Topics in Spectral Theory

Publicações Matemáticas

Topics in Spectral Theory

Carlos Tomei
PUC-Rio

30º Colóquio Brasileiro de Matemática

Impresso no Brasil / Printed in Brazil
Capa: Noni Geiger / Sérgio R. Vaz

30º Colóquio Brasileiro de Matemática

- Aplicações Matemáticas em Engenharia de Produção - Leonardo J. Lustosa e Fernanda M. P. Raupp
- Boltzmann-type Equations and their Applications - Ricardo Alonso
- Dissipative Forces in Celestial Mechanics - Sylvio Ferraz-Mello, Clodoaldo Grotta-Ragazzo e Lucas Ruiz dos Santos
- Economic Models and Mean-Field Games Theory - Diogo A. Gomes, Levon Nurbekyan and Edgard A. Pimentel
- Generic Linear Recurrent Sequences and Related Topics - Letterio Gatto
- Geração de Malhas por Refinamento de Delaunay - Marcelo Siqueira, Afonso Paiva e Paulo Pagliosa
- Global and Local Aspects of Levi-flat Hypersurfaces - Arturo Fernández Pérez e Jiří Lebl
- Introdução às Curvas Elípticas e Aplicações - Parham Salehyan
- Métodos de Descida em Otimização Multiobjetivo - L. M. Graña Drummond e B. F. Svaiter
- Modern Theory of Nonlinear Elliptic PDE - Boyan Slavchev Sirakov
- Novel Regularization Methods for Ill-posed Problems in Hilbert and Banach Spaces - Ismael R. Bleyer e Antonio Leitão
- Probabilistic and Statistical Tools for Modeling Time Series - Paul Doukhan
- Tópicos da Teoria dos Jogos em Computação - O. Lee, F. K. Miyazawa, R. C. S. Schouery e E. C. Xavier
- **Topics in Spectral Theory - Carlos Tomei**

Distribuição: IMPA
Estrada Dona Castorina, 110
22460-320 Rio de Janeiro, RJ
E-mail: ddic@impa.br
http://www.impa.br

ISBN: 978-85-244-0413-9

Contents

1 Introduction **5**

1.1 Contents 7
1.2 Texts 8
1.3 Basic notation 8
1.4 Why spectral theory? 9

2 Some basic facts **11**

2.1 Linear transformations, matrices 11
2.2 Krylov spaces, companion matrices 16
2.3 Lanczos's procedure, Jacobi matrices 17
2.3.1 Jacobi inverse variables 19
2.4 Genericity and density arguments 20
2.4.1 The resultant 20
2.4.2 Density arguments 21
2.5 Tensors and spectrum 23
2.6 Wedges 26
2.7 Some applications 27
2.7.1 Roots of polynomials are eigenvalues 27
2.7.2 The resultant revisited 28
2.7.3 Algebraic numbers form a field 28
2.8 Some examples: adjacency matrices 29
2.8.1 Polygons and second derivatives 29
2.8.2 Regular polytopes 32
2.8.3 Semi-regular polytopes 36

3 Some analysis **39**
3.1 Algebras of matrices and operators 39
3.2 Smoothness of eigenpairs 42
3.2.1 Bi-orthogonality, derivatives of eigenpairs . . . 43
3.2.2 Continuity of eigenvalues 45
3.3 Some variational properties 45
3.4 Approximations of small rank 48
3.5 Isospectral manifolds 50
3.5.1 More isospectral manifolds 52
3.5.2 Two functionals on $\mathcal{S}_\Lambda$ 53

4 Spectrum and convexity **57**
4.1 The Schur-Horn theorem 57
4.2 Mutations, the high and low roads 65
4.3 Interlacing and more 66
4.3.1 Rank one perturbations 66
4.3.2 The sum of two Hermitian matrices 71
4.3.3 Weinstein-Aronsjan, Sherman-Morrison 71

5 The spectral theorem **74**
5.1 The Dunford-Schwartz calculus 74
5.2 Orthogonal polynomials 80
5.3 A quadrature algorithm 84
5.4 The spectral theorem — a sketch 86

Chapter 1

Introduction

Linear relations are unavoidable — one doubles the cause and the effect doubles, at least on a first guess. A substantial amount of the mathematics used in modeling is linear, which does not mean that it is trivial. The calculus of many variables, like optimizing in hundreds, millions of unknowns, is frequently a linear algebra problem. Besides, nonlinear problems are hard and substantial information may be obtained by linearization at points of interest.

It is not clear that this is how we approach the teaching of linear algebra however. Sometimes, students are introduced to the subject as a sophisticated version of analytic geometry, for essentially visual purposes. Few students have the opportunity of relating linear algebra to... all things linear. The task is considered in engineering courses, but rarely within mathematics departments.

There are historical reasons for this attitude. Linear algebra at some moment must have looked like the golden opportunity to present students to the axiomatic approach. Possibly the very first consequence of this point of view was the utter separation between linear and nonlinear theory: Jacobians and Hessians from Calculus hardly relate to matrices in linear algebra courses.

Then there were the computational difficulties — a 3×3 matrix should have a simple eigenvalue or somehow it is inappropriate to fit in an exercise. Galois theory provides one of the most interesting no-go theorems in mathematics: radicals and the usual arithmetic

symbols are not sufficient to write down the solutions of a polynomial of degree 5 with integer coefficients. Still, this is not the end of the world — one might invent new symbols or simply live with arbitrarily good approximations. Few students (possibly few professionals) are conscious of the fact that most real numbers cannot even be described, a simple cardinality argument.

And worse, among the important concepts in linear algebra lie ... nonlinear objects, eigenvalues, functions and groups of matrices. Derek Hacon, a colleague from PUC-Rio, used to say that fiber bundle theory is linear algebra with parameters. Few math students know how to take a derivative of an eigenvalue $\lambda(t)$ of a matrix $M(t)$. The standard analysis course in $\mathbb{R}^n$ interacts poorly with matrix theory.

The highlights from last century — quantum mechanics [1], the role of spectral theory in pure and applied dynamical systems; the numerical analysis of differential equations; the spectral theory of graphs, or in a more general setting, numerical linear algebra as a whole, being confronted with larger symmetric and non-symmetric matrices — indicate a combination of theory, practice and technology which should be a source of enthusiasm to any mathematician. Just to stick to one inevitable example, any Google search is a very large (numerical) problem in spectral graph theory.

There is so much to choose from, what should be said in five short lectures? The topics intend to stimulate the interaction among different disciplines within mathematics, having in mind a public of graduate students. There are arguments involving algebra, basic real analysis, geometry, some complex variable, a bit of measure theory, a couple of algorithms, differential equations... and extensive pointers to more sophisticated material in algebraic topology, symplectic geometry, numerical and functional analysis.

Alas, everything is deterministic: there is nothing about random matrices or eigenvalue distributions. Also, there is nothing about the integrable systems associated to spectral theory.

Acknowledgements abound. The Departamento de Matemática at PUC-Rio allowed me to teach a number of courses in these subjects, and students from different departments contributed with a large

[1] According to Reed and Simon ([45]), the fact that the point spectrum of the Schrödinger operator of the hydrogen atom describes with spectacular precision the frequencies of its emission spectrum borders on the scientifically embarrassing.

spectrum of opinions. Some colleagues are friends and mentors — Percy Deift, Charlie Epstein, Nicolau Saldanha. Peter Lax, Beresford Parlett, Barry Simon are sources of inspiration. Years of subsidies from CNPq, CAPES and FAPERJ are also gratefully acknowledged.

1.1 Contents

There are threads across the text. Chapter 2 contains some basic algebraic constructions which are extensively used. The fact that matrices form an algebra lead to cyclic vectors, companion and Jacobi matrices, inverse variables for Jacobi matrices, an introduction to orthogonal polynomials and eventually an indication of how the spectral theorem for self-adjoint operators in infinite dimension follows from its counterpart for tridiagonal matrices. Tensor products and their spectral properties give rise to the resultant, the SVD decomposition, small rank approximations of symmetric and nonsymmetric matrices. There is frequent interplay between the invariant formalism and the use of coordinates. From the very start, density arguments are used to simplify proofs: matrices with distinct eigenvalues are usually easier to handle.

The study of eigenvalues and eigenvectors as objects which depend smoothly on matrices extends to some basic geometry of isospectral manifolds, which in turn are presented as natural phase spaces of algorithms for the computation of spectrum. Other geometric methods in the study of eigenvalues are exemplified by standard results — the Schur-Horn theorem, spectral interlacing — for which elementary proofs are given, as opposed to the current symplectic approach.

The standard road to the spectral theorem is the construction of a powerful functional calculus. Since good presentations are available, we decided instead to stop along the road, in particular the Dunford-Schwartz calculus, for better appreciation of some details.

The text contains a number of references for further study. And this is perhaps the real motivation for these notes: to convince the reader of the vitality of the subject.

1.2 Texts

There are periodicals, libraries, dedicated to the subject. Here I just quote a few classics, of great mathematical and pedagogical value. The basic spectral theory of differential operators is exquisitely described in the books by Reed and Simon ([45]) and Kato ([28]). Great functional analysis texts are Dunford-Schwartz ([19]), Lax ([33]) and for linear algebra, [34] and [27]. From the numerical analysis literature I choose a minimal sample, Wilkinson ([64]), Parlett ([44]) and Trefethen and Embree, as a source to more recent material ([63]).

1.3 Basic notation

Some sets are just too frequent. Let $\mathcal{M}(n, \mathbb{K})$ denote the vector space of $n \times n$ matrices with entries in the field $\mathbb{K}$ (usually $\mathbb{R}$ or $\mathbb{C}$). Similarly $\mathcal{S}(n, \mathbb{K})$ and $\mathcal{A}(n, \mathbb{K})$ will denote symmetric and skew symmetric matrices for $\mathbb{K} = \mathbb{R}$, and Hermitian and skew Hermitian matrices for $\mathbb{K} = \mathbb{C}$. The reader should get used to dropping a few indices and dimensions: GL is the group of invertible matrices, SO, the real orthogonal matrices with determinant equal to 1 — the context will naturally specify the dimension and the underlying field.

There is a systematic, irresistible, abuse of notation which we accept as a fact of life. An $n \times n$ matrix M has n eigenvalues λ_i counted with multiplicity. Thus, the appropriate concept to describe the *spectrum* of a matrix is a *multiset*, i.e., a set on which repetitions of elements are allowed and give rise to different objects. We avoid such considerations and refer to the spectrum $\sigma(M)$ as

$$\sigma(M) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\},$$

where eigenvalues may be equal, and then they should be repeated in the list. The order in which the eigenvalues are presented, i.e. their labeling, is an irrelevant matter. In particular, one cannot in principle talk about the first eigenvalue of a matrix.

1.4 Why spectral theory?

This section is not about the contributions of spectral theory to science in general and mathematics in particular. This is a very specific example, an excellent starting point for a second course in linear algebra. *Leslie models* are frequently used in biology: they are concerned with the evolution of a population divided in age groups ([39]). Say, for example, that a population splits in $n+1$ groups of ages $0-1, 1-2, 2-3, \ldots, n-\infty$ and consider the simplest possible model for demographic variation. Thus, for example, for $n=4$, one might consider a transition matrix

$$M = \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \\ s_0 & 0 & 0 & 0 & 0 \\ 0 & s_1 & 0 & 0 & 0 \\ 0 & 0 & s_2 & 0 & 0 \\ 0 & 0 & 0 & s_3 & s_4 \end{pmatrix}$$

relating the population vector between consecutive (integer) times,

$$p(t+1) \;=\; M\, p(t)\,.$$

The α_i's are *fertility* rates, indicating the contribution of each age group to newborns. The s_i's instead are *survival* rates, which again may vary among age groups.

One merit of this simple model is that it may be described graphically, with boxes and weighted arrows representing the information coded in the matrix. Another is its versatility: it allows for sexual distinctions and immigrations, which would introduce a non-homogeneous term. The students may develop a concrete feeling for the values of the parameters.

The natural question is: given M, what happens in the long run to a population? Things get especially theatrical if the class is facing computers (and why isn't it?!). Each student enters with a different initial population and, by comparing results with the teacher's choice, the following beautiful fact comes through.

Theorem 1. *For an open, dense set of positive parameters, $p(n)$ for n large is essentially given by $c\lambda_M^n v$, where $c \in \mathbb{R}$ is a fixed number, λ_M is the eigenvalue of M of largest module and v is a normalized eigenvector associated to λ_M.*

The vector v is the *pyramid distribution* of the population if it is normalized so as to have its components adding to one: in the long range, it is essentially independent of the initial population! A *single* eigenvalue specifies if the population increases (exponentially) or faces extinction. More, in order for the model to make sense, one is forced to believe that the eigenvalue of largest module of M is a positive number, and has an eigenvector with nonnegative entries — after all, the pyramid distribution consists of fractions of the population. A more experienced practitioner would identify these last statements as consequences of the celebrated Perron-Frobenius theorem.

Thus, not only eigenvalues and eigenvectors come up as the natural vocabulary to answer an interesting question, but the key information is concentrated in very little data of that kind — one seldom cares for many eigenvalues and eigenvectors on a realistic model. What may happen is that the relevant eigenvalues have different geometric properties: largest real part (recall that they might be complex numbers), or they should belong to a certain set in the complex plane. By the way, no symmetric matrices were required in the example.

Chapter 2

Some basic facts

2.1 Linear transformations, matrices

Consider the following two versions of the matrix spectral theorem.

Theorem 2. *Every real, symmetric matrix S may be written as a product $S = Q^T \Lambda Q$, where $Q \in SO$ and Λ is a real diagonal matrix.*

Theorem 3. *Let V be a finite dimensional vector space over $\mathbb{R}$ endowed with an inner product. Then every symmetric linear transformation $T : V \to V$ admits a basis of orthonormal eigenvectors.*

First, are we talking about the same theorem? For starters, the word *symmetry* here plays different roles. We all know what a symmetric matrix is, and a symmetric transformation is still quite familiar — for any vectors $u, v \in V$, one should have

$$\langle T\,u\,,\,v\,\rangle \;=\; \langle\,u\,,\,T\,v\,\rangle\,.$$

Now, we also know that linear transformations incarnate in matrices, once bases are chosen in the domain and counterdomain. Here we have to be careful: the subjects in which (finite dimensional) spectral theory is relevant require the same choice of vector space for both roles, and this strongly suggests that only one choice can be made.

Indeed, if M represents a transformation $T : V_1 \to V_2$ for choices of two bases, one for V_1, the other for V_2, and then other two bases

are chosen, the new matrix $\tilde{M}$ representing the same T satisfies

$$\tilde{M} \;=\; P\,M\,R$$

for two invertible matrices P and R — and these matrices arbitrary, once invertibility is preserved. This flexibility may be used to convert T into a matrix M which consists only of zeros and ones, so that the ones are along the diagonal entries — in particular, all eigenvalues of such an M are zeros and ones. Said differently, there is only one number which is left of T when all possible matrix representations are considered: its rank.

On the other hand, if the same basis is chosen on V_1 and V_2, then matrices M and $\tilde{M}$ are related by

$$\tilde{M} \;=\; P^{-1}\,M\,P$$

and we know that conjugation does not change the eigenvalues of a matrix (besides changing the eigenvectors in a controlled fashion).

Also, one might start with a symmetric transformation $T : V \to V$ and get to a non-symmetric matrix. A possible amend is choosing a common orthonormal basis (by the way, an *orthogonal* basis suffices).

Exercise 1. Under these requirement, M is indeed a symmetric matrix. Changes of bases between orthonormal bases give rise to orthogonal matrices $Q \in SO$, so that

$$\tilde{M} \;=\; Q^{-1}\,M\,Q \;=\; Q^{T}\,M\,Q$$

and matrix symmetry is preserved.

The first version of the spectral theorem above essentially states that eigenvalues are the only invariant information, once all matrix representations of this form are allowed. Actually, it says another simple fact: by writing $MQ^T = Q^T\Lambda$ and comparing columns, we see that the columns of Q^T are eigenvectors associated to eigenvalues which sit along the diagonal of Λ.

The two versions suggest different proofs of the spectral theorem. The invariant version, i.e., the one about linear transformations, is usually proved by induction on the dimension: once an eigenvector v is found, restrict T to the invariant subspace given by the orthogonal

complement of v, on which, trivially, T is still symmetric. This last statement does not convert easily to matrix representations.

Thinking of a matrix as a box of numbers leads to different approaches — a paradigmatic example is Gaussian elimination for solving linear systems. A proof of the spectral theorem from this point of view follows from Jacobi's algorithm, described in Section 3.5.2.

One might think that for numerical purposes one should stick to matrix representations, but this would be an exaggeration. Many algorithms in numerical linear algebra are based on the assumption that the only manifestation of the matrix M is through a routine that computes the value Mv for an input vector v.

We get back to a more conceptual standpoint. Real, symmetric matrices are diagonalizable and all matrices admit a Jordan form. One might be tempted to search for other hypotheses implying diagonalizability. We know, for example, that *normal* matrices $M \in \mathcal{M}(n, \mathbb{C})$, for which

$$M\,M^* \;=\; M^*\,M\,, \quad \text{where} \quad M^* = \overline{M}^T\,,$$

are diagonalizable. From the spectral theorem for normal matrices, $M = U^*\,\Lambda U$, where now U is a *unitary* matrix (i.e., $UU^* = I$) and Λ is a diagonal matrix with possibly complex entries.

But there is still a substantial gap between normal and arbitrary matrices. The invariant point of view is especially convenient to handle this issue. Say $M = P\Lambda P^{-1}$: think of the columns of the invertible matrix P as orthonormal vectors *for some inner product* of $\mathbb{C}^n$. In other words, *define* the inner product in $\mathbb{C}^n$ by requiring that the columns of P are orthonormal. The invariant point of view then forces you to believe the following result, which closes the gap above.

Proposition 1. *Any diagonalizable matrix is normal for some appropriate inner product.*

Let us be clear about the statement of the theorem. For a complex vector space V with an inner product, there is an appropriate generalization of transposing a matrix: a linear transformation $T : V \to V$ has a unique *adjoint* $T^* : V \to V$, defined by

$$\langle\, u\,, T\,v\,\rangle \;=\; \langle\, T^*\,u\,, v\,\rangle\,, \quad \forall\; u\,,\, v \;\in\; V\,.$$

In invariant terms, a *normal* transformation satisfies $TT^* = T^*T$. This definition also defines normality for an $n \times n$ complex matrix, where we think of $\mathbb{C}^n$ being endowed with an inner product.

We finish with another confrontation of the two points of view. Some people find Sylvester's law of inertia nonintuitive.

Theorem 4 (Sylvester)**.** *Let S be a real, $n \times n$ symmetric matrix, and take an $n \times n$ real invertible P. Then S and PSP^T have the same number of positive, zero and negative eigenvalues.*

The eigenvalues of both matrices would be equal for an orthogonal matrix P. Now, consider a smooth n-manifold $\mathcal{M}$ and a smooth function $f : \mathcal{M} \to \mathbb{R}$ with a critical point m. Following the reflex inherited from calculus, once the derivative at a point is zero, we search for the sign of the second derivative there (in order to classify the critical point, as they say). In this case, we have to compute the *Hessian* $S = Hf(m)$ of f at m, and different charts would give rise to different symmetric matrices representing the Hessian.

Patient use of the chain rule shows that the Hessian associated to two different charts are of the form S and PSP^T (and what is P, untiring reader?). Say f represents the temperature on the surface of a planet: the fact that, say, m is a local minimum is independent of the chart being used, so Sylvester's law is inevitable — the signs of the eigenvalues of the Hessian are invariants. They provide the so called *signature* at a critical point of a Morse function.

One should neither overestimate spectrum, nor underestimate coordinates. As a final comment, in many variable calculus one frequently classifies local extrema by computing eigenvalues of the Hessian, which boils down to diagonalizing it. Sylvester's law, or better, one of its computational implementations, the LDL^T decomposition (or, closely, the Cholesky's decomposition), can be actually performed, unlike the computation of eigenvalues — it is an extended version of the celebrated 'complete the square' trick. Thus, for example, for $H = LDL^T$ as below,

$$H = \begin{pmatrix} 2 & 4 & -2 \\ 4 & 5 & -7 \\ -2 & -7 & 1 \end{pmatrix} = L \begin{pmatrix} 2 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 2 \end{pmatrix} L^T,$$

for

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix}.$$

For $v = (x, y, z)$ the quadratic form

$$\langle H\, v\, , v \rangle = \langle\, D\, L^T\, v\, , L^T\, v \rangle$$

becomes

$$2x^2 + 8xy - 4xz + 5y^2 - 14yz + z^2 = 2X^2 - 3Y^2 + 2Z^2$$

for $(X, Y, Z) = L^T\, v = (x + 2y - z, y + z, z)$.

Matrix factorizations are so natural that they are incorporated in the more invariant Lie group theory ([18]).

Exercise 2. This factorization counts the positive eigenvalues of an invertible real symmetric matrix S. How would you count the number of eigenvalues of S in an arbitrary closed interval whose endpoints are not in $\sigma(S)$? This simple idea gives rise to a *bisection method* to compute eigenvalues. It is rather cumbersome, but robust ([44]).

A linear transformation $T : V \to W$ between vector spaces V and W (over the same field $\mathbb{R}$ or $\mathbb{C}$) gives rise to a matrix M once bases are chosen for V and W. In particular, taking the canonical bases for $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, the entries of M are the numbers

$$M_{ij} \;=\; \langle\, e_i\, ,\, M\, e_j\, \rangle,$$

for the usual inner product in $\mathbb{R}^m$. We define *generalized entries*

$$T_{vw} \;=\; \langle\, w\, ,\, T\, v\, \rangle \quad v \,\in\, V\, ,\, w \in W,$$

for some inner product in W. If X is a Banach space and $T : X \to X$ is a linear bounded map, consider

$$T_{vw} \;=\; w(\, T\, v\,), \quad v \,\in\, X\, ,\, w \,\in\, X^*,$$

where X^* is the dual space of X. When X is a Hilbert space, we are back to the previous definition, by the Riesz representation theorem.

Generalized entries sometimes are eigenvalues — this is what diagonalization is all about. But this happens away from the diagonalizable context — an example is given by the variational interpretation of the eigenvalues and of their indispensable and underrated cousins, the *singular values* of a transformation (Section 3.3).

Frequently in descriptions of the (infinite dimensional) spectral theorem, one performs operations on a transformation T by specifying what goes on with these many scalars T_{uv}. The practice is common among (quantum) physicists — generalized matrix entries include the so called *observables* of a system. But this is another story.

2.2 Krylov spaces, companion matrices

As we all know, a linear transformation is determined by its action on a basis, a fundamental fact of linear algebra. We should give more thought, as algebraists do, to the possibility of a more appropriate concept when we deal with square matrices, which correspond to linear transformations from a space to itself — the *multiplication structure* (composition, for transformations) might be exploited. To simplify matters, assume all vector spaces to be real, even if eigenvalues might be complex occasionally.

Let $T : V \to V$ be a linear transformation from a vector space V to itself and consider $v \in V$. The *Krylov subspace* $K(T, v) \subset V$ is the space generated by the vectors $T^k v, k = 0, 1, \ldots$. Clearly, once a vector $T^{k_0} v$ is a combination of the previous ones, the same happens to the subsequent vectors. The set $\{v, Tv, \ldots, T^{k_0-1} v\}$ forms a basis $B(T, v)$ for $K(T, v)$, which is clearly an invariant subspace of T. If V is finite dimensional and $K(T, v) = V$, then v is a *cyclic vector*.

Exercise 3. Suppose V is of dimension $n < \infty$ and $T : V \to V$ has simple spectrum (i.e., all its eigenvalues are distinct). Show that a vector $v \in V$ is not cyclic if and only if it is a linear combination of less than n eigenvectors of T. If instead T has a basis of eigenvectors and some double eigenvalue, then there is no cyclic vector.

Say V is of dimension $n < \infty$ with a cyclic vector v. Endow domain and counterdomain with the basis $B(T, v)$ and the matrix associated to T is a *companion matrix*, which for $n = 5$ looks like

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 & -c_0 \\ 1 & 0 & 0 & 0 & -c_1 \\ 0 & 1 & 0 & 0 & -c_2 \\ 0 & 0 & 1 & 0 & -c_3 \\ 0 & 0 & 0 & 1 & -c_4 \end{pmatrix} .$$

and whose characteristic polynomial is

$$\det(M - \lambda I) = \lambda^5 + c_4\lambda^4 + c_3\lambda^3 + c_2\lambda^2 + c_1\lambda^1 + c_0 \, .$$

This used to be the standard method to compute the characteristic polynomial of a matrix M. The usual algorithm to compute determinants by expanding along a line requires essentially $n!$ multiplications for an $n \times n$ matrix, while the process above takes something of order n^3, depending of your favorite way of solving a linear system (so as to expand $T^n v$ in the basis $B(T, v)$). The process would be applied to reduce an arbitrary matrix M to a companion form (or even simpler, if the vector v happened not to be cyclic!), from which one could search for eigenvalues by solving for the roots of a polynomial.

2.3 Lanczos's procedure, Jacobi matrices

If V is a real finite dimensional Hilbert space and $T : V \to V$ is a bounded symmetric operator, one might wish to obtain a real, symmetric matrix to represent T, but this is not expected for a basis $B(T, v)$. Probably, the first thing that comes to mind to circumvent this difficulty is submitting $B(T, v)$ to the Gram-Schmidt orthonormalization process, obtaining an orthonormal basis

$$GS(T, v) = \{v_0, v_1, \ldots, v_{n-1}\}$$

(say v is cyclic, to simplify matters). It is clear that that each v_k is a linear combination of the first $k+1$ vectors in $B(T, v)$ (said differently, both bases are related by a triangular matrix). The representation of T in the basis $GS(T, v)$ in principle is a so called *Hessenberg matrix*, a matrix whose entries (i, j) with $i - j > 1$ are all equal to zero. Here is the pattern of zeros of a 5×5 Hessenberg matrix.

$$H = \begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix} .$$

On the other hand, this matrix should also be symmetric: after all, this is why we decided to use the basis $GS(T, v)$. The upshot is that T in the $GS(T, v)$ basis is a *real, symmetric, tridiagonal matrix.*

One can do slightly more: because of cyclicity, the entries (i, j) with $i = j + 1$ can be shown to be nonzero, and even more, they are strictly positive numbers. Thus, T in this case is represented by a so called *Jacobi* matrix. If $n = 5$,

$$J = \begin{pmatrix} a_0 & b_0 & 0 & 0 & 0 \\ b_0 & a_1 & b_1 & 0 & 0 \\ 0 & b_1 & a_2 & b_2 & 0 \\ 0 & 0 & b_2 & a_3 & b_3 \\ 0 & 0 & 0 & b_3 & a_4 \end{pmatrix} ,$$

where $b_k > 0$.

Let us write the computations above in matrix form. Say T is $n \times n$, with a cyclic vector v_0 giving rise to an orthonormal basis $GS(T, V_0) = v_0, v_1, \ldots, v_{n-1}$. Let W be the (orthogonal) matrix with columns given by the v_k's.

Proposition 2. *W tridiagonalizes T by conjugation,*

$$T\,W = W\,J\,, \qquad \textit{so that} \qquad J = W^T\,T\,W\,,$$

$$a_0 = \langle\, v_0\,, T\,v_0\,\rangle\,, \quad a_k = \langle\, v_k\,, T\,v_k\,\rangle\,, \quad b_{k-1} = \langle\, v_{k-1}\,, T\,v_k\,\rangle \quad k \geq 1\,.$$

We extend the result.

Theorem 5. *Let H be a Hilbert space, $T : H \to H$ a bounded symmetric operator. Then there is a unitary transformation $U : H \to H$ such that $\tilde{J} = U^* T U$ splits into Jacobi blocks.*

The operator T may induce infinitely many blocks. Thus H splits in a sum of orthogonal invariant subspaces

$$H = \text{closure}\,\big(\bigoplus_{\alpha} H_\alpha\big)\,,$$

(only countably many, if H is separable) and on each such H_α there is a cyclic vector v_α, in the sense that H_α is the closure of the span of the vectors $T^k v_\alpha, k = 0, 1, \ldots$.

Proof. Take any nonzero $v \in H$ to obtain the first invariant subspace, given by $\overline{span(T^k v)}$. Now proceed by taking another vector in the orthogonal complement of this subspace. This works if H is finite dimensional, otherwise use Zorn's lemma. □

This is one of the reasons why tridiagonal (in particular Jacobi) matrices are relevant. In particular, the spectral theorem for bounded symmetric operators follows once it is proved for Jacobi matrices.

Exercise 4. No eigenvector v of a Jacobi matrix has its first coordinate v_1 equal to zero (write $Jv = \lambda v$ in coordinates). The spectrum of a Jacobi matrix is always simple (i.e., all eigenvalues are distinct).

The *Lanczos method* tridiagonalizes a symmetric matrix S by splitting it in a sum of Jacobi matrices J_k so that $\cup_k \sigma(J_k) = \sigma(S)$. It is used, for example, in the conjugate gradient algorithm.

2.3.1 Jacobi inverse variables

What does it take do describe an $n \times n$ Jacobi matrix J? Say J has the same eigenvalues of a *diagonal* matrix Λ with simple spectrum (this is automatic, from Exercise 4) and write $J = W^T \Lambda J$, where again by the same exercise, we may suppose that the first row of W^T (hence, the first column of W) has strictly positive entries. Clearly, knowing J is equivalent to knowing W, which in turn is obtained from the cyclic vector v_0 from the Lanczos procedure applying successively Λ to v_0. The entries of v_0 are the first coordinates of the eigenvectors of J, which, from exercise 4, may be taken to be strictly positive. Also, since W is orthogonal, the vector v_0 must be normal.

Thus, Jacobi matrices are described by eigenvalues (yielding Λ) and *norming constants* c_k, which are the entries of v_0 — these are the so called *Jacobi inverse variables*. In a sense we are tridiagonalizing a diagonal matrix Λ, and this allows for the additional parameters c_k: we provide details. Define two natural geometric objects,

$$\mathbb{R}^n_+ = \{x \in \mathbb{R}^n \,|\, x_1 > x_2 > \ldots > x_n\},$$

$$Q_+^n = \{ c \in \mathbb{R}^n \mid c_k > 0 \text{ and } \sum_k c_k^2 = 1 \}.$$

Theorem 6. *The map taking a Jacobi matrix J to its ordered spectrum and norming constants is a diffeomorphism to $\mathbb{R}_+^n \times Q_+^n$.*

Moser stated this result as a discrete analogue of the inverse scattering variables for the Schrödinger equation in the line. In the continuous case, such variables were used to linearize the celebrated Korteweg-de Vries equation. Moser used their discrete counterpart to essentially linearize the Toda lattice ([42]). He credits the result to Stieltjes and gives a different proof from the one presented here.

Proof. The map $J \mapsto (\Lambda, c)$ is smooth (even real analytic): Jacobi matrices for an open set of the vector space of symmetric, tridiagonal matrices with simple spectrum, so that eigenvalues and eigenvectors vary smoothly (from Proposition 10 in Section 3.2). The inverse map $(\Lambda, c) \mapsto J$, from which the rest of the statement follows immediately, is just the Lanczos procedure starting from Λ and c. □

2.4 Genericity and density arguments

It is very frequent that a statement about matrices is simpler to prove if we add some generic hypothesis. The additional hypothesis is then removed by taking limits. Let us provide examples of this technique, which will be used extensively in this text.

2.4.1 The resultant

The *resultant* of two polynomials is one of those algebraic jewels everybody should know. Say $p(x)$ and $q(x)$ have degrees n and m. If they have a common root r, their greatest common divisor (gcd) has $(x - r)$ as a factor. On the other hand, if they are mutually prime, their gcd is 1 and from the Euclidean algorithm, there are polynomials $a(x)$ and $b(x)$ with degrees at most $m - 1$ and $n - 1$ for which

$$a(x)\, p(x) \, + \, b(x)\, q(x) = 1 \,.$$

The coefficients of a and b can be obtained from this equation by solving a system with $m + n$ unknowns and $m + n$ equations — its

determinant $R(p,q)$ then is zero if and only if p and q are mutually prime. Clearly $R(p,q)$ is a polynomial in the coefficients of p and q. In particular, p has a double root if and only if $R(p,p')=0$.

The next step is more interesting: the resultant provides a nonlinear version of Gaussian elimination. Indeed, consider

$$p(x,y)=0\,,q(x,y)=0\,.$$

In the linear case, we solve for one variable and replace it in the other equation. Here, we compute the resultant $R(p,q;y)$ (think of $x=x_0$ fixed and compute the resultant of two polynomials in y) to obtain a polynomial $\tilde{R}(x)$ which is zero exactly when $p(x_0,.)$ and $q(x_0,.)$ have a common root y_0. So, the roots of $\tilde{R}(x)$ are exactly the values of x_0 for which one obtains common roots y_0 of p and q! Clearly the procedure holds for larger systems — get rid of a variable per step!

A very natural construction of the resultant is presented in Section 2.7.2. It uses a basic fact of spectral theory of tensor products.

2.4.2 Density arguments

As usual, let $\mathcal{M}(n,\mathbb{K})$ denote the algebra of $n\times n$ matrices with entries in the field $\mathbb{K}$ (which is either $\mathbb{R}$ or $\mathbb{C}$). Here are two generic properties of matrices which are commonly used. As usual, $GL(n,\mathbb{K})$ is the set of invertible matrices. Let $\mathcal{M}_d$ be the set of matrices with distinct eigenvalues (i.e., of simple spectrum).

Proposition 3. *$GL(n,\mathbb{K})$ and $\mathcal{M}_d$ are open, dense sets of $\mathcal{M}(n,\mathbb{K})$.*

Proof. If there is a nontrivial ball $B\subset\mathcal{M}(n,\mathbb{K})$ in which all matrices are not invertible, then the determinant function $\det(M)$, which is a polynomial in the entries of M, is identically zero throughout $\mathcal{M}(n,\mathbb{K})$, which is not true.

Similarly if in such a ball B all matrices have a double eigenvalue, then the resolvent $R(p(M),p'(M))$ between $p(M)=\det(M-\lambda I)$ and its derivative in λ, $p'(\lambda)$, is also zero in B — again, this expression is a polynomial in the entries of M and is not identically zero, since there are matrices with simple spectrum.

Openness of both sets is trivial: det and R are continuous maps. □

Exercise 5. Sometimes, genericity is simply not true: this exercise is harder. Let A and B be real, skew-symmetric matrices (so that $A^T = -A$ and $B^T = -B$). Then the product AB never has a simple eigenvalue. More, eigenvalues of AB always have even multiplicity. Hint: learn about Pfaffians.

Let H be a separable infinite dimensional Hilbert space. Within $\mathcal{B}(H)$, the algebra of bounded operators from H to itself, invertible operators form an open subset, but they are not dense. This role is taken by *Fredholm operators*: they somehow subsume rectangular matrices (how can one distinguish $\mathbb{R}^N$ and $\mathbb{R}^n$ in infinite dimensions?). We don't handle such issues in this text ([34] is a beautiful reference).

We now provide an archetypical density argument.

Proposition 4. *Let A and B be square matrices of the same dimension. Then $\sigma(AB) = \sigma(BA)$. If A is $n \times N$ and B is $N \times n$, for $N > n$, then $\sigma(AB) \setminus \{0\} = \sigma(BA) \setminus \{0\}$.*

This result holds for appropriate closed operators (in the rectangular version...) — see [12] for a proof and some very interesting applications. Actually, one can even obtain the celebrated *KdV solitons* using this result ([15]).

Proof. Suppose A is invertible: then the result is trivial:

$$AB = A\,(BA)\,A^{-1}\,.$$

Take an arbitrary square matrix A, and $A_n \to A$, where the A_n's are invertible. Then $\sigma(A_n\,B) = \sigma(B\,A_n)$, which we rewrite as

$$\det(A_n\,B - \lambda\,I) = \det(B\,A_n - \lambda\,I)\,.$$

The coefficients of the polynomials on both sides are continuous functions of the input matrices, so equality is preserved in the limit.

In general, obtain $N \times N$ matrices $\tilde{A}$ and $\tilde{B}$ from A and B by adding blocks of zeros and apply the result for $\tilde{A}$ and $\tilde{B}$. □

2.5 Tensors and spectrum

There is nothing wrong with thinking of vectors in $\mathbb{R}^n$ as strings of n real numbers, but sometimes this is simply not convenient. Consider the following important example — solve for u in

$$\Delta u(x,y) = u_{xx}(x,y) + u_{yy}(x,y) = f(x,y)\,, \quad u = 0 \,\text{in}\ \ \partial R$$

for $(x,y) \in R = (0,\,(n+1)\,h) \times (0,(m+1)\,h)$. A good starting point for this enormous subject is [16], here we just point out a relevant issue and bifurcate. Define the grid of equally spaced

$$(x_i, y_j) = (i,j)\,, \quad i = 1\,,\ldots,\,n \quad i = 1\,,\ldots,\,m\,.$$

The unknowns $\hat{u}_{ij}$ are the approximations of $u(x,y)$ on these points. Frequently the Laplacian Δ is approximated by

$$\Delta u(x_i,y_j) \sim \frac{1}{h^2}\left(\hat{u}_{i-1,j} - 2\hat{u}_{ij} + \hat{u}_{i+1,j} + \hat{u}_{i,j-1} - 2\hat{u}_{ij} + \hat{u}_{i,j+1}\right),$$

and one solves the linear system for $\hat{u}_{ij}$,

$$\frac{1}{h^2}\left(\hat{u}_{i-1,j} + \hat{u}_{i+1,j} + \hat{u}_{i,j-1} + \hat{u}_{i,j+1} - 4\hat{u}_{ij}\right) = f_{ij}\,,$$

where $\hat{u}$ is taken to be zero in the grid points in ∂R (i.e., $i = 0, n+1$ or $j = 0, m$) and f_{ij} is the value of f on grid point (x_i, y_j).

Suppose that there are n and m points in the interior of R on each horizontal or vertical lines of the grid, respectively. The unknown is naturally a vector in $\mathbb{R}^{nm}$ and the matrix L associated to the discrete Laplacian is of dimension $nm \times nm$. Still, the unknown is not really a string of numbers, it is a *box* of numbers, precisely, an association of a number to each grid point. If we think of the unknown $\hat{U}$ as an $m \times n$ matrix , L is given by $A\hat{U} + \hat{U}\,B$, where A is $m \times m$ and B is $n \times n$: A and B are discretizations of the second order (one dimensional) derivative, respectively along the y and the x axes.

It is true that A and B have much less entries than L, but a numerical analyst would argue that $L\hat{u}$, the discrete counterpart of Lu, should not be computed by writing the discrete Laplacian as a matrix: one might simply obtain Lu by using the discretization formula above. So clever programming circumvents this issue — but how

do you realize that there is a similar clever programming associated to the linear transformation of another problem? The Fast Fourier Transform, for example, is essentially a trick using tensor products, combined with exquisite programming ([46]).

There is more to this representation of L. As usual, $\mathcal{M}_{rs}$ is the vector space of real $r \times s$ matrices.

Theorem 7. *Let $A \in \mathcal{M}_{nn}$ and $B \in \mathcal{M}_{mm}$ have eigenvalues*

$$\alpha_i, i = 1, \ldots, n \quad \textit{and} \quad \beta_j, j = 1, \ldots, m\,.$$

The linear transformations

$$T_s, T_p : \mathcal{M}_{nm} \to \mathcal{M}_{nm}\,, \quad T_s M = A\,M - M\,B\,, \quad T_p(M) = A\,M\,B$$

have spectra given by

$$\sigma(T_s) = \{\, \alpha_i \,-\, \beta_j \,\}\,, \quad \sigma(T_p) = \{\, \alpha_i\,\beta_j \,\}\,,$$

for $i = 1, \ldots, n\,, \quad j = 1, \ldots, m.$

Thus, one obtains the spectrum of the Laplacian with Dirichlet conditions on a rectangle from the spectrum of the second derivative acting on functions which satisfy Dirichlet conditions in an interval, both in the continuum and discrete cases (use Exercise 10).

Proof. Suppose that A and B have simple spectrum. Fix eigenvalues and eigenvectors

$$A\,v_i = \alpha_i\,v_i\,, \quad B^T\,w_j = \beta_j\,w_j$$

and define the matrix $Z_{ij} = v_i\,w_j^T$. We have

$$T_s Z_{ij} = A\,v_i\,w_j^T \,-\, v_i\,w_j^T\,B = (\alpha_i \,-\, \beta_j)\,v_i\,w_j^T = (\alpha_i \,-\, \beta_j)\,Z_{ij}\,,$$

$$T_p Z_{ij} = A\,v_i\,w_j^T\,B = \alpha_i\,\beta_j\,v_i\,w_j^T = \alpha_i\,\beta_j\,Z_{ij}\,,$$

so not only we computed eigenvalues but eigenvectors of T_s and T_p. We still have to show that the Z_{ij}'s are independent. A possibility

is the following: take bases $\{\tilde{v}_k\}$ and $\{\tilde{w}_\ell\}$ so that $\langle \tilde{v}_k \, , v_i \rangle = \delta_{ki}$ and $\langle \tilde{w}_\ell \, , w_j \rangle = \delta_{\ell j}$. If

$$\sum_{ij} c_{ij} Z_{ij} = \sum_{ij} c_{ij} v_i \, w_j^T = 0 \, ,$$

multiply the linear combination on the left by $(\tilde{v}_i)^T$ and on the right by $\tilde{w}_j$ to conclude that $c_{ij} = 0$.

For arbitrary A and B, use a density argument (Section 2.4.2). □

Bases $\{\tilde{v}_k\}$ and $\{\tilde{w}_\ell\}$ are *bi-orthogonal*: they will come up again in Section 3.2.1. Notice that rows of a matrix and columns of its inverse are bi-orthogonal.

In a nutshell, tensor products are related to matrix properties which are easily described in terms of rank one matrices uv^T. Say $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$. The *tensor product* $V \times W$ is $\mathcal{M}_{nm}$, the vector space of $n \times m$ real matrices. A natural basis is given by the matrices $E_{ij} = e_i e_j^T$, whose only nonzero entry is $e_{ij} = 1$.

More generally, one may define $V \otimes W$ as linear combinations of *symbolic expressions* $v_i \otimes w_j$, for basis elements of both spaces, which are interpreted as a basis for the product. For Hilbert spaces, start from orthonormal bases and expressions like that define the inner product structure of $V \otimes W$. In the infinite dimensional Hilbert case, linear combinations are replaced by (convergent) series.

Exercise 6. As every physicist knows ([45]),

$$L^2(\, \mathbb{R} \times \mathbb{R}, \, dxdy \,) \, \simeq \, L^2(\, \mathbb{R}, \, dx \,) \otimes L^2(\, \mathbb{R}, \, dy \,) \, .$$

For orthonormal bases of functions $f_i \in L^2(\mathbb{R}, dx)$ and $g_j \in L^2(\mathbb{R}, dy)$, the expressions $f_i \otimes g_j$ are identified with $f_i(x) g_j(y)$. The equivalence states that functions in two variables $u(x, y) \in L^2(\mathbb{R} \times \mathbb{R}, dxdy)$ are limits of linear combinations of monomials of the form $f_i(x) g_j(y)$. This is what makes the method of separation of variables work.

Even more generally, one may define $V \otimes W$ in invariant terms, without invoking explicit bases, but this is another story ([32]). Notice that extensions of the symbolic approach, like using expressions

$u_i \otimes v_j \otimes w_k$, are equally amenable, and correspond to vectors which are associated to grid points in *three dimensional* boxes.

Equally tempting, what about different patterns, what if for example the entries of a vector naturally correspond to vertices of an icosahedron? This would lead us to *representation theory*, a fascinating subject outside of the scope of these notes ([49], [50]). For some simple examples, see Section 2.8.2.

One also takes tensor products of linear transformations. Say $A \in \mathcal{M}_{nn}$ and $B \in \mathcal{M}_{mm}$: define $A \otimes B : \mathbb{R}^n \otimes \mathbb{R}^m \to \mathbb{R}^n \otimes \mathbb{R}^m$ to be the map T_p in the theorem above. In particular, $T_s = A \otimes I_n + I \otimes I_m$ where the I_k denotes the identity $I_k : \mathbb{R}^k \to \mathbb{R}^k$.

Exercise 7. For $A \in \mathcal{M}_{nn}$ and $B \in \mathcal{M}_{mm}$, the *Kronecker product* ([27]) (actually, there are two forms of it) yields an $nm \times nm$ matrix associated to $T_p = A \otimes B$ defined in Theorem 7, when one identifies M (in the notation of the theorem) with either the nm-vector obtained by writing sequentially all its rows or all its columns. How does it look? In particular, if the entries of A and B are integer numbers, the same happens to their Kronecker products.

Tensor products are now a hot topic in numerical analysis. Don't expect you favorite linear transformation to be a tensor product, like the discrete Laplacian on a rectangle. But perhaps it is well approximated by a sum of few tensor product monomials. We will have more to say about this in Section 3.4. The interested should consult [3], which approaches from this point of view the numerics of Schrödinger operators, and the review article [29].

2.6 Wedges

Once we consider linear operations like T_s and T_p acting on two sides of matrices M, we may think of special cases on which the matrices M have additional structure. We are interested in $M \in \mathcal{A}(n, \mathbb{R})$, a real skew-symmetric matrix — this is a possible starting point to *exterior algebras* ([53]). We limit our attention to a unique example.

Take $S \in \mathcal{S}(n, \mathbb{R})$ (symmetric) with eigenvalues λ_k and define

$$T_S : \mathcal{A}(n, \mathbb{R}) \to \mathcal{A}(n, \mathbb{R}), \quad A \mapsto S\,A + A\,S\,.$$

Clearly, this is is a well defined linear map and it is so similar to T_s in Theorem 7 that we should be able to compute its eigenvalues.

Theorem 8. $\sigma(T_S) = \{ \lambda_k + \lambda_\ell \, , \, k > \ell \}$

Proof. Take an orthonormal eigenvectors v_i so that $S\, v_i = \lambda_i \, v_i$. Now $v_i \otimes v_j$ is not an eigenvector of T_S, simply because it is not a skew symmetric matrix. Take instead $Z_{ij} = v_i \otimes v_j - v_j \otimes v_i$: to get a basis, we must stick to $i > j$. And the rest is the same:

$$T_S \, Z_{ij} = S \, Z_{ij} + Z_{ij} \, S$$

$$= (S \, v_i) \otimes v_j - (S \, v_j) \otimes v_i + v_i \otimes (S \, v_j) - v_j \otimes (S \, v_i)$$

$$= \lambda_i \, v_i \otimes v_j - \lambda_j \, v_j \otimes v_i + \lambda_j \, v_i \otimes v_j - \lambda_i \, v_j \otimes v_i$$

$$= (\lambda_i + \lambda_j \,) \, (\, v_i \otimes v_j - v_j \otimes v_i \,) = (\, \lambda_i + \lambda_j \,) \, Z_{ij} \, .$$

□

Thus, if $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, the smallest eigenvalues of S and T_S are λ_1 and $\lambda_1 + \lambda_2$. Iterate the process (how?) to prove properties about extremal eigenvalues (an example is in Section 3.2.2).

But then what are the wedges? They are the expressions

$$u \wedge v = u \otimes v - v \otimes u \, .$$

Linear combinations of such monomials span $\mathcal{A}(n, \mathbb{R})$. taking longer strings of wedges leads to the *exterior algebra*, a very elegant formalism to handles areas and volumes, but this is another story ([53]).

2.7 Some applications

2.7.1 Roots of polynomials are eigenvalues

Consider the following natural question: given a polynomial p with roots r_k, find another polynomial q with roots r_k^2. The key point is to notice that the roots are not given, and in principle are not obtainable. But indeed, this is not necessary. Given p, find a matrix having p for its characteristic polynomial — this is accomplished by a companion matrix A. The required polynomial q is the characteristic

polynomial of A^2. Indeed, the eigenvalues of A^2 are simply the square of the eigenvalues of A (by Jordan's theorem, for example, or by proving first for diagonalizable matrices and then taking a limit).

There is nothing sacred about squaring, one might ask for q with roots $f(r_i)$: it is the characteristic polynomial of $f(A)$.

2.7.2 The resultant revisited

From Section 2.4.1, the resultant $R(p, q)$ is a polynomial which is zero exactly when p and q have a common root — by checking its degree, we are forced to have, up to a nonzero constant (which is actually 1),

$$R(p, q) = c\,\Pi^{x_i, y_j}(x_i - y_j),$$

where x_i and y_j are the roots of p and q. From Theorem 7, $R(p, q)$ is the characteristic polynomial of $A \otimes I_m - I_n \otimes B$, where A and B are companion matrices with characteristic polynomials p and q.

2.7.3 Algebraic numbers form a field

An *algebraic number* is a root of a polynomial with integer coefficients: we prove the familiar fact that algebraic numbers are a subfield of $\mathbb{C}$. Thus for example, if x is an algebraic number and $p(x) = 0$, for a polynomial p with integer coefficients, then $1/x$ is a root of a polynomial q obtained by getting rid of denominators in the coefficients of $p(1/x)$. The only nontrivial facts to check are related to closure: the sum and product of two algebraic numbers is another one.

Take x and y roots of p and q with integer coefficients, and consider companion matrices A and B with characteristic polynomials p and q. Both A and B have integer entries, and do not have necessarily the same dimension. Now, as seen in Theorem 7, the transformations $A \otimes I + I \otimes B$ and $A \otimes B$ have respectively $x + y$ and xy among their many roots, and are represented by Kronecker products consisting of integer valued matrices (exercise 7) — we are done.

The reader might be intrigued by the fact that if p and q are of degree n and m, then a polynomial of large degree mn pops up, but this is frequently necessary to accommodate all the roots $x + y$ and xy, — the resulting polynomial is usually irreducible over the rationals.

2.8 Some examples: adjacency matrices

Given a graph G, enumerate its vertices $1, 2, \ldots, n$ and consider the $n \times n$ *adjacency matrix* A whose entry $a_{i,j}$ is 1 or 0, depending if vertices i and j have a common edge or not. Notice that this is simple idea — a matrix as a table — is a natural technique to convey what looks like visual information to a computer.

The reader may imagine a number of generalizations: one could think about directed or undirected graphs (i.e, one or two-way streets) or weighted edges (which give rise to *Markov chains*). Most of what we do admit trivial adaptations to these contexts, but we stick to the simple case of undirected graphs.

The following well known theorem is a nice motivation for adjacency matrices. A path in a graph G with endpoints i and j is of *length* k if it consists of k (possibly repeated) edges.

Theorem 9. *Let G be a graph with adjacency matrix A. Then the number of paths of length k with endpoints i and j is $(A^k)_{ij}$.*

2.8.1 Polygons and second derivatives

We compute the eigenvalues (and eigenvectors) of some special adjacency matrices. Let us start with something simple: G is a bracelet, i.e., a graph with n vertices so that vertex i is adjacent to vertices $i-1$ and $i+1$, where we identify labels 0 and $n+1$. Set $n = 5$:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

We use the first trick in representation theory. Consider the shift

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

Then $A\,S = S\,A$, a fact that can be checked by matrix multiplication, or which might be phrased in words. The shift S essentially moves the indices $1, 2, \ldots, 5$ as $5 \to 4 \to 3 \to 2 \to 1 \to 0 \sim 5$. Now, $S^{-1}AS$ describes adjacencies of the same graph after the vertices have been renamed by the shift, and clearly nothing changes.

Also, it should be clear that we know that $S^5 = I$, the identity matrix and it is not hard to write an explicit diagonalization

$$S = U^*\Lambda U\,, \quad U \in SU(5)\,.$$

Such a diagonalization of S follows from the spectral theorem for normal matrices: S commutes with its transpose $S^T = S^{n-1}$. The matrix Λ contains the fifth roots of unity along the diagonal, say,

$$1,\, \omega = \exp(2\pi/5),\, \omega^2,\, \omega^3,\, \omega^4\,,$$

and the unitary matrix U^* has for columns the associated normalized eigenvectors — it is the *discrete Fourier transform of dimension* 5.

The reader may think that something went wrong: we were interested in A and deviated in order to compute eigenvalues and eigenvectors of the simpler matrix S. We now compute eigenpairs of A in two different ways. The simpler one is to realize that $A = S + S^T = S + S^{n-1}$, so that if v_k is an eigenvector of S associated to S, then v_k is an eigenvector of A associated to $\omega_k + \omega_k^{n-1} = \omega_k + \overline{\omega}_k = 2\cos(2k\pi/n)$. In particular, most of the eigenvalues of A are double.

Exercise 8. From the spectral theorem, A should have only real eigenvalues associated to an orthonormal basis of real eigenvectors — show that this is indeed the case, by taking seriously the fact that most of its eigenvalues are double.

Exercise 9. Consider *circulant matrices*, polynomials in the shift S:

$$p(T) = \sum_{k=1}^{n} c_k\, S^k\,, \quad c_k \in \mathbb{C}\,.$$

Find a pattern in its entries. Compute eigenvalues and eigenvectors.

The second approach is more conceptual. Since A and S commute and S has simple spectrum, A is a function of S, which might

even be taken as a polynomial p, $A = p(S)$. So again we learn that eigenvectors of S are eigenvectors of A: computing eigenvalues once one has eigenvectors is a triviality. One of the first proofs in the representation theory, the so called *Schur's lemma*, resolves the issue as follows. If v is an eigenvector of S, then $S\,v = \lambda\,v$ for some $\lambda \in \mathbb{C}$. Since $A\,S = S\,A$, we must have $\lambda\,A, v = S\,(Av)$, so that Av is either zero (and then v is an eigenvector of A associated to the eigenvalue 0) or $Av \neq 0$ is another eigenvector of S associated to λ. Since λ is a simple eigenvalue, $A\,v = \mu\,v$ and again v is an eigenvector of A.

Exercise 10. Let $f : [0, \pi] \to \mathbb{R}$ be a smooth function with $f(0) = f(\pi) = 0$. The mesh with equally spaced points

$$x_0 = 0 \;<\; x_1 \;<\; \ldots \;<\; x_n \;<\; x_{n+1} = 1\,,$$

yields sub-intervals of size $h = \pi/(n+1)$ and a discretization of the second derivative acting on such functions, which, for $n = 5$, is

$$L_5 = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -2 \end{pmatrix}.$$

To compute its spectrum, it clearly suffices to know the spectrum of

$$A_5 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

The answer is remarkable. The second derivative clearly has eigenfunctions $\sin(kx)$ and eigenvalues k^2 for $k = 1, 2, \ldots$. The eigenvectors $v_k, k = 1, \ldots, n$ of the discrete problem are the evaluations of the continuous eigenfunctions at the points of the grid,

$$v_k = (\,\sin(kx_i)\,) \in \mathbb{R}^n,\; i = 1, \ldots, n$$

and eigenvalues $\lambda_k = 2\cos(k\,x_1)$. This is easy to check — can you *obtain* this result from the periodic case, which is the adjacency matrix of a bracelet ? Show also that, as $n \to \infty$, the k-th discrete eigenpair converges to the k-th continuous eigenpair.

2.8.2 Regular polytopes

It is a remarkable fact, known already in the nineteenth century, that one can enumerate all possible regular polytopes (for a precise definition and an enormous amount of fascinating material, see [9]).

In two dimensions, they are just the regular polygons, with adjacency matrices whose spectra were computed above. In three dimensions, the Greeks knew about the five platonic solids: the tetrahedron, the cube, the octahedron, the dodecahedron and the icosahedron. Things are less familiar in four dimensions, where one still has the counterparts of the tetrahedron (a 4-simplex, or more concretely, the convex span of the five canonical vectors in $\mathbb{R}^5$), the cube, the octahedron (the span of the centers of the faces of the cube) and three other regular polytopes, with 24, 120 and 600 vertices respectively. Rather surprisingly, for dimensions larger than 4, only the three simpler types of polytopes remain.

Simplexes

The n-simplex is the convex span of the canonical vectors in $\mathbb{R}^{n+1}$. In particular its $(n+1)\times(n+1)$ adjacency matrix A satisfies $A = \mathbf{1} - I$, where $\mathbf{1}$ is the matrix all of whose entries are equal to 1. Now, all the canonical vectors are taken by $\mathbf{1}$ to the same vector, which then must be an eigenvector of $\mathbf{1}$,associated to $n+1$. Since $\dim \operatorname{Ran} \mathbf{1} = 1$, the kernel must be of dimension n— the remaining eigenvalues of $\mathbf{1}$ are 0, with multiplicity n. The upshot is that $\sigma(A)$ consists of the eigenvalue -1 with multiplicity n and the simple eigenvalue n.

Cubes

We present two different arguments. The vertices of the n-cube may be taken to be the points in $\mathbb{R}^n$ all of whose coordinates are equal to zero or one. More, there is a natural underlying inductive construction for the adjacency matrix A_n. The list of vertices of the n-cube consists of two copies of the list of vertices of the $(n-1)$-cube to which one appends a 0 or a 1 as last coordinate of each copy. Thus, a labeling of the vertices of the $(n-1)$-cube induces a labeling for the n-cube and the adjacency matrices are related in a simple fashion,

$$A_n = \begin{pmatrix} A_{n-1} & I \\ I & A_{n-1} \end{pmatrix}.$$

A simple computation now shows that if $A_{n-1}v_{n-1} = \lambda_{n-1}v_{n-1}$ then the eigenpairs $(v_N^{\pm}, \lambda_n^{\pm})$ of A_n are given by

$$v_n^+ = (v_{n-1}, v_{n-1}), \quad \lambda_n^+ = \lambda_{n-1} + 1$$

and

$$v_n^- = (v_{n-1}, -v_{n-1}), \quad \lambda_n^- = \lambda_{n-1} - 1.$$

It is easy to see that all eigenvectors constructed in such fashion are orthogonal, hence the juxtaposition of the v_n^+ and v_n^- form a basis.

The inductive step yields the eigenvalues of cubes. In one dimension, they are simply $\{-1, 1\}$. Adding and subtracting 1 to these eigenvalues, we obtain for the 2-cube the eigenvalues $-2, 0$ and 2, where 0 is a *double* eigenvalue. More generally, the n-cube has $n+1$ distinct eigenvalues in arithmetic progression from $-n$ to n, with consecutive numbers differing by two. The multiplicities, in ascending order, are given by the binomial numbers. Thus, the 4-cube, for example, has eigenvalues $-4, -2, 0, 2, 4$, with multiplicities $1, 4, 6, 4, 1$.

We consider a second argument, which is more complicated but more flexible. The formula $Av = \lambda v$ for the eigenpair of an adjacency matrix A has a geometric interpretation. Think of v as the obvious distribution of numbers at the vertices of A — the coordinates of v are labeled by the indices of the vertices. Now Av is a similar distribution obtained by adding the neighboring values of v at a vertex. Thus, for example, for the 3-cube, the vector v which corresponds to the number 1 at each of the eight vertices, gives for Av the vector for which there is a 3 on each vertex — we have just found an eigenvector associated to the eigenvalue 3.

We start with $n = 3$. Hold the cube from a vertex (or bend your head) and think of $(1, 1, 1)$ as a vertical vector. Vertices lie in three different planes of $\mathbb{R}^3$ according to the number of nonzero coordinates:

$$\{1, 1, 1\}, \quad \{(1, 1, 0), (1, 0, 1), (0, 1, 1)\},$$

$$\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}, \quad \{(0, 0, 0)\}.$$

Now, a rotation R by $2\pi/3$ around the vertical axis keeps the two extreme vertices fixed and permutes the vertices of the intermediate levels. More, if $A_3 v = \lambda v$ for $v \neq 0$, then Rv and $R^2 v$ are also

eigenvectors of A_3 associated to the same λ. Without loss (why?) we may suppose that v at the vertex $(1,1,1)$ is not equal to zero. Thus,

$$v_m \,=\, \frac{1}{3}\big(\, v \,+\, R\, v \,+\, R^2\, v\,\big)$$

is also a (nonzero) eigenvector of A_3 associated to λ with the additional property that it is constant on each level — v_m is the average of an orbit of a $\mathbb{Z}_3$ action which keeps the levels invariant and is *transitive* on each level (i.e., there are group elements which take one point of a level to any other in the same level).

In a nutshell, every eigenvalue of A_3 is associated to an eigenvector which is constant on each of the four levels. Let V be the subspace of vectors v which are constant on each level, clearly a vector space of dimension 4. The fact that A commutes with the rotation R (why? think visually about the effect that A and R have on distributions of numbers at vertices) implies that V is an invariant subspace of A. Thus, to obtain the eigenvalues of A_3, it suffices to look at the eigenvalues of the restriction of A_3 to V. Label the levels A, B, C and D. A vertex in A has three neighbors in B, a vertex in B has one neighbor in A and two in C, one in C has two in B and one in D and finally the vertex in D has three neighbors in C. A distribution of values (a,b,c,d) giving rise to a vector in V would be taken by A_3 to the vector in V associated to $(3b, a+2c, 2b+d, 3c)$, so that a matrix representation of the restriction of A_3 to V is given by

$$\begin{pmatrix} 0 & 3 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 3 & 0 \end{pmatrix}.$$

We can do better. This matrix has a special symmetry: the entry (i,j) equals the entry $(n+1-i, n+1-j)$ (surely your eyes have a simple description of this symmetry). This in turn implies that V splits in two additional invariant subspaces $V = V^e \oplus V^o$ of *even* and *odd* vectors of the form (a,b,b,a) and $(a,b,-b,-a)$ (this could be inferred abstractly and pedantically, by an averaging argument). Also, on V^e and V^o, A_3 acts as follows:

$$(a,b,b,a) \,\mapsto\, (3b, a+2b, b+2a, 3b)\,,$$

$$(a, b, -b, -a) \mapsto (3b, a - 2b, b - 2a, -3b)\,,$$

and representations of A_3 restricted to such smaller subspaces are

$$\begin{pmatrix} 0 & 3 \\ 1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 0 & 3 \\ 1 & -2 \end{pmatrix}.$$

So, each eigenvalue of the 8×8 matrix A_3, is an eigenvalue of one of these 2×2 matrices, which are respectively $\{-1, 3\}$ and $\{-3, 1\}$.

For the general case A_n, there are $n + 1$ levels, defined as above, and the issue is the existence of a group of special orthogonal matrices (like the rotation R) which acts transitively on the levels. Indeed, there is such a group: it is the group of even permutations on n symbols (for $n = 3$, it is $\mathbb{Z}_3$), represented as $n \times n$ permutation matrices. The action is just the permutation of the entries of a vector: it clearly preserves levels and is indeed transitive, so the subspace of vectors which are constant on levels plays the same role as before. Notice the implicit use of the fact that, on each level (orbit action), the number of group elements taking one coordinate to another is the same (the isotopy group at each orbit element is of the same size).

As a nice corollary, we obtain a collection of tridiagonal matrices with simple spectrum having integer eigenvalues in arithmetic progression — extend the 3×3 example above. These matrices come up in Lie algebra theory, as for example in the theory of Verma modules.

By the way, once the eigenvalues are computed, how does one go about computing multiplicities? A simple answer goes as follows. From Theorem 9, the number of closed paths of length k of such a regular graph with adjacency matrix A is given by $\operatorname{tr} A^k$ (why?), an easily computable number. But $\operatorname{tr} A^k = \sum_i \lambda_i^k$, so for each k one obtains a linear relation among the multiplicities of the many eigenvalues.

The polytope with 600 vertices

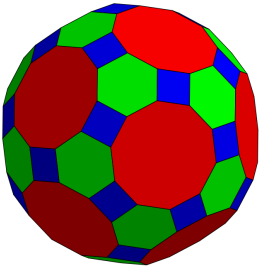
The spectra of all regular polytopes was obtained in [47]. The spectrum of the adjacency matrix of the four-dimensional polytope with 600 vertices, for example, has been computed with the many levels technique described above for the cube. There are 46 levels which, using an additional $\mathbb{Z}_2$-involution, give rise to two invariant

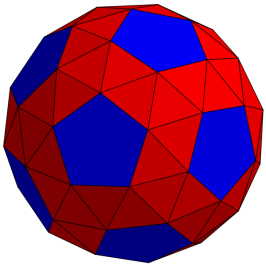
subspaces of dimension 23 containing the original eigenvalues. Rather surprisingly, the eigenvalues of all adjacency matrices of all regular polytopes can be explicitly calculated, in the sense that all the characteristic polynomials have no Galois-type obstructions.

Let us add some details. One might use a numerical algorithm to approximate eigenvalues to compute the 46 eigenvalues above. It takes some... inspiration (but there are algorithms for this also, nowadays) to conclude that some of these numbers are *surds*, a very old word to describe numbers of the form $a + b\sqrt{c}$, $a, b \in \mathbb{Q}$, $c \in \mathbb{N}$. How does one prove precisely that a surd is indeed an eigenvalue, using only integer arithmetic? Well, we all know that a real matrix has eigenvalues coming in complex conjugate pairs. If a matrix M with integer entries has an eigenvalue $a + b\sqrt{c}$ it will also have $a - b\sqrt{c}$ for eigenvalue, so that the matrix $(M - aI)^2 - b^2\, c\, I$ should have two eigenvalues equal to zero, a check which is easily performed in $\mathbb{Z}$.

2.8.3 Semi-regular polytopes

The usual soccer ball consisting of hexagons and pentagons is an example of a semi-regular polyhedron. There are thirteen of them in three dimensions, the so called Kepler polyhedra. The list of semi-regular polytopes in all dimensions has been known since the nineteenth century, but the confirmation that it was indeed complete is a result from the early nineties, by Blind and Blind ([4]).



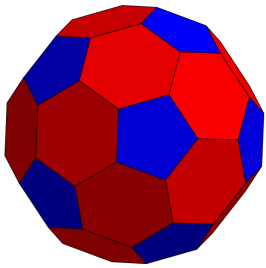


These spectra were computed in [48]. The subfamily of *Gosset polytopes* only has integer eigenvalues, hundreds of them. In the pic-

ture, we show the only two semi-regular polytopes whose eigenvalues are beyond solutions by radicals. Both are three dimensional.

Additional techniques from representation theory were required. We all know that if a collection of diagonalizable matrices commute, then they are simultaneously diagonalizable, i.e., they have a diagonal representation on a common basis. Representation theory studies with spectacular success the following generalization: how far can one go towards diagonalizing a finite group of matrices? A small part of the answer is that full diagonalization is usually not possible, but one may achieve a common block-diagonal form, where the sizes of the block may be known in advance.

A related technique, also used in [48], handles the so called *Cayley graphs*, which are graphical descriptions of multiplication tables of groups described by generators and relations. The spectrum of adjacency matrices associated to graphs — *spectral graph theory* — is an intense field of research, associated to issues in contagion of diseases, and more generally, the propagation of information ([10],[7]).



As a final remark, we slightly expand our collection of graphs. There is a special configuration of atoms of carbon, *fullerene*, consisting of 60 atoms arranged as vertices of the soccer ball formed by hexagons and pentagons. Carbon usually attaches to four neighbors, and each vertex has only three, but the edges between hexagons have two links between them, so one performs a slight modification on the adjacency matrix to take into account this information. The spec-

trum of this modified adjacency matrix is of practical interest, and may be computed with the techniques described before ([8], [48]).

Chapter 3

Some analysis

3.1 Algebras of matrices and operators

Square matrices (or for the matter linear transformations, or even bounded operators from a Banach space to itself) are not just a normed vector space — they form an *algebra*, i.e., they can be multiplied. In particular, we can evaluate squares, cubes, polynomials of square matrices and transformations of a space to itself. More general functions are also relevant.

Let X be a complex Banach space and consider

$$\mathcal{B} = \mathcal{B}(X) = \{ \text{ linear bounded transformations } T : \mathcal{B} \to \mathcal{B} \} ,$$

which we endow with the *operator norm*

$$|| T || = \sup_{||v||=1} || Tv || .$$

With this norm, $\mathcal{B}$ is known to be a Banach space, with a special feature: this norm is *multiplicative* ,

$$|| TS || \leq || T || \, || S || , \quad \forall T , S \in \mathcal{B} .$$

Norms are not supposed to handle multiplications, being defined on vector spaces. This property is an extra feature — $\mathcal{B}$ now is a *Banach algebra*, but we shall not deal in such generality. The great advantage

of working with a multiplicative norm is that *the estimates performed on* $\mathbb{C}$ *transfer to* $\mathcal{B}$. This vague statement deserves an illustration.

What is e^π ? The answer through the series

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} \ldots ...$$

is so good that actually much more is obtained. First, of course, we have to give meaning to the limit indicated by the dots. This follows by verification that the sequence of partial sums is a Cauchy sequence. It requires the triangular inequality, the fact that $|cz^k| \leq |c|\,|z|^k$ (well, actually an equality) and estimates in terms of geometric progressions, which is how we usually estimate series in an open disk.

There is nothing wrong in replacing z in the series by an $n \times n$ matrix or by an operator $T \in \mathcal{B}$ — all the steps used to give sense to the limit still make sense! This is where the multiplicative property of the norm is handy: indeed, $||T^k|| \leq ||T||^k$. In the series, 1 has to be replaced by the neutral element of multiplication, namely I. The argument also shows that e^T is indeed a *bounded* operator.

Exercise 11. The imitation goes further. If one has to justify the convenience of being able to compute e^T, one may prove that the (unique) solution to the differential equation

$$v'(t) = Tv(t),\, v(0) = v_0$$

is given, as usual, by $v(t) = e^{tT}\, v_0$ — simply show that derivatives may be taken term by term as in the scalar case.

Exercise 12. In the notation of Theorem 7, consider

$$X'(t) = T_s X(t) = A\, X(t) - X(t)\, B\,, \quad X(0) = X_0\,.$$

Exponentiate (with the Taylor series) T_s to obtain

$$X(t) = e^{t\, T_s}\, X_0 = e^{t\, A}\, X_0\, e^{t\, B}\,.$$

Now take $X' = T_p X = A\, X\, B\,, \quad X(0) = X_0$. What now?

Let us identify another such operator. Consider the exponential

$$F : \mathcal{M}(n, \mathbb{R}) \to \mathcal{M}(n, \mathbb{R})\,, \quad M \mapsto e^M\,.$$

(everything works for complex numbers too). What is the derivative of F at M along the direction A ? We have to compute the standard Newton-type limit. All possible rearrangements are permitted since the series are absolutely convergent (just like real series...):

$$\lim_{t\to 0}\frac{F(M+tA)-F(M)}{t}=\lim\frac{e^{M+tA}\,-\,e^{M}}{t}$$

$$=\lim\frac{1}{t}\,(\,I\,-\,I\,+\,(M+tA\,-\,M)$$

$$+\frac{1}{2!}\big((M+tA)^2\,-\,M^2\big)\,+\,\frac{1}{3!}\big((M+tA)^3\,-\,\frac{1}{3!}M^3\big)+\ldots)$$

$$=A+\frac{1}{2!}\left(M\,A+A\,M\right)+\frac{1}{3!}\left(M^2\,A+M\,A\,M+A\,M^2\right)+\ldots=T\,A$$

which in principle is the answer to the problem — notice by the way that T is indeed a *linear* transformation T on A.

There is something to learn by computing eigenvalues of this transformation. Say M has distinct eigenvalues μ_k and eigenvectors

$$M\,v_k=\mu_k\,v_k\,,\quad w_\ell^T M=\mu_\ell\,w_\ell^T\,,\quad ,k,\ell=1,\ldots,n\,.$$

As in the proof of Theorem 7, the matrices $Z_{k\ell}=v_k\otimes w_\ell=v_k w_\ell^T$ are eigenvectors of T, with eigenvalues

$$1\,+\,\frac{1}{2!}\left(\mu_k+\mu_\ell\right)\,+\,\frac{1}{3!}\left(\mu_k^2+\mu_k\,\mu_\ell+\mu_\ell^2\right)\,+\,\ldots$$

$$=\frac{1}{\mu_k-\mu_\ell}\big(\mu_k-\mu_\ell+\frac{1}{2!}\,(\mu_k^2-\mu_\ell^2)+\frac{1}{3!}\,(\mu_k^3-\mu_\ell^3)\big)$$

$$=\frac{e_k^\mu-e_\ell^\mu}{\mu_k-\mu_\ell}\,,$$

a beautiful formula. In particular, if M has two eigenvalues differing by $2\pi i\in\mathbb{C}$, the Jacobian $DF(M)$ is *not* a bijection.

3.2 Smoothness of eigenpairs

We present a short argument which implies smoothness of eigenvalues and eigenvectors, combined with the formulas for the derivatives.

Let $\mathbb{K}$ be either $\mathbb{R}$ or $\mathbb{C}$: $\mathbb{C}$ is the usual context for spectral theory, but we are also interested in real eigenvalues. Consider $X \subset Y$ Banach spaces over $\mathbb{K}$ and let $\mathcal{B} = \mathcal{B}(X, Y)$ be the space of linear transformations from X to Y, bounded in the sup norm. Let $\lambda_0 \in \mathbb{K}$ and $\phi_0 \in X$ be eigenvalue and eigenvector of $T_0 \in \mathcal{B}$, so that $(T_0 - \lambda_0 I)\phi_0 = 0$.

In order for ϕ_0 to be well defined, we need some kind of normalization. Let $\ell \in X^*$ be a linear functional for which $\ell(\phi_0) = 1$. This choice is better than normalizing by the norm, since the unit ball is not necessarily a manifold for a general Banach space (think of balls with spikes, for example).

Another difficulty for a theorem stating smooth variation of eigenpairs are double eigenvalues. This is easy to see already on two dimensional examples. We need additional hypotheses which somehow exclude this possibility.

Theorem 10. *Suppose that $T_0 - \lambda_0 I$ is a Fredholm operator of index zero with one dimensional kernel and that $\phi_0 \notin \operatorname{Ran}(T_0 - \lambda_0 I)$. Set $Z = \phi_0 + \operatorname{Ker} \ell$. Then there is an open neighborhood $U \subset \mathcal{B}$ of T_0 and unique maps $\lambda : U \to \mathbb{K}$ and $\phi : U \to Z$ with $(T - \lambda(T) I)\, \phi(T) = 0$. Such maps are analytic.*

For finite dimensional spaces, the hypothesis $\phi_0 \notin \operatorname{Ran}(T_0 - \lambda_0 I)$ is equivalent to the statement that λ_0 has algebraic multiplicity one.

Proof. Clearly Z is a closed, affine subspace of X of codimension 1. We use the implicit function theorem on the analytic map

$$H : \mathcal{B} \times Z \times \mathbb{K} \to Y\,, \qquad T,\, \phi,\, \lambda \mapsto (T - \lambda I)\, \phi.$$

We must show the invertibility at (T_0, ϕ_0, λ_0) of the derivative of H with respect to the last two variables,

$$DH_{\phi,\lambda}(T_0, \phi_0, \lambda_0)(v, c) = (T_0 - \lambda_0 I)v - c\phi_0\,.$$

Let $\langle \phi_0 \rangle$ be the vector space generated by ϕ_0. By hypothesis,

$$X = \operatorname{Ker} \ell \oplus \langle \phi_0 \rangle\,, \quad Y = \operatorname{Ran}(T_0 - \lambda_0 I) \oplus (DF(u) - \lambda(u))$$

and $T_0 - \lambda_0 I$ respects the decompositions. To solve

$$(T_0 - \lambda_0 I)\, v - c\, \phi_0 = y\,,$$

split $v = k + c\,\phi_0$ and $y = r + d\,\phi_0$: $k = (T_0 - \lambda_0 I)^{-1}\, r$ and $c = -d$. □

3.2.1 Bi-orthogonality, derivatives of eigenpairs

In Section 2.5, we used two sided eigenvectors extensively, they are convenient when matrices are not symmetric. On $\mathbb{C}^n$, take the usual complex inner product which is skew-linear in the first coordinate.

Proposition 5. *Let $M \in \mathcal{M}(n, \mathbb{C})$ have distinct eigenvalues $\{\lambda_k\}$, and two sided eigenvectors*

$$M\, v_k = \lambda_k\, v_k\,, \quad w_\ell^* M = \lambda_\ell\, w_\ell^*\,, \quad k, \ell = 1, \ldots, n\,.$$

Then $\langle\, v_k\,, w_\ell\,\rangle = 0$ for $k \neq \ell$. Also, $\langle\, v_k\,, w_k\,\rangle \neq 0$.

The orthonormal basis of eigenvectors of self-adjoint matrices is a manifestation of this *bi-orthogonality* of both eigenvector bases.

Proof. The proof follows the standard self-adjoint counterpart:

$$\overline{\lambda_k}\langle\, v_k\,, w_\ell\,\rangle = \langle \lambda_k\, v_k\,, w_\ell\,\rangle = \langle M\, v_k\,, w_\ell\,\rangle$$

$$= \langle\, v_k\,, M^*\, w_\ell\,\rangle = \langle\, v_k\,, \overline{\lambda_\ell}\, w_\ell\,\rangle = \overline{\lambda_\ell}\,\langle\, v_k\,, w_\ell\,\rangle\,,$$

and now use that $\lambda_k \neq \lambda_\ell$. If $\langle\, v_k\,, w_k\,\rangle = 0$, then v_k would be orthogonal to all the elements of the basis of the w_ℓ's, a contradiction. □

Exercise 13. Prove something harder — use the Jordan form. Say λ is a simple eigenvalue of M with left and right eigenvectors v and w. Then $\langle\, v\,, w\,\rangle \neq 0$. State and prove something similar in infinite dimensions: consider the Dunford-Schwartz calculus of Section 5.1.

We compute derivatives of eigenvalues and eigenvectors of matrices, but they extend easily to infinite dimensional operators: one only has to add the hypotheses of Theorem 10.

In the finite dimensional case, if λ_0 is a simple eigenvalue of M_0 with eigenvectors v_0 and w_0,

$$M_0\, v_0 = \lambda_0\, v_0\,, \quad w_0^* M = \lambda_0\, w_0^*\,,$$

we saw in Theorem 10 that that there is a neighborhood of M_0 for which, under appropriate normalizations, there are (analytic) functions $\lambda(M)$, $v(M)$ and $w(M)$ which equal λ_0, v_0 and w_0 at $M = M_0$. We simply take derivatives on a parameter of the eigenvalue equation

$$M\, v = \lambda\, v\,,$$

yielding, in very convenient notation,

$$\dot{M}\, v \,+\, M\, \dot{v} = \dot{\lambda}\, v \,+\, \lambda \dot{v}\,.$$

Now take inner products with w to get

$$\langle\, w\,, \dot{M}\, v\,\rangle \,+ \langle\, w\,, M\, \dot{v}\,\rangle = \langle\, w\,, \dot{\lambda}\, v\,\rangle \,+ \langle\, w\,, \lambda \dot{v}\,\rangle\,,$$

so that, imitating the proof of Proposition 5,

$$\langle\, w\,, \dot{M}\, v\,\rangle \,+ \langle\, M^*\, w\,,\, \dot{v}\,\rangle = \dot{\lambda}\, \langle\, w\,,\, v\,\rangle \,+\, \lambda\, \langle\, w\,, \dot{v}\,\rangle$$

and we obtain

$$\dot{\lambda} = \frac{\langle\, w\,, \dot{M}\, v\,\rangle}{\langle\, w\,,\, v\,\rangle}\,,$$

which is well defined from the previous exercise. Notice also that the expression is independent of the normalizations we impose on both eigenvectors. To clarify matters, denote by ∂_A the directional derivative along direction A. This equation states that

$$\partial_A \lambda(M) = \frac{\langle\, w\,, \partial_A M\, v\,\rangle}{\langle\, w\,,\, v\,\rangle}\,.$$

We now compute the derivative $\dot{v}$ of the eigenvector v. From the derivative of the eigenvalue equation, we have

$$(M - \lambda\, I)\, \dot{v} = -\, c\,,$$

so that the apparent obstruction —the fact that $(M - \lambda\, I)$ is not invertible — has to be irrelevant. The right hand side $(\dot{M} - \dot{\lambda}\, I)\, v$ is indeed in the range of $(M - \lambda\, I)$, which is of rank $n - 1$. Indeed,

$$\mathrm{Ran}(M - \lambda\, I) = \big(\,\mathrm{Ker}(M^* - \overline{\lambda}\, I)\big)^{\perp} = w^{\perp}\,,$$

so we have to check that $\langle\, w\,, \dot{M} - \dot{\lambda}\, I)\, v\,\rangle = 0$, which is exactly what we get from the formula for $\dot{\lambda}$. Thus we know $\dot{v}$ up to a kernel vector $c\, v$: this is where we need to specify a normalization for v.

We consider only a simple case: say M is a real, symmetric matrix, and we take $||v|| = 1$. Then it is easy to see that the restriction $(M - \lambda_I)|_{v^\perp} : v^\perp \to v^\perp$ is a bijection, and the normalization forces $c = 0$: in a nutshell,

$$\dot{V} = -\,(M - \lambda_I)|_{v^\perp}^{-1}\,(\dot{M} - \dot{\lambda}\, I)\, v\,.$$

The simple details are left to the reader.

3.2.2 Continuity of eigenvalues

Self-adjoint matrices have real eigenvalues and can be ordered, say

$$\lambda_1 \,\leq\, \lambda_2 \,\leq\, \ldots \,\leq\, \lambda_n\,.$$

We are thus entitled to ask about continuity of $\lambda_k(S)$ as a function of $S \in \mathcal{S}$ — this is a well known fact. We present a brief argument. The standard min-max definition of λ_1 follows by simplifying the computations in the previous section,

$$\lambda_1 = \min_{||v||=1} \langle\, S\, v\,, v\,\rangle\,,$$

and continuity is immediate. For the remaining eigenvalues, one may consider the formulas for λ_k associated to min-max, or simply invoke the fact that $T_S : \mathcal{A}(n, \mathbb{R}) \to \mathcal{A}(n, \mathbb{R})$ in Section 2.6 has minimal eigenvalue $\lambda_1 + \lambda_2$, from which continuity of $\lambda_1 + \lambda_2$ and hence λ_2 follow. The argument extends for all eigenvalues.

3.3 Some variational properties

A real, $n \times n$ symmetric matrix S gives rise to a quadratic form

$$Q : S^{n-1} \subset \mathbb{R}^n \to \mathbb{R}\,, \quad v \,\mapsto\, \langle\, v\,, S\, v\,\rangle\,,$$

whose critical points (i.e., the vectors $v \in S^{n-1}$ for which $DQ(v) = 0$) are the eigenvectors of S and critical values are its eigenvalues. We can do almost the same for general matrices.

Proposition 6. *Consider the function*

$$F : S^{n-1} \times S^{n-1} \to \mathbb{R}\,, \quad u\,, v \mapsto \langle\, u\,, M\, v\,\rangle\,.$$

Then a point (u, v) is critical if and only if u and v are eigenvectors respectively of $M\,M^T$ and $M^T\,M$ associated to the same eigenvalue.

Recall from Exercise 4 that $\sigma(\,M\,M^T\,) = \sigma(\,M^T\,M\,)$.

Proof. Take directional derivatives for $\langle\, a\,, u\,\rangle = 0$ and $\langle\, b\,, v\,\rangle = 0$,

$$DF(u,v)\,(a,b) = \langle\, a\,, M\, v\,\rangle + \langle\, u\,, M\, b\,\rangle\,,$$

which is always zero only if

$$\forall\, a \in u^{\perp}\,, \langle\, a\,, M\, v\,\rangle = 0 \quad \text{and } \forall\, b \in v^{\perp}\,, \langle\, u\,, M\, b\,\rangle = 0\,.$$

Thus, there are constants α and β such that

$$M\, v = \alpha\, u \qquad M^T\, u = \beta\, v$$

and

$$M^T\, M\, v = \alpha\beta\, v \quad \text{ and } \quad M\, M^T\, u = \alpha\beta\, u\,.$$

□

In the symmetric case, the variational description of eigenvalues and eigenvectors is related to the spectral theorem: $S = Q\Lambda Q^T$, or better, $S\,Q^T = Q\,\Lambda$, so that the columns of Q are the eigenvectors of S. To show that Q may be taken orthogonal start with a generic matrix with simple spectrum, conclude that Q may be taken orthogonal (from bi-orthogonality, if you want). For an arbitrary symmetric S, take limits, keeping in mind that the orthogonal group is compact.

Similarly, the variational principle above is related to the *singular value decomposition* (SVD) of a matrix M: $M = Q\Sigma U$, where now Q and U are orthogonal matrices and Σ is a non-negative diagonal.

Theorem 11. *Every $M \in \mathcal{M}(n, \mathbb{R})$ is a product*

$$M = Q^T\,\Sigma\, U\,, \quad Q,\, U \,\in\, O(n), \quad \Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)\,,\, \sigma_k > 0\,.$$

Proof. Use the spectral theorem and Exercise 4 to write

$$M\,M^T = Q\,\Sigma^2\,Q\,, \quad M^T\,M = U^T\,\Sigma^2\,U\,.$$

Here we make the generic hypothesis

$$\sigma(M\,M^T) = \sigma(M^T\,M) = \{\sigma_1^2 > \sigma_2^2 > \ldots > \sigma_n^2, \sigma_k \in \mathbb{R}\}\,,$$

and take $\Sigma = \ \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$. Indeed, if we want $M = Q^T\,\Sigma\,U$, we must have spectral decompositions of $M\,M^T$ and $M^T\,M$ as above: we must show that the Q, U and Σ obtained by the spectral decompositions indeed yield M. Using that M is invertible ($\sigma_k > 0$), write

$$M\,M^t = M\,(\,M^T\,M\,)\,M^{-1}$$

so that comparing spectral decompositions (and using the generic hypothesis) we learn that

$$Q\,D = M\,U^T$$

for some real diagonal matrix D (each eigenvector of $M\,M^T$ is well defined up to normalization), and thus $M = Q\,D\,U$. Now

$$M\,M^T = \ Q\Sigma^2 Q^T = Q\,D^2\,Q^T$$

so $D^2 = \Sigma^2$, since both matrices are positive. But from the hypothesis, $\Sigma > 0$ and we want $D > 0$: we must have $D = \Sigma$. Get rid of the generic hypothesis by taking limits. □

The SVD may be interpreted as an extension of the spectral theorem. It is probably the most important neglected subject in linear algebra courses. It is also a the fundamental tools in applied mathematics (a good example is [20]). We will see why, from a theoretical standpoint, in the next section. The SVD of rectangular (and complex) matrices is equally important: a very clear presentation is [61].

The *singular values* σ_k of M have a very simple geometric interpretation, which is indicative of their relevance. The (Euclidean) unit ball is taken to an ellipsoid by M — its semi-axes are the σ_k's. Clearly, when M is symmetric, then $\sigma_k = |\lambda_k|$.

3.4 Approximations of small rank

We start rewriting the spectral theorem and the SVD.

Theorem 12. *Let S be a real, $n \times n$ symmetric matrix, with eigenvalues λ_k and normalized eigenvectors v_k, so that it has a spectral decomposition $S = V \Lambda V^T$. Then*

$$S = \lambda_1 \, v_1 \otimes v_1 + \lambda_2 \, v_2 \otimes v_2 + \ \ldots \ + \lambda_n \, v_n \otimes v_n \, ,$$

where the v_k's are the columns of V. For M a real $n \times n$ matrix with SVD decomposition $M = Q \, \Sigma \, U^T$,

$$M = \sigma_1 \, q_1 \otimes u_1 + \sigma_2 \, q_2 \otimes u_2 + \ \ldots \ + \sigma_n \, q_n \otimes u_n + \ ,$$

where the q_k's and u_k's are the columns of Q and U.

Said differently, S and M are decomposed in a sum of rank one maps. In both cases, the monomials are orthogonal to each other with respect to the usual matrix inner product, $\langle X \, , Y \rangle = \operatorname{tr} X^T \, Y$.

Proof. In the symmetric case, we show that the sum of monomials acting on each v_k gives $\lambda_k \, v_k$, which is obvious, since $u_\ell \otimes u_\ell = u_\ell \, u_\ell^T$. In the general case, the action of the sum on u_k has to be equal to $M \, u_k = Q \, \Sigma \, U^T u_k$, and again orthogonality (of U) does it. □

The matrix inner product defines a distance between matrices. Let $\mathcal{M}_k$ be the set of real $n \times n$ matrices with rank at most k. Given a matrix M, which is the matrix $M_k \in \mathcal{M}_k$ closest to it?

Theorem 13. *Given the SVD $M = Q \Sigma U^T$, with ordered singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$,*

$$M_k = \sum_{i=1}^{k} \sigma_i q_i \otimes u_i \, .$$

Similarly, for S a real, symmetric matrix with spectral decomposition $S = Q \Lambda Q^T$ and $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$,

$$S_k = \sum_{i=1}^{k} \lambda_i v_i \otimes v_i \, .$$

Hopefully, M_k and M differ by a small amount even for small k: the distance between both matrices is given by Pythagoras:

$$|| M - M_k ||^2 = \sum_{i=k+1}^{n} \sigma_i^2 .$$

This is a fascinating opening to random matrix theory ([11], [55]): given a measure in the space of matrices, how do the singular values (or the eigenvalues, in the symmetric case) distribute? The answers are surprisingly benign: substantial truncation is frequently possible. We do not treat these issues in this text.

Proof. We handle the nonsymmetric case, the other is similar. For any matrices X and Y and orthogonal matrices Q and U, we have

$$\langle Q\, X\, U^t\, , Q\, Y\, U^T \rangle = \langle X\, , Y \rangle ,$$

so that multiplication on the left and on the right by orthogonal matrices are isometries in $\mathcal{M}(n, \mathbb{R})$. Thus, without loss, instead of the general $M = Q\, \Sigma\, U$ we only consider $M = \Sigma$.

We make the generic hypothesis that the singular values of Σ (which are also its eigenvalues) are distinct and greater than 0 — the general Σ is handled by taking limits, using the continuity of eigenvalues proved in Section 3.2.2 and the compactness of $SO(n, \mathbb{R})$.

Since $\mathcal{M}_k \subset \mathcal{M}(n, \mathbb{R})$ is a closed set, a closest matrix $\hat{M}_k \in \mathcal{M}_k$ to Σ exists by compactness. Notice that we don't know that $\hat{M}_k$ is diagonal, or for the matter, symmetric. Indeed, once once this is proved, the rest is trivial — we have to find a diagonal matrix with at most k nonzero diagonal entries closest to Σ: the answer is exactly what the statement of the theorem says, and is trivially proved.

We show that $\hat{M}_k$ is diagonal. Any invertible matrix, like Σ, is best approximated in $\mathcal{M}_k$ by a matrix of rank *equal* to k (why?). Take an SVD $\hat{M}_k = \hat{Q}\, \hat{\Sigma}_k\, \hat{U}^T$, where the first k diagonal entries of $\hat{\Sigma}_k$ are nonzero. The matrices in $\mathcal{M}_k$ near $\hat{M}_k$ are parameterized by

$$N(A, B, E) = e^A\, \hat{Q}\, (\, \hat{\Sigma}_k + E_k\,)\, \hat{U}^T\, e^B ,$$

for skew-orthogonal matrices A, B near 0 and a diagonal matrix E_k with nonzero entries only along the first k diagonal entries. (For more about this parametrization, see the exercise below).

We are now reduced to a calculus problem: derivatives along directions A, B and E_k of the function $||\,\Sigma - N(A,B,E_k)\,||^2$ at the point $N(0,0,0) = \hat{M}_k$ should be equal to zero:

$$||\,\Sigma - N(A,B,E_k)\,||^2 = \mathrm{tr}\,(\Sigma - N)^T\,(\Sigma - N)$$

$$= \mathrm{tr}\;\Sigma^2 + \mathrm{tr}\;N(A,B,E_k)^T\,N(A,B,E_k) - 2\;\mathrm{tr}\,\Sigma\,N(A,B,E_k)$$

$$= \mathrm{tr}\;\Sigma^2 + \mathrm{tr}\,(\hat{\Sigma}_k + E_k)^2 - 2\;\mathrm{tr}\,\Sigma\,N(A,B,E_k)\,.$$

Take the derivative with respect to A at $(0,0,0)$ — or more precisely, replace A by tA and take the derivative with respect to t, to get

$$\mathrm{tr}\;\Sigma\,A\,\hat{Q}\;\hat{M}_k\,\hat{U}^T = 0 \quad \text{or} \quad \mathrm{tr}\;A\,\hat{Q}\;\hat{M}_k\,\hat{U}^T\,\Sigma = 0\,, \quad \forall\, A \in \mathcal{A}\,,$$

and we learn that $\hat{Q}\;\hat{M}_k\,\hat{U}^T\,\Sigma = \hat{M}_k\,\Sigma$ is a symmetric matrix. In the same fashion, taking the derivative with respect to B, we learn that $\Sigma\,\hat{M}_k$ is also symmetric. Write down in coordinates the equalities

$$\hat{M}_k\,\Sigma = \Sigma\hat{M}_k^T \quad \text{and} \quad \Sigma\,\hat{M}_k = \hat{M}_k^T\,\Sigma$$

(recall that Σ has simple spectrum) to see that $\hat{M}_k$ is diagonal. □

Exercise 14. The parametrization $N(A,B,E_k)$ is not injective, and this is not relevant: we only want to take directional derivatives in variables for which they make sense. All we need is the smoothness of the matrices Q, Σ and U, which follow from the smoothness of simple eigenvalues and eigenvectors proved in Section 10.

3.5 Isospectral manifolds

Sets of matrices with a given fixed spectrum are frequently manifolds. We begin with the first nontrivial manifold of matrices, $SO(n,\mathbb{R})$.

Indeed, $SO(n,\mathbb{R}) = SO \subset \mathcal{M}(n,\mathbb{R})$ is a (compact) submanifold of dimension equal to the dimension of the vector space $\mathcal{A}$, which happens to be its tangent space at the origin. The argument is surely familiar: the matrix $I \in \mathcal{S}$ is a regular value of the map

$$F : \mathcal{M} \to \mathcal{S}\,, \quad M \mapsto M^T M\,,$$

so that $SO = F^{-1}(I)$ is a manifold. The *tangent space* at I, $T_I SO$, is the vector space of values $Q'(0)$ for curves $Q(t) \in SO$, $Q(0) = I$. Taking derivatives of $Q^T(t)\ Q(t) = I$ yields $Q'(0) \in \mathcal{A}$. Also, the curve $Q(t) = e^{tA} \in SO$, for $A \in \mathcal{A}$, satisfies $Q'(0) = A$ so $T_I SO = \mathcal{A}$.

We now consider an *isospectral* manifold of matrices. Let Λ be a real $n \times n$ diagonal matrix with simple, ordered eigenvalues along the diagonal, $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. We consider the set of real, symmetric matrices with eigenvalues equal to those of Λ,

$$\mathcal{S}_\Lambda = \{Q^T \Lambda\ Q\ ,\ Q \in SO(n, \mathbb{R})\}.$$

The set $\mathcal{S}_\Lambda$ would be a compact manifold even if some eigenvalues were equal, by general arguments about group actions ([1]). Here, we start from scratch, require simplicity of spectrum, and provide more information. Let $\mathcal{S}$ and $\mathcal{A}$ denote respectively the vector space of real, symmetric and skew-symmetric matrices of dimension n. We also denote by $\mathcal{C}(S)$ the vector space of all polynomials of S, i.e., of matrices of the form $p(S)$, for arbitrary polynomials.

Exercise 15. $\mathcal{C}(S)$ consists of the matrices $p(S)$, where p is a polynomial of degree at most $n-1$. Hint: without loss, take $S = \Lambda$.

Theorem 14. *$\mathcal{S}_\Lambda$ is a compact, oriented manifold. At $S_0 \in \mathcal{S}_\Lambda$, the tangent space of $\mathcal{S}_\Lambda$ is $\{[S_0, A],\ A \in \mathcal{A}\}$ and the normal space is $\mathcal{C}(S)$.*

The *bracket* between to matrices is $[X, Y] = XY - YX$. To make sense of normal spaces, we need an inner product in $\mathcal{M}$: take

$$\langle X, Y \rangle = \operatorname{tr} X^T Y\ , \quad X, Y \in \mathcal{M}(n, \mathbb{R})\ ,$$

which is invariant under translations by orthogonal matrices,

$$\langle X, Y \rangle = \langle Q_1\ X\ Q_2,\ Y \rangle\ , \forall\ Q_1,\ Q_2 \in SO\ .$$

Proof. Diagonalize $S_0 \in \mathcal{S}_\Lambda$ as $S_0 = Q_0^T \Lambda Q_0$ and consider the map

$$F : SO(n, \mathbb{R}) \times \mathcal{D} \to \mathcal{S}\ , \quad (Q, D) \mapsto Q^T Q_0^T (\Lambda + D)\ Q_0 Q\ .$$

Clearly F is smooth and $F(I, 0) = S_0$. We use the inverse function theorem to show that F is a local diffeomorphism between neighborhoods of $(I, 0)$ and S_0.

Derivatives along directions $A \in \mathcal{A}$ are obtained by taking the derivative at $t = 0$ of the expression $(e^{tA})^T S_0 e^{tA}$, yielding $[S_0, A]$. For a curve $t\, d \in \mathcal{D}$ we obtain for derivative the matrix $Q_0^T\, d\, Q_0 = p(S_0) \in \mathcal{C}(S_0)$. Adding up, the Jacobian of F at $(I, 0)$ is

$$DF(I,0) : \mathcal{A} \times \mathcal{D} \to \mathcal{S}\,, \quad (A, d) \mapsto (\ [S_0, A]\ ,\ Q_o^T\, d\, Q_0\ )\,.$$

To show invertibility of $DF(I, 0)$, we first show that matrices $[S_0, A]$ are orthogonal to matrices $p(S_0) \in \mathcal{C}(S_0)$:

$$\langle\, [S_0\,, A]\,, p(S_0)\,\rangle = \mathrm{tr}(S_0 A^- A S_0)^T\, p(S_0)$$

$$= \mathrm{tr}\,(S_0 A - A S_0)\, p(S_0) = \mathrm{tr}\ A\,[p(S_0)\,, S_0] = 0,$$

since $\mathcal{C}(S_0)$ is a commutative algebra. We prove that $DF(I, 0)$ is injective on the restrictions to $\mathcal{A}$ and $\mathcal{D}$: the general case $Q_0 \in SO$ reduces to $Q_0 = I$, for which injectivity is a simple verification.

As for compactness, $\mathcal{S}_\Lambda$ lies in the sphere centered at 0 through Λ. Finally, $\mathcal{S}_\Lambda$ is not only orientable, it is parallelizable: for each $S \in \mathcal{S}_\Lambda$, consider the independent vector fields $[S, A], A \in \mathcal{A}$. □

Exercise 16. Show by using tensor products that if S has simple spectrum, $A \in \mathcal{A} \mapsto [S, A] \in \mathcal{S}$ is injective.

Exercise 17. The manifold $\mathcal{S}_\Lambda$ is very similar to SO: $\mathcal{S}_\Lambda$ is a quotient of SO by the (discrete) subgroup of its diagonal matrices. In particular, SO covers (2^{n-1} times) $\mathcal{S}_\Lambda$.

3.5.1 More isospectral manifolds

Keep Λ with simple spectrum. It turns out that for certain vector spaces $V \subset \mathcal{S}$ the intersection $V \cap \mathcal{S}_\Lambda$ is still a (compact) manifold. This is the case when $\mathcal{T}_\Lambda = V \cap \mathcal{S}_\Lambda$ for V consisting of tridiagonal matrices. The result is harder: the group structure is not available any more. The first proof was given in [59] and a second yielding charts in [36], which extends the result for vector spaces of matrices with a fixed *profile*: all entries below a *staircase* are zero. The dimension of the manifold is the number of possibly nonzero entries strictly below the diagonal. Here is one example of dimension 5. the charts consist of extensions of the Jacobi inverse variables to larger domains and

more general scenarios — notice, for example, that diagonal matrices are not parameterized by Jacobi inverse variables.

$$J = \begin{pmatrix} * & * & * & 0 & 0 \\ * & * & * & 0 & 0 \\ * & * & * & * & 0 \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix}.$$

3.5.2 Two functionals on $\mathcal{S}_\Lambda$

From Galois theory, the roots of polynomials of degree greater than four are rarely written in terms of radicals and the usual arithmetic operations on the coefficients of the polynomial. Thus, the eigenvalues of a matrix can only be approximated with standard arithmetic. In particular, few symmetric matrix are explicitly diagonalized, despite of the fact that tridiagonalizing without changing the spectrum is feasible, from Lanczos's method. Actually, his algorithm is a construction with compass and straight edge, in the sense that only quadratic extensions of the original field of entries are required.

Jacobi had the following wonderful idea to compute (i.e., to approximate arbitrarily well) the spectrum of a symmetric matrix. As an example, we compute the spectrum of S:

$$S = \begin{pmatrix} 3 & 2 & -4 \\ 2 & 3 & 2 \\ -4 & 2 & 3 \end{pmatrix}.$$

First, conjugate S by a rotation on the plane generated by the canonical vectors e_1 and e_3 so as to diagonalize the 2×2 block in the intersection of lines and columns 1 and 3. More generally, define a *Jacobi step* to be the conjugation of a $n \times n$ matrix S by the (Jacobi) rotation $R_{ij}(\theta)$ on the plane spanned by the canonical vectors e_i and e_j of an angle θ. A few simple observations are in order. The only diagonal entries which change value are in positions (i,i) and (j,j). Also, the sum of the squares of these new diagonal entries equal the sum of the squares of the four entries S_{ii}, S_{ij}, S_{ji} and S_{jj} (why? you may think in terms of the matrix inner product on 2×2 matrices).

Consider now the sum of squared diagonal entries

$$J : \mathcal{S}_\Lambda \to \mathbb{R}\,, \quad S \mapsto \sum_i S_{ii}^2\,.$$

From the remarks above, a Jacobi step, taking a matrix $S_0 \in \mathcal{S}_\Lambda$ to $S_1 \in \mathcal{S}_\Lambda$ typically increases the value of J, $J(S_1) \geq J(S_0)$. By compactness of $\mathcal{S}_\Lambda$, J achieves a maximum $S^{\max}$, which is necessarily a diagonal matrix: indeed, if $S^{\max}_{i,j} \neq 0, i \neq j$, then a Jacobi step associated to a rotation $R_{i,j}(\theta)$ increases J.

Thus, we proved the spectral theorem: a real symmetric matrix is orthogonally conjugate to a diagonal matrix. Also, we have a numerical scheme to approximate eigenvalues by diagonal entries. The algorithm provides a simple cumulative measure for how much we deviate from spectrum if we decide to stop at matrix S_k: if S has eigenvalues λ_i and S_k has diagonal entries μ_i, then

$$\sum(\lambda_i^2 - \mu_i^2) = \sum_{i \neq j} (S_k)_{ik}^2\,.$$

There are finer issues related to implementation: one might set to zero the largest off-diagonal entry in each iteration, but finding it is rather expensive. Also, the algorithm is extremely compatible with parallel processing: conjugating by $R_{i,j}$ and $R_{k,\ell}$ can be made with minimal interaction (think of the dimension n as being in the scale of hundreds, thousands). The underlying compactness guarantees that the algorithm is very stable numerically.

Exercise 18. There is an invariant description of the Jacobi step which may look pedantic now but will be convenient later. The rotation $R_{ij}(\theta)$ lies in the curve $e^{tA_{i,j}}$, where $A_{i,j}$ is the skew-symmetric matrix having only two nonzero entries, (i,j) and (j,i), respectively equal to 1 and -1. Said differently, $A = e_i \wedge e_j = e_i e_j^T - e_j e_i^T$, a wedge monomial as in Section 2.6. More generally, rotations on the two dimensional plane spanned by the two orthonormal vectors $u, v, \in \mathbb{R}^n$ are of the form e^{tA}, for $A = u \wedge v$.

Conceptually, there is something peculiar about the Jacobi algorithm: it might be thought as a dynamical system (with some discrete

freedom at each step) converging to one of the $n!$ maxima of the function J — all the diagonal matrices. But what about other critical points of J, say, its minima ? This is not an easy question.

For a simple matrix Λ, consider instead the *weighted trace*

$$T : \mathcal{S}_\Lambda \to \mathbb{R}\,, \quad S \mapsto w_1.S_{1,1} + w_2.S_{2,2} + \cdots + w_n.S_{n,n}$$

for a vector $w \in \mathbb{R}^n$ with $w_1 > w_2 > \cdots > w_n$.

One might think of T as being a *height function* on the manifold $\mathcal{S}_\Lambda$ — in particular, it achieves extrema and one might compute its critical points: they are exactly the $n!$ diagonal matrices of $\mathcal{S}_\Lambda$. The *signature* at a critical point D (i.e., the number of negative eigenvalues of the Hessian of T at the point D) is also an interesting number.

Exercise 19. How does the signature at D relate to the number of *inversions* of the eigenvalues of Λ in D? For example, eigenvalues in descending (resp. ascending) order correspond to the maximum (resp. minimum) of T.

One can obtain a substantial amount of topological information about $\mathcal{S}_\Lambda$ by studying the *Morse decomposition* associated to T. In this case, there are two issues to take into account. First, since $\mathcal{S}_\Lambda$ is so close to SO, this is not necessarily the simplest way of doing it. On the other hand, the same T, when restricted to $\mathcal{T}_\Lambda$, the *tridiagonal isospectral manifold* from Section 3.5.1, provides information which is not amenable from Lie group arguments ([59], [36], [22]).

The second issue is more... serendipitous — is there a numerical algorithm associated to T in the same fashion that Jacobi rotations are related to J ? Welcome to the *Toda lattice* ([21],[13], [60]). Every subject in mathematics has its surprising moments, but the Toda lattice is really a collection of fireworks, one of the great combinations of linear and nonlinear phenomena. Alas, we will not handle the subject in this text. Suffices to say that, for matrices with simple spectrum, it is a vector field in $\mathcal{S}_\Lambda$, i.e., a differential equation which preserves symmetry, the eigenvalues and even the original profile (from Section 3.5.1) of the initial condition. The diagonal matrices are its equilibrium points and more, the weighted trace T is a height function (for any weight w!), i.e., given a solution $S(t) \in \mathcal{S}_\Lambda$, the function $T(S(t))$ is strictly increasing, unless the orbit is a single point, i.e., the initial condition is an equilibrium.

So now it is a differential equation to prove the spectral theorem for matrices with simple spectrum: to diagonalize S_0, follow its orbit $S(t)$ and ... wait: $S(\infty)$ is diagonal! For the general case, take limits.

For a different example, consider the *Wielandt-Hoffmann theorem.*

Theorem 15. *Let A and B be real, symmetric $n \times n$ matrices with eigenvalues α_i and β_j. Then there is a permutation π for which*

$$\sum_i (\alpha_i - \beta_{\pi(i)})^2 \leq ||A - B|| = \mathrm{tr}(A - B)^2 \,.$$

Frequently, a property of a real symmetric matrix is easily verified using the spectral theorem. A property of *two* matrices is more complicated: one can rarely diagonalize two matrices simultaneously.

Proof. Without loss (why?), take A and B with simple spectrum and A diagonal with eigenvalues $a_{11} = \alpha_1 > a_{22}\alpha_2 > \cdots > a_{nn} = \alpha_n > 0$. We need to show the inequality

$$\sum_i (\alpha_i - \beta_{\pi(i)})^2 = \sum_i \alpha_i^2 - 2\sum_i \alpha_i \, \beta_{\pi(i)} + \sum \beta_i^2$$

$$\leq \mathrm{tr}(A - B)^2 = \mathrm{tr}\ A^2 - 2\,\mathrm{tr}\ A\,B + \mathrm{tr}\ B^2 \,,$$

where we used the fact that $\mathrm{tr}\, A\, B = \mathrm{tr}\, B\, A$. Since

$$\sum_i \alpha_i^2 = \mathrm{tr}\ A^2 \quad \text{and} \quad \sum_i \beta_i^2 = \mathrm{tr}\ B^2 \,,$$

we are left with showing that

$$\sum_i \alpha_{\pi(i)} \beta_i \ \geq\ \mathrm{tr}\ A\,B \,.$$

Now, the left hand side is a weighted trace T with weights α_i, so that $\mathrm{tr}\, AB = T(B)$. Let $B = B(0)$ flow with Toda. From the properties above, $\mathrm{tr}\, AB(t)$ increases and obtains is maximum at

$$\mathrm{tr}\ A\,B(\infty) = \sum \alpha_u \, \beta_{\pi(i)} \,,$$

where $B(\infty)$ is a diagonal matrix with diagonal entries given by the eigenvalues of B in some order. □

This proof is from [14].

Chapter 4

Spectrum and convexity

4.1 The Schur-Horn theorem

We start with a classic. As usual, Λ is an $n \times n$ real diagonal matrix with spectrum $\sigma(\Lambda) = \{\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n\}$.

Theorem 16. *(Schur-Horn) The image of the map*

$$H : \mathcal{S}_\Lambda \to \mathbb{R}^n \cap \mathcal{H}_\Lambda \,, \quad S \mapsto \operatorname{diag} S \,,$$

is $\mathcal{P}_\Lambda$, the convex closure of the $n!$ points

$$\{v_\pi = (\lambda_{\pi(1)}, \lambda_{\pi(2)}, \ldots, \lambda_{\pi(n)}, \,) \,, \pi \in S_n\} \subset \mathcal{H}_\Lambda \,.$$

Here, $\mathcal{H}_\Lambda$ is the hyperplane

$$\mathcal{H}_\Lambda = \{\, x \in \mathbb{R}^n \,, \sum_i x_i = \sum_i \lambda_i \,\} \,,$$

the expression $\operatorname{diag} S \in \mathcal{H}_\Lambda$ is the vector with the diagonal entries of S and S_n is the permutation group on the numbers $1, 2, \ldots, n$.

For $n+2$, $\mathcal{P}_\Lambda$ is a segment in $\mathbb{R}^2$. For $n = 3$, a hexagon in $\mathbb{R}^3$ (or in two dimensions, once we restrict its ambient space to be $\mathcal{H}_\Lambda$), which projects injectively to the first two coordinates (see the figure in Step 2 of the proof). For $n = 4$, it is a *permutohedron*, a polyhedron in $\mathbb{R}^3$ displayed in the end of the section.

The inclusion $\operatorname{Ran} H \subset \mathcal{P}_\Lambda$ is due to Schur ([51]). Horn ([25]) proved the harder inclusion. The argument below was presented in the dissertation of Leite ([35]).

The untiring reader might enjoy proving that all the vectors v_π are indeed vertices of $\mathcal{P}_\Lambda$, and hence they are all the vertices of $\mathcal{P}_\Lambda$. Convex polytopes may be described by their vertices or by their faces, or more generally, by the intersection of a collection of half-spaces (i.e., one of the two closed sides of space defined by an affine hyperplane). It turns out that $\operatorname{Ran} H = \mathcal{P}_\Lambda$ is the intersection of the sets

$$x_i \le \lambda_1\,, \quad i=1,\dots,n$$

$$x_i + x_j \le \lambda_1 + \lambda_2\,, \quad i,j=1,\dots,n \quad i \ne j\,,$$

$$\dots$$

$$\Big(\sum_1^n x_i\Big) - x_k \le \sum_1^{n-1} \lambda_i\,, \quad k=1,\dots,n$$

$$\sum_i x_1^n = \sum_1^n \lambda_i\,.$$

This more balanced list of restrictions also describes $\mathcal{P}_\Lambda$:

$$\lambda_n \le x_i\,, \quad i=1,\dots,n$$

$$\lambda_n + \lambda_{n-1} \le x_i + x_j\,, \quad i,j=1,\dots,n\, i \ne j\,,$$

$$\dots$$

$$\sum_2^n \lambda_i \le \Big(\sum_i x_i\Big) - x_k\,, \quad k=1,\dots,n\,.$$

The equivalence may be proved with bare hands (as in [59]) — it is a matter of manipulating inequalities in an appropriate fashion. The reader might instead have a look at [41], where the problem is made to fit into the interesting field of majorization inequalities. Or one could identify the polytope $\mathcal{P}_\Lambda$ as a *dual Weyl chamber*, and the equivalence of both descriptions would be a theorem in Lie algebra theory. It is not surprising that there should be other convex polytopes hanging around associated to similar theorems ([25], [35]).

Before proceeding with the proof, it is worth considering the really simple case $n=2$. Without loss (why?), suppose $\Lambda = \mathrm{diag}(1,0)$. In this case, the image of F lies in the line $x+y=1+0$ in the plane. The two diagonal matrices with spectrum equal to Λ correspond to the vectors $(0,1)$ and $(1,0)$. Perhaps the simplest way to confirm the theorem is to use the spectral theorem,

$$S \in \mathcal{S}_\Lambda \Leftrightarrow S = \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix} = \begin{pmatrix} c^2 & cs \\ cs & s^2 \end{pmatrix},$$

where $c = \cos\theta, s = \sin\theta$ for $\theta \in [0, \pi/2]$.

In the general case, the scheme of the proof is geometric and simple. Since the domain of H is compact, $\mathrm{Ran}\, H$ must be a compact set in $\mathcal{H}_\Lambda$ and probably its boundary $\partial \mathrm{Ran}\, H$ identifies it. Suppose without loss (...) that Λ has simple spectrum. For a generic matrix $S \in \mathcal{S}_\Lambda$, we should expect the derivative $DH(S) : \mathbb{R}^N \to \mathcal{H}_\Lambda \simeq \mathbb{R}^{n-1}$ to be surjective, since $N \gg n$ — at such *regular* points S_r, from the local form of a surjection, H is open, i.e., an open neighborhood of S_r is taken to an open set containing $H(S_r)$:

Only critical (i.e., non-regular) points can be taken to $\partial \mathrm{Ran}\, H$.

Some notation will be helpful. A subspace $V \subset \mathbb{R}^n$ is *canonical* if it is spanned by vectors of the canonical basis of $\mathbb{R}^n$. Canonical subspaces come in pairs: V and $V^\perp$ are simultaneously canonical. A matrix S which has a nontrivial invariant canonical subspace *splits*. Indeed, consider a simple example: if V is spanned by the first k canonical vectors, only the entries on the $k \times k$ top principal minor and on the $(n-k) \times (n-k)$ bottom principal minor can be nonzero.

The orthogonal complement of $\mathcal{H}_\Lambda$ is spanned $\mathbf{1} = (1, 1, \ldots, 1)$ and we denote by $\mathcal{H}_0$ the parallel subspace $\mathbf{1}^\perp$.

Step 1. Computing the critical set $\mathcal{C}$ of H.

We prove that a matrix S is critical if and only if it splits. First, we compute $DH(S) : T_S\mathcal{S}_\Lambda \to \mathcal{H}_0$. From Theorem 14,

$$T_S\mathcal{S}_\Lambda = \{\, [S, A] = SA - AS,\ A \in \mathcal{A}(n, \mathbb{R}) = \mathcal{A} \,\}.$$

Then $DH(S).[S, A] = \mathrm{diag}[S, A]$ and $DH(S)$ is not surjective if and only if there is a nonzero $w \in \mathcal{H}_0$ for which

$$\langle\, \mathrm{diag}\,[S, A]\,,\, w \rangle = \mathrm{tr}\,[\,S\,,\,A]\,W = 0\,, \quad \forall\, A \in \mathcal{A}$$

where W is the diagonal matrix having w along its diagonal (by the way, show that $\mathrm{diag}[S,A] \in \mathbf{1}^\perp$). Thus $S \in \mathcal{C}$ if and only if

$$\mathrm{tr}\ A\,[S,W] = 0\,, \quad \forall\, A \in \mathcal{A}\,.$$

Thus, the matrix $[S,W]$ is orthogonal to all real skew-symmetric matrices under the matrix inner product considered in Section 3.5: it is easy to show that such matrices are exactly the real symmetric matrices, so $[S,W] \in \mathcal{S}$. On the other hand, both S and W are symmetric, which trivially implies that $[S,W]$ is *skew*-symmetric: there is only one possibility, $[S,W]=0$ — in words, S and W commute.

The orthogonality of w and $\mathbf{1}$ implies that $\mathrm{tr}\,WI=0$, so that the diagonal matrix W has at least two distinct eigenvalues. Split the eigenvalues of W in two nonempty disjoint sets sharing no common eigenvalue and take a polynomial g take one set to 0 and the other to 1. Then $g(W)$ is a diagonal matrix with spectrum $\{0,1\}$.

Since S commutes with W, it must commute with $g(W)$ (proof?). But then $V = \mathrm{Ran}\,g(W)$ is a nontrivial invariant canonical subspace of S (write down the matrices if you don't see why): S splits.

The converse — if S splits then $S \in \mathcal{C}$ — is easy: any invariant canonical subspace of S gives rise to a diagonal matrix W with eigenvalues 0 and 1 so that $[S,W]=0$ and then

$$\mathrm{tr}\,[S,A]\,W = -\,\mathrm{tr}\,[S,W]\,A = 0\,, \quad \forall\, A \in \mathcal{A}\,.$$

We consider $n=3$, $\Lambda = \mathrm{diag}(4,2,1)$. The critical set decomposes into nine ways of splitting. Each split yields a pair $(V, V^\perp) \simeq (V^\perp, V)$ for which V may taken to be of dimension one, and hence spanned by some canonical vector. Having fixed V, one has to choose the eigenvalue of the restriction of S to V. The x's stand for real numbers.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & x & x \\ 0 & x & x \end{pmatrix} \quad \begin{pmatrix} 2 & 0 & 0 \\ 0 & x & x \\ 0 & x & x \end{pmatrix} \quad \begin{pmatrix} 4 & 0 & 0 \\ 0 & x & x \\ 0 & x & x \end{pmatrix}$$

$$\begin{pmatrix} x & 0 & x \\ 0 & 1 & 0 \\ x & 0 & x \end{pmatrix} \quad \begin{pmatrix} x & 0 & x \\ 0 & 2 & 0 \\ x & 0 & x \end{pmatrix} \quad \begin{pmatrix} x & 0 & x \\ 0 & 4 & 0 \\ x & 0 & x \end{pmatrix}$$

$$\begin{pmatrix} x & x & 0 \\ x & x & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} x & x & 0 \\ x & x & 0 \\ 0 & 0 & 2 \end{pmatrix} \quad \begin{pmatrix} x & x & 0 \\ x & x & 0 \\ 0 & 0 & 4 \end{pmatrix}$$

Each such set is a circle and they meet at diagonal matrices.

In general, the critical set $\mathcal{C}$ decomposes into *chunks* $\mathcal{C}_{V,\sigma_V}$ consisting of matrices which have V as invariant canonical subspace and whose restrictions $S_V : V \to V$ have spectrum $\sigma_V \subset \sigma(\Lambda)$. The *partial trace* $\operatorname{tr}_V S$ of S in V is equal to

$$\operatorname{tr}_V S = \operatorname{tr} S_V = \sum_{\lambda_i \in \sigma_V} \lambda_i \,.$$

Step 2. Computing the image $H(\mathcal{C})$.

The image of chunk $\mathcal{C}_{V,\sigma_V} \subset \mathcal{C}$ lies in the (affine) hyperplane

$$\mathcal{H}_{V,\sigma_V} = \{y \in \mathcal{H}_\Lambda \,|\, \langle\, e_V \,,\, y \,\rangle = \sum_{\lambda_i \in \sigma_V} \lambda_i\} \,,$$

where e_V is the vector with coordinate i equal to 0 and 1, depending if $e_i \in V$ or not. Then

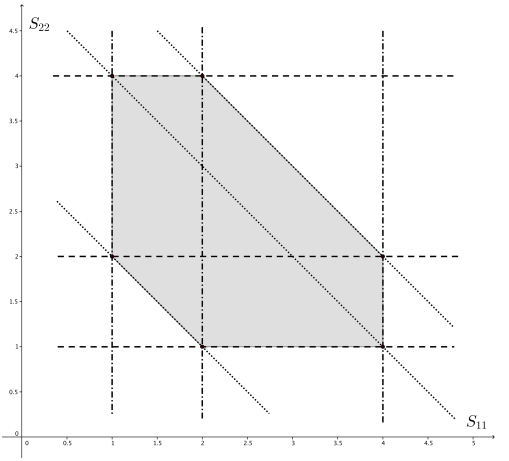
$$\operatorname{tr}_V S = \operatorname{tr} E_V S E_V = \operatorname{tr} S E_V = \langle\, e_V \,,\, \operatorname{diag} S \,\rangle \,.$$

The orthogonal projection E_V on V is the diagonal matrix having the vector e_V along its diagonal. Clearly S commutes with E_V.

Chunks associated to the same V are taken to parallel hyperplanes, differing only by the choice of eigenvalues of S_V. For a given V, such family of parallel hyperplanes has two extremal elements, when the partial trace equals the smallest and the greatest possible sums of σ_V. If V is of dimension k, those numbers are the sum of the k smallest or the k largest eigenvalues of Λ.

We have just proved Schur's inclusion: $\operatorname{Ran} H \subset \mathcal{P}_\Lambda$, where $\mathcal{P}_\Lambda$ is expressed in terms of (balanced) inequalities.

Again, the case $n = 3$ is informative: take for spectrum the set $\{1, 2, 4\}$. In the picture, we see the three families of parallel hyperplanes. Notice that only the first two diagonal entries are plotted:



the last one is obtained from the (fixed) trace. The hyperplanes of the form $S_{33} = c$ correspond to lines of the form $S_{11} + S_{22} = 7 - c$.

So far, we know that $\partial \operatorname{Ran} H$ is contained in a *grid* $\mathcal{G}$ consisting of families of parallel hyperplanes. This is already intriguing: since a dense set of points in the domain consists of regular points (why?), the interior of $\operatorname{Ran} H$ is dense in $\mathcal{H}_\Lambda$, so $\operatorname{Ran} H$ itself consists of the closure of the union of a collection of connected component of the set $\mathcal{H}_\Lambda \setminus \mathcal{G}$, each component a parallelotope in $\mathcal{H}_\Lambda$. Thus, if one of such parallelotopes meets $\operatorname{Ran} H$ then it is completely included in $\operatorname{Ran} H$.

Step 3. Getting rid of fake walls.

We now show that in each family of parallel hyperplanes, only the two extreme hyperplanes act as barriers to $\operatorname{Ran} H$ — this essentially proves the theorem. The reader should return to the picture for $n = 3$. Each family consists of three hyperplanes (in this case, lines) — removal of the central line on each family yields the theorem.

First, notice that the intersection of two such hyperplanes is an (affine) plane of codimension two in $\mathcal{H}_\Lambda$. Let $\mathcal{D}$ be the set of points in the grid $\mathcal{G}$ which belong to at least two different hyperplanes: not only $\mathcal{D}$ has empty interior in $\mathcal{H}_\Lambda$, but $\mathcal{P}_\Lambda \setminus \mathcal{D}$ is still connected. A point in $\mathcal{G} \setminus \mathcal{D}$ belongs to a single hyperplane $\mathcal{H}_{V,\sigma_V}$.

If $\mathcal{H}_{V,\sigma_V}$ is not extreme within its family, then σ_V consists of numbers which are not the smallest or the largest in $\sigma(\Lambda)$. Say σ_V does not yield the smallest (resp. largest) sum: there is $\lambda_{in} \in \sigma_V$ and $\lambda_{out} \notin \sigma_V$ with $\lambda_{in} > \lambda_{out}$ (resp. $\lambda_{in} < \lambda_{out}$).

Suppose that $S \in \mathcal{S}_\Lambda$ obtains $H(S) = \operatorname{diag} S \in \mathcal{H}_{V,\sigma_V} \cap \mathcal{G} \setminus \mathcal{D}$ for a hyperplane $\mathcal{H}_{V,\sigma_V}$ which is not an extremal of this family. We show that there are matrices $S_+, S_- \in \mathcal{S}_\Lambda$ close to S for which

$$\langle\, e_V\,,\, \operatorname{diag} S_-\,\rangle \,<\, \langle\, e_V\,,\, \operatorname{diag} S\,\rangle = \sum_{\lambda_i \in \sigma_V} \lambda_i \,<\, \langle\, e_V\,,\, \operatorname{diag} S_+\,\rangle,$$

and thus the parallelotopes on both sides of the hyperplane $\mathