)), X_0), \tag{B.36}$$

where $\Omega : [0, \infty) \to \mathfrak{g}$ satisfies

$$\dot{\Omega}(t) = \mathrm{d}\varphi_\Omega^{-1} A(t, \Lambda(\varphi(\Omega(t)), X_0)), \quad \Omega = O. \tag{B.37}$$

The three layers involved are summarised in a table:

| Manifold $\mathcal{M}$ | Lie group $\mathcal{G}$ | Lie algebra $\mathfrak{g}$ |
|---|---|---|
| $\dot{X} = \lambda(B(X))(X)$ | $\dot{P} = B(X)P$ | $\dot{\Omega} = \mathrm{d}\varphi_{\Omega(t)}^{-1} B(X)$ |
| $X(0) = X_0$ | $P(0) = I$ | $\Omega(0) = O$ |
| $X = \Lambda(P, X_0)$ | $P = \varphi(\Omega)$ | |

Let's see what this looks like for an isospectral flow $\dot{X} = [A(X), X]$ with $\varphi = \exp$:

$$\begin{aligned} \dot{\Omega}(t) &= \mathrm{d}\exp_{\Omega(t)}^{-1} A(\mathrm{Ad}_{\exp(\Omega(t))} X_0) \\ &= \sum_{k,j=0}^{\infty} \frac{B_k}{k!j!} \left(\mathrm{ad}_{\Omega(t)}^k \circ A \circ \mathrm{ad}_{\Omega(t)}^j\right) X_0 \end{aligned}$$

In some cases, we can calculate this series exactly. For example, if $\mathfrak{g} = \mathfrak{so}(3)$, then $\mathrm{ad}_\Omega^3 A = -\|\Omega\|_F^2 \mathrm{ad}_\Omega A$ for all $A \in \mathfrak{so}(3)$, so that

$$\mathrm{d}\exp_\Omega^{-1} A = A - \frac{1}{2}\mathrm{ad}_\Omega A + \left(\frac{1}{2}\|\Omega\|_F \cot(\frac{1}{2}\|\Omega\|_F) - 1\right) \mathrm{ad}_\Omega^2 A. \tag{B.38}$$

Otherwise, we must resort to a truncation to an acceptable error margin.

### B.3.1 Quadratic Lie groups and the Cayley transform

A restricted class of matrix Lie groups is that of the quadratic Lie groups. Such Lie groups have a particularly nice mapping $\varphi : \mathfrak{g} \to \mathcal{G}$ called the Cayley transform.

**Definition B.3.10** (Quadratic Lie group)**.** A Lie group $\mathcal{G}$ is *quadratic* if it is of the form

$$\mathcal{G} = \left\{Q \in \mathrm{GL}(n) : QRQ^T = R\right\}, \tag{B.39}$$

where $R \in \mathrm{GL}(n)$ is a prescribed matrix.

The most obvious example of a quadratic Lie group is the orthogonal group $\mathrm{O}(n)$. Other examples include the symplectic group $\mathrm{Sp}(2n)(\subset \mathrm{GL}(2n))$ for which

$$R = \begin{pmatrix} O_n & I_n \\ -I_n & O_n \end{pmatrix}, \tag{B.40}$$

and the Lorentz group $\mathrm{SO}(3,1) \subset \mathrm{SL}(4)$, for which

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{B.41}$$

The former is appears in the geometry of Hamiltonian systems and the latter appears in Special Relativity.

**Proposition B.3.11** (Quadratic Lie algebra)**.** *The Lie algebra associated to the quadratic Lie group B.39 is*

$$\mathfrak{g} = \left\{B \in \mathfrak{gl}(n) : BR + RB^T = O\right\} \tag{B.42}$$

*Any Lie algebra of this form is a called a quadratic Lie algebra.*

**Definition B.3.12** (Cayley transform)**.** Let $\mathfrak{g}$ be a quadratic Lie algebra. Then the *Cayley transform* of $\Omega \in \mathfrak{g}$ is defined to be

$$\mathrm{Cay}(B) = (I - \frac{1}{2}B)^{-1}(I + \frac{1}{2}B) = (I + \frac{1}{2}B)(I - \frac{1}{2}B)^{-1}. \tag{B.43}$$

We stress that although the Cayley transform takes the form of the Padé approximant $\phi(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$ for the exponential function, it has nothing to do with the exponential function. It would be just as valid to change $\frac{1}{2}$ to any real nonzero constant.

**Lemma B.3.13.** *If $\mathcal{G}$ is quadratic, then the Cayley transform maps $\mathfrak{g}$ into $\mathcal{G}$.*

*Proof.* Let $B \in \mathfrak{g}$. Then

$$\begin{aligned}
\mathrm{Cay}(B)R\mathrm{Cay}(B)^T &= (I - \frac{1}{2}B)^{-1}(R + \frac{1}{2}BR)\mathrm{Cay}(B^T) \\
&= (I - \frac{1}{2}B)^{-1}R(I - \frac{1}{2}B^T)\mathrm{Cay}(B^T) \\
&= (R^{-1} - \frac{1}{2}R^{-1}B)^{-1}(I - \frac{1}{2}B^T)\mathrm{Cay}(B^T) \\
&= (R^{-1} + \frac{1}{2}B^TR^{-1})^{-1}(I - \frac{1}{2}B^T)\mathrm{Cay}(B^T) \\
&= R\mathrm{Cay}(B^T)^{-1}\mathrm{Cay}(B^T) \\
&= R.
\end{aligned}$$

Therefore $\mathrm{Cay}(B) \in \mathcal{G}$. □

**Proposition B.3.14** (Differential of Cay)**.** *The right trivalised differential of the Cayley transform is the map*

$$\mathrm{dCay}_\Omega A = (I - \frac{1}{2}\Omega)^{-1}A(I + \frac{1}{2}\Omega)^{-1}. \tag{B.44}$$

*Hence the inverse is the map*

$$\mathrm{dCay}_\Omega^{-1} A = A - \frac{1}{2}[\Omega, A] - \frac{1}{4}\Omega A\Omega. \tag{B.45}$$

*Proof.* A direct calculation of $\frac{\mathrm{d}}{\mathrm{d}s}\mathrm{Cay}(\Omega(s))$ for a smooth path $\Omega$ in $\mathfrak{g}$ confirms the result. □

The inverse differential $\mathrm{dCay}_\Omega^{-1}$ is quadratic in $\Omega$, no matter what the dimension of the Lie algebra is. Hence we can always evaluate it exactly, unlike with the exponential, where we usually have to do some sort of approximation.

# Appendix C
# Useful matrix identities

## C.1 Derivatives

**Lemma C.1.1** (Derivative of an inverse)**.** *Let* $P \in \mathrm{GL}(n)$*. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t}P(t)^{-1} = -P(t)^{-1}P'(t)P(t)^{-1}. \tag{C.46}$$

**Lemma C.1.2** (Jacobi's formula)**.** *Let* $P \in \mathrm{GL}(n)$*. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\det(P(t)) = (\det(P(t)))\mathrm{tr}(\dot{P}(t)P(t)^{-1}). \tag{C.47}$$

## C.2 Frobenius Inner Product

**Lemma C.2.1** (Adjoint of the adjoint map)**.** *Let* $A, B, C \in \mathbb{R}^{n\times n}$*. Then*

$$\langle A, \mathrm{ad}_C B\rangle_F = \langle \mathrm{ad}_{C^T} A, B\rangle_F. \tag{C.48}$$

# Appendix D
# Some results in analysis

## D.1 The Radon–Nikodym derivative

A reference for this section is [Tes16].

**Definition D.1.1** (Absolute continuity of measures)**.** Let $\mu$ and $\nu$ be measures on a measurable space $(X, \Sigma)$. We say that $\nu$ is absolutely continuous with respect to $\mu$, if,

$$\mu(A) = 0 \Rightarrow \nu(A) = 0, \tag{D.49}$$

for all sets $A \in \Sigma$. We write $\nu \ll \mu$.

**Theorem D.1.2** (Radon-Nikodym)**.** *Let $\mu$, $\nu$ be two $\sigma$-finite measures on a measurable space $(X, \Sigma)$. Then $\nu$ is absolutely continuous with respect to $\mu$ if and only if there exists $f \in L^1_\mu(X, \Sigma)$ such that*

$$\nu(A) = \int_A f(s)\, \mathrm{d}\mu(s) \tag{D.50}$$

*for every $A \in \Sigma$. The function $f$ is called the Radon-Nikodym derivative of $\nu$ with respect to $\mu$, and is denoted $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.*

## D.2 Fredholm operators

A reference for this section is [Sim79], or many other good functional analysis books.

**Definition D.2.1** (Fredholm operator)**.** An operator $A \in \mathcal{B}(\ell^2)$ is *Fredholm* if the range $\mathrm{Ran}(A)$ is closed and both $\mathrm{Ker}(A)$ and $\mathrm{Ker}(A^T)$ are finite dimensional. The

*index* of a Fredholm operator is

$$\mathrm{ind}(A) = \dim(\mathrm{Ker}(A)) - \dim(\mathrm{Ker}(A^T))$$

*Remark* D.2.2. An equivalent definition in the textbooks is that the kernel and cokernel of $A$ are both finite dimensional.

**Theorem D.2.3.** *An operator $A \in \mathcal{B}(\ell^2)$ is Fredholm with index 0 if and only if it is a compact perturbation of an invertible operator.*

**Theorem D.2.4** (Fredholm alternative)**.** *Let $A \in \mathcal{B}(\ell^2)$ be Fredholm with index 0. Then $A$ is injective if and only if it is surjective.*

**Theorem D.2.5.** *Let $A, B \in \mathcal{B}(\ell^2)$ be Fredholm operators with indices $k_a$ and $k_b$ respectively. Then both $AB$ and $BA$ are Fredholm with index $k_a + k_b$.*

# Bibliography


[ADH97] Ashlock, D.A., Driessel, K.R. and Hentzel, I.R. *On matrix structures invariant under Toda-like isospectral flows.* Linear Algebra Appl. (1997). 254, no. 1: 29–48.

[AH05] Atkinson, K. and Han, W. *Theoretical numerical analysis*, vol. 39. Springer (2005).

[AK65] Akhiezer, N.I. and Kemmer, N. *The classical moment problem: and some related questions in analysis*, vol. 5. Oliver & Boyd Edinb. (1965).

[AMVW15] Aurentz, J.L., Mach, T., Vandebril, R. and Watkins, D.S. *Fast and backward stable computation of roots of polynomials.* SIAM J. Matrix Anal. A. (2015). 36, no. 3: 942–973.

[Arv94a] Arveson, W. *C*-algebras and numerical linear algebra.* J. Funct. Anal. (1994). 122, no. 2: 333–360.

[Arv94b] Arveson, W. *The role of c∗-algebras in infinite dimensional numerical linear algebra.* In *Contemp. Math* (1994) .

[Ask75] Askey, R. *Orthogonal polynomials and special functions.* SIAM (1975).

[BAHNS15a] Ben-Artzi, J., Hansen, A.C., Nevanlinna, O. and Seidel, M. *Can everything be computed? on the solvability complexity index and towers of algorithms.* arXiv preprint arXiv:1508.03280 (2015).

[BAHNS15b] Ben-Artzi, J, Hansen, A.C., Nevanlinna, O. and Seidel, M. *New barriers in complexity theory: on the solvability complexity index and the towers of algorithms.* C. R. Acad. Sci. Paris, Ser. I (2015). 353, no. 10: 931 – 936.

[Bat90] Batterson, S. *Convergence of the shifted QR algorithm on 3× 3 normal matrices.* Numer. Math. (1990). 58, no. 1: 341–352.

[Bat94] Batterson, S. *Convergence of the Francis shifted QR algorithm on normal matrices.* Linear Algebra Appl. (1994). 207: 181–195.

[BBI+09] Bloch, A.M., Brînzănescu, V., Iserles, A., Marsden, J.E. and Ratiu, T.S. *A class of integrable flows on the space of symmetric matrices.* Comm. Math. Phys. (2009). 290, no. 2: 399–435.

[BG98] Bloch, A.M. and Gekhtman, M.I. *Hamiltonian and gradient structures in the Toda flows.* J. Geom. Phys. (1998). 27, no. 3: 230–248.

[BH11] Brouwer, A.E. and Haemers, W.H. *Spectra of graphs.* Springer (2011).

[BI06] Bloch, A.M. and Iserles, A. *On an isospectral Lie–Poisson system and its Lie algebra.* Found. Comput. Math. (2006). 6, no. 1: 121–144.

[Blo90] Bloch, A.M. *Steepest descent, linear programming and Hamiltonian flows.* Contemp. Math. AMS (1990). 114: 77–88.

[Bof10] Boffi, D. *Finite element approximation of eigenvalue problems.* Acta Numerica (2010). 19: 1–120.

[BP94] Berman, A. and Plemmons, R.J. *Nonnegative matrices in the mathematical sciences.* SIAM (1994).

[BP98] Brin, S. and Page, L. *The anatomy of a large-scale hypertextual web search engine.* Comput. Networks ISDN (1998). 30, no. 1: 107–117.

[Bro91] Brockett, R.W. *Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems.* Linear Algebra Appl. (1991). 146: 79–91.

[Bro93] Brockett, R.W. *Differential geometry and the design of gradient algorithms.* In *Proc. Symp. Pure Math.*, vol. 54 (1993) 69–92.

[BS] Benet, L. and Sanders, D.P. *ValidatedNumerics.jl Julia package.* github.com/dpsanders/ValidatedNumerics.jl.

[BS13] Böttcher, A. and Silbermann, B. *Analysis of Toeplitz operators.* Springer (2013).

[BV90] Babelon, O. and Viallet, C.M. *Hamiltonian structures and Lax equations.* Phys. Lett. B (1990). 237, no. 3: 411–416.

[CB76] Cantoni, A. and Butler, P. *Eigenvalues and eigenvectors of symmetric centrosymmetric matrices.* Linear Alg. Appl. (1976). 13, no. 3: 275–288.

[CD89] Chu, M.T. and Driessel, K.R. *Can real symmetric Toeplitz matrices have arbitrary real spectra?* Tech. rep., Ida. State Univ. (1989).

[CD90] Chu, M.T. and Driessel, K.R. *The projected gradient method for least squares matrix approximations with spectral constraints.* SIAM J. Numer. Anal. (1990). 27, no. 4: 1050–1060.

[CG02] Chu, M.T. and Golub, G.H. *Structured inverse eigenvalue problems.* Acta Numer. (2002). 11, no. 0: 1–71.

[CG05] Chu, M.T. and Golub, G.H. *Inverse eigenvalue problems: theory, algorithms, and applications.* Oxf. Univ. Press (2005).

[Chu93] Chu, M.T. *On the differential equation $dx/dt = [X, k(X)]$ where $k$ is a Toeplitz annihilator.* Tech. rep., N.C. State Univ. (1993). URL http://www4.ncsu.edu/~mtchu/Research/Papers/ode.pdf.

[Chu94] Chu, M.T. *A list of matrix flows with applications.* Fields Inst. Commun. (1994). 3: 87–97.

[Chu95] Chu, M.T. *Scaled Toda-like flows.* Linear Algebra Appl. (1995). 215: 261–273.

[Chu98] Chu, M.T. *Inverse eigenvalue problems.* SIAM Rev. (1998). 40, no. 1: 1–39.

[Chu08] Chu, M.T. *Linear algebra algorithms as dynamical systems.* Acta Numer. (2008). 17: 1–86.

[Cip00] Cipra, B.A. *The best of the 20th century: editors name top 10 algorithms.* SIAM news (2000). 33, no. 4: 1–2.

[CIZ97] Calvo, M., Iserles, A. and Zanna, A. *Numerical solution of isospectral flows.* Math. Comput. (1997). 66, no. 220: 1461–1486.

[Cle55] Clenshaw, C.W. *A note on the summation of chebyshev series.* Math. Comput. (1955). 9, no. 51: 118–120.

[CT65] Cooley, J.W. and Tukey, J.W. *An algorithm for the machine calculation of complex Fourier series.* Math. Comput. (1965). 19, no. 90: 297–301.

[Day96] Day, D. *How the shifted QR algorithm fails to converge and how to fix it.* Tech. rep., Tech. Report 96–0913, Sandia National Laboratories, Albuquerque, NM (1996).

[DE15] Dubbs, A. and Edelman, A. *Infinite random matrix theory, tridiagonal bordered Toeplitz matrices, and the moment problem.* Linear Algebra Appl. (2015). 467: 188–201.

[Dei00] Deift, P. *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*, vol. 3. Am. Math. Soc. (2000).

[Dei08] Deift, P. *Some open problems in random matrix theory and the theory of integrable systems.* Contemp. Math. (2008). 458: 419.

[DHT14] Driscoll, T.A., Hale, N. and Trefethen, L.N. *Chebfun Guide.* Pafnuty Publications (2014).

[DKS10] Damanik, D., Killip, R. and Simon, B. *Perturbations of orthogonal polynomials with periodic recursion coefficients.* Ann. Math. (2010). 1931–2010.

[DLT85] Deift, P., Li, L.C. and Tomei, C. *Toda flows with infinitely many variables.* J. Funct. Anal. (1985). 64, no. 3: 358–402.

[DN86] Dombrowski, J. and Nevai, P. *Orthogonal polynomials, measures and recurrence relations.* SIAM J. Math. Anal. (1986). 17, no. 3: 752–759.

[DNT83] Deift, Percy A., Nanda, Tara and Tomei, Carlos. *Ordinary differential equations and the symmetric eigenvalue problem.* SIAM J. Numer. Anal. (1983). 20, no. 1: 1–22.

[DP04] Davies, E.B. and Plum, M. *Spectral pollution.* IMA J. Numer. Anal. (2004). 24, no. 3.

[DPS08] Damanik, D., Pushnitski, A. and Simon, B. *The analytic theory of matrix orthogonal polynomials.* Surveys in Approximation Theory (2008). 4: 1–85.

[Dri87] Driessel, K.R. *On finding the eigenvalues and eigenvectors of a matrix by means of an isospectral gradient flow.* In *Tech. Rep. 541.* Dept. of Mathematical Sciences, Clemson Univ Clemson, SC (1987).

[DS99] Diele, F. and Sgura, I. *Isospectral flows and the inverse eigenvalue problem for Toeplitz matrices.* J. Comput. Appl. Math. (1999). 110, no. 1: 25–43.

[DS02] Diele, F. and Sgura, I. *The Cayley method and the inverse eigenvalue problem for Toeplitz matrices.* BIT Numer. Math. (2002). 42, no. 2: 285–299.

[DS03] Diele, F. and Sgura, I. *Centrosymmetric isospectral flows and some inverse eigenvalue problems.* Linear Algebra Appl. (2003). 366: 199–214.

[DS06a] Damanik, D. and Simon, B. *Jost functions and Jost solutions for Jacobi matrices, I. a necessary and sufficient condition for Szegő asymptotics.* Invent. Math. (2006). 165, no. 1: 1–50.

[DS06b] Damanik, David and Simon, Barry. *Jost functions and Jost solutions for Jacobi matrices, II. decay and analyticity.* Int. Math. Res. Notices (2006). 2006.

[DSBB71] Dunford, N., Schwartz, J.T., Bade, W.G. and Bartle, R.G. *Linear operators.* Wiley (1971).

[EH75] Eberlein, P.J. and Huang, C.P. *Global convergence of the QR algorithm for unitary matrices with some results for normal matrices.* SIAM J. Numer. Anal. (1975). 12, no. 1: 97–104.

[Fla74] Flaschka, H. *The Toda lattice II. Existence of integrals.* Phys. Rev. B (1974). 9, no. 4: 1924.

[FLSL66] Feynman, R.P., Leighton, R.B., Sands, M. and Lindsay, R.B. *The Feynman lectures on physics, vol. 3: Quantum mechanics* (1966).

[FNO87] Friedland, S., Nocedal, J. and Overton, M.L. *The formulation and analysis of numerical methods for inverse eigenvalue problems.* SIAM J. Numer. Analysis (1987). 24, no. 3: 634–667.

[Fra61] Francis, J.G.F. *The QR transformation. A unitary analogue to the LR transformation—part 1.* Comput. J. (1961). 4, no. 3: 265–271.

[Fra99] Frank, J. *Introduction to computational chemistry.* Wiley (1999).

[Gau04] Gautschi, W. *Orthogonal polynomials: computation and approximation.* Oxf. Univ. Press (2004).

[GC80] Geronimo, J.S. and Case, K.M. *Scattering theory and polynomials orthogonal on the real line.* T. Am. Math. Soc. (1980). 258, no. 2: 467–494.

[GM09] Golub, G.H. and Meurant, G. *Matrices, moments and quadrature with applications.* Princet. Univ. Press (2009).

[GNR16] Gamboa, F., Nagel, J. and Rouault, A. *Sum rules via large deviations.* Journal of Functional Analysis (2016). 270, no. 2: 509–559.

[Gra86] Gragg, W.B. *The QR algorithm for unitary hessenberg matrices.* J. Comput. Appl. Math. (1986). 16, no. 1: 1–8.

[Gra06] Gray, R.M. *Toeplitz and circulant matrices: A review.* Found. Trends Comm. Info. Theory (2006). 2, no. 3: 155–239.

[GV64] Gelfand, I.M. and Vilenkin, N.Y. *Generalized functions vol. 4: applications of harmonic analysis (translated from the Russian by Amiel Feinstein).* Acad. Press (1964).

[GVA86] Geronimo, J.S. and Van Assche, W. *Orthogonal polynomials with asymptotically periodic recurrence coefficients.* J. Approx. Theory (1986). 46, no. 3: 251–283.

[GVL12] Golub, G.H. and Van Loan, C.F. *Matrix computations*, vol. 3. JHU Press (2012).

[GWW92] Gordon, C., Webb, D.L. and Wolpert, S. *One cannot hear the shape of a drum.* Bull. Am. Math. Soc. (1992). 27, no. 1: 134–138.

[Han08] Hansen, A.C. *On the approximation of spectra of linear operators on Hilbert spaces.* J. Funct. Anal. (2008). 254, no. 8: 2092–2126.

[Han09] Hansen, A.C. *The infinite dimensional QR algorithm.* Tech. rep., University of Cambridge (2009). URL http://www.damtp.cam.ac.uk/research/afha/anders/Inf_QR1.pdf.

[Han10] Hansen, A.C. *Infinite-dimensional numerical linear algebra: theory and applications.* P. Roy. Soc. Lond. A (2010).

[Han11] Hansen, A.C. *On the solvability complexity index, the n-pseudospectrum and approximations of spectra of operators.* J. Am. Math. Soc. (2011). 24, no. 1: 81–124.

[Hay08] Haykin, S.S. *Adaptive filter theory.* Pearson (2008).

[Hei01] Heinig, G. *Not every matrix is similar to a toeplitz matrix.* Linear Algebra Appl. (2001). 332: 519–531.

[Hig08] Higham, N.J. *Functions of matrices: theory and computation.* SIAM (2008).

[HJB85] Heideman, M.T., Johnson, D.H. and Burrus, C.S. *Gauss and the history of the fast Fourier transform.* Arch. Hist. Exact Sci. (1985). 34, no. 3: 265–277.

[HM94] Helmke, U. and Moore, J.B. *Optimization and dynamical systems.* Springer (1994).

[HO09] Huybrechs, D. and Olver, S. *Highly oscillatory quadrature.* Highly oscillatory problems (2009). 366: 25–50.

[HSS01] Hubbard, J., Schleicher, D. and Sutherland, S. *How to find all roots of complex polynomials by Newton's method.* Invent. Math. (2001). 146, no. 1: 1–33.

[IMKNZ00] Iserles, A., Munthe-Kaas, H.Z., Nørsett, S.P. and Zanna, A. *Lie-group methods.* Acta Numer. (2000). 9, no. 1: 215–365.

[IQ16] Iserles, A. and Quispel, G.R.W. *Why geometric integration?* In *Discrete Mechanics, Geometric Integration and Lie–Butcher Series* (2016) .

[Ise02] Iserles, A. *On the discretization of double-bracket flows.* Found. Comput. Math. (2002). 2, no. 3: 305–329.

[Kac66] Kac, M. *Can one hear the shape of a drum?* Am. Math. Month. (1966). 73, no. 4: 1–23.

[Kat95] Kato, T. *Perturbation theory for linear operators*, vol. 132. Springer (1995).

[Kau16] Kaur, A. *On solving an isospectral flow.* J. Comput. Appl. Math. (2016). 308: 263–275.

[Kos79] Kostant, B. *The solution to a generalized Toda lattice and representation theory.* Adv. Math. (1979). 34, no. 3: 195–338.

[Kro] Kropf, E. H. *ComplexPhasePortrait.jl Julia package.* github.com/ehkropf/ComplexPhasePortrait.jl.

[KS03] Killip, R. and Simon, B. *Sum rules for Jacobi matrices and their applications to spectral theory.* Ann. Math. (2003). 253–321.

[Kub62] Kublanovskaya, V.N. *On some algorithms for the solution of the complete eigenvalue problem.* USSR Comput. Math. Math. Phys. (1962). 1, no. 3: 637–657.

[KV95] Korteweg, D.J. and de Vries, G. *On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves.* Lond., Edinb., Dublin Philos. Mag. J. Sci. (1895). 39, no. 240: 422–443.

[Lan94] Landau, H.J. *The inverse eigenvalue problem for real symmetric Toeplitz matrices.* J. Am. Math. Soc. (1994). 7, no. 3: 749–767.

[Lau88] Laurie, D.P. *A numerical approach to the inverse Toeplitz eigenproblem.* SIAM J. Sci. Stat. Comput. (1988). 9, no. 2: 401–405.

[Lau91] Laurie, Dirk P. *Solving the inverse eigenvalue problem via the eigenvector matrix.* Journal of Computational and Applied Mathematics (1991). 35, no. 1-3: 277–289.

[Lau01] Laurie, D.P. *Initial values for the inverse Toeplitz eigenvalue problem.* SIAM J. Sci. Comput. (2001). 22, no. 6: 2239–2255.

[Lax68] Lax, P.D. *Integrals of nonlinear equations of evolution and solitary waves.* Comm. Pure Appl. Math. (1968). 21, no. 5: 467–490.

[LS04] Levitin, M. and Shargorodsky, E. *Spectral pollution and second-order relative spectra for self-adjoint operators.* IMA J. Numer. Anal. (2004). 24, no. 3: 393–416.

[LSL99] Li, H., Stoica, P. and Li, J. *Computationally efficient maximum likelihood estimation of structured covariance matrices.* IEEE T. Signal Proces. (1999). 47, no. 5: 1314–1323.

[Mar13] Marion, J.B. *Classical dynamics of particles and systems.* Academic Press (2013).

[MH16] Matthysen, R. and Huybrechs, D. *Fast algorithms for the computation of fourier extensions of arbitrary length.* SIAM J. Sci. Comput. (2016). 38, no. 2: A899–A922.

[MMD03] Mackey, D.S., Mackey, N. and Dunlavy, D.M. *Structure preserving algorithms for perplectic eigenproblems.* Electron. J. Linear Algebra (2003). 1: 1.

[MMP99] Mackey, D.S., Mackey, N. and Petrovic, S. *Is every matrix similar to a toeplitz matrix?* Linear Algebra Appl. (1999). 297, no. 1-3: 87–105.

[Mos75] Moser, J. *Finitely many mass points on the line under the influence of an exponential potential – an integrable system.* In *Dyn. Sys. Theory Appl.*, vol. 38 of *Lecture Notes in Physics, Berlin Springer Verlag* (1975) 467–497.

[Nor98] Norris, J.R. *Markov chains.* 2. Camb. Univ. press (1998).

[NVA92] Nevai, P. and Van Assche, W. *Compact perturbations of orthogonal polynomials.* Pac. J. Math. (1992). 153, no. 1: 163–184.

[olva] *NIST Digital Library of Mathematical Functions.* http://dlmf.nist.gov/, Release 1.0.14 of 2016-12-21. URL http://dlmf.nist.gov/. Olver, F.W.J., Olde Daalhuis, A.B., Lozier, D.W., Schneider, B.I., Boisvert, R.F., Clark, C.W., Miller, B. R. and Saunders, B.V., eds.

[Olvb] Olver, S. et al. *ApproxFun Julia package.* URL github.com/ApproxFun.

[Olv67] Olver, F.W.J. *Numerical solution of second-order linear difference equations.* J. Res. Nat. Bur. Standards Sect. B (1967). 71: 111–129.

[OT13] Olver, S. and Townsend, A. *A fast and well-conditioned spectral method.* SIAM Rev. (2013). 55, no. 3: 462–489.

[OT14] Olver, S. and Townsend, A. *A practical framework for infinite-dimensional linear algebra.* In *Proceedings of the 1st First Workshop for High Performance Technical Computing in Dynamic Languages.* IEEE Press (2014) 57–62.

[Par80] Parlett, B.N. *The symmetric eigenvalue problem*, vol. 7. SIAM (1980).

[Pro96] Proakis, J.G. *Digital signal processing: principles, algorithms, and applications* (1996).

[RS02] Rahman, Q.I. and Schmeisser, G. *Analytic theory of polynomials.* Oxf. Univ. Press (2002).

[RT92] Reichel, L. and Trefethen, L.N. *Eigenvalues and pseudo-eigenvalues of Toeplitz matrices.* Linear Algebra Appl. (1992). 162: 153–185.

[Saa03] Saad, Y. *Iterative methods for sparse linear systems.* SIAM (2003).

[Sim79] Simon, B. *Trace ideals and their applications*, vol. 35. Camb. Univ. Press (1979).

[Sle78] Slepian, D. *Prolate spheroidal wave functions, fourier analysis, and uncertainty—V: The discrete case.* Bell Labs Tech. J. (1978). 57, no. 5: 1371–1430.

[SM03] Süli, Endre and Mayers, David F. *An introduction to numerical analysis.* Cambridge university press (2003).

[Smi07] Smith, J.O. *Introduction to Digital Filters with Audio Applications.* W3K Publishing (2007). URL http://www.w3k.org/books/.

[SO17] Slevinsky, R.M. and Olver, S. *A fast and well-conditioned spectral method for singular integral equations.* J. Comput. Phys. (2017). 332: 290–315.

[Sti94] Stieltjes, T.-J. *Recherches sur les fractions continues.* In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, vol. 8 (1894) 1–122.

[Str14] Strogatz, S.H. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering.* Westview press (2014).

[Sym82] Symes, W.W. *The QR algorithm and scattering for the finite nonperiodic Toda lattice.* Phys. D: Nonlinear Phenom. (1982). 4, no. 2: 275–280.

[TBI97] Trefethen, L.N. and Bau III, D. *Numerical linear algebra.* 50. SIAM (1997).

[TE05] Trefethen, L.N. and Embree, M. *Spectra and pseudospectra: the behavior of nonnormal matrices and operators.* Princ. Univ. Press (2005).

[Tes00] Teschl, G. *Jacobi operators and completely integrable nonlinear lattices.* 72. Am. Math. Soc. (2000).

[Tes01] Teschl, G. *Almost everything you always wanted to know about the Toda equation.* Jahresber. Deutsch. Math.-Verein. (2001). 103, no. 4: 149–162.

[Tes16] Teschl, G. *Topics in Real and Functional Analysis* (2016). URL http://www.mat.univie.ac.at/~gerald/ftp/book-fa/.

[TO16] Trogdon, T. and Olver, S. *Riemann–Hilbert Problems, Their Numerical Solution and the Computation of Nonlinear Special Functions.* SIAM (2016).

[Tod67] Toda, M. *Vibration of a chain with nonlinear interaction.* J. Phys. Soc. Jpn. (1967). 22, no. 2: 431–436.

[TP91] Turk, M. and Pentland, A. *Eigenfaces for recognition.* J. Cognitive Neurosci. (1991). 3, no. 1: 71–86.

[Tre97] Trench, W.F. *Numerical solution of the inverse eigenvalue problem for real symmetric Toeplitz matrices.* SIAM J. Sci. Comput. (1997). 18, no. 6: 1722–1736.

[Tre00] Trefethen, L.N. *Spectral methods in MATLAB.* SIAM (2000).

[Tre08] Trefethen, L.N. *Is Gauss quadrature better than Clenshaw–Curtis?* SIAM review (2008). 50, no. 1: 67–87.

[Tre13] Trefethen, L.N. *Approximation theory and approximation practice.* SIAM (2013).

[Tuc11] Tucker, W. *Validated numerics: a short introduction to rigorous computations.* Princeton University Press (2011).

[VA90] Van Assche, W. *Asymptotics for Orthogonal Polynomials and Three-Term Recurrences*, 435–462. Springer Netherlands (1990).

[VA91] Van Assche, W. *Orthogonal polynomials, associated polynomials and functions of the second kind.* J. Comput. Appl. Math. (1991). 37, no. 1: 237–249.

[VA94] Van Assche, W. *Chebyshev polynomials as a comparison system for orthogonal polynomials.* In *Proceedings of the Cornelius Lanczos International Centenary Conference.* SIAM (1994) 365–367.

[VAG89] Van Assche, W. and Geronimo, J.S. *Asymptotics for orthogonal polynomials with regularly varying recurrence coefficients.* Rocky Mt. J. Math. (1989). 19.

[VBL+16] Vasil, G.M., Burns, K.J., Lecoanet, D., Olver, S., Brown, B.P. and Oishi, J.S. *Tensor calculus in polar coordinates using jacobi polynomials.* J. Comput. Phys. (2016). 325: 53–73.

[Wan01] Wang, T.-L. *Convergence of the tridiagonal QR algorithm.* Linear Algebra Appl. (2001). 322, no. 1-3: 1–17.

[Wat84] Watkins, D.S. *Isospectral flows.* SIAM Rev. (1984). 26, no. 3: 379–391.

[Wat07] Watkins, D.S. *The matrix eigenvalue problem: GR and Krylov subspace methods.* SIAM (2007).

[Wat08] Watkins, D.S. *The QR algorithm revisited.* SIAM Rev. (2008). 50, no. 1: 133–145.

[Weg12] Wegert, E. *Visual complex functions: an introduction with phase portraits.* Springer (2012).

[Wil65] Wilkinson, J.H. *Convergence of the LR, QR, and related algorithms.* Comput. J. (1965). 8, no. 1: 77–84.

[Zan98] Zanna, Antonella. *On the numerical solution of isospectral flows.* Ph.D. thesis, University of Cambridge (1998).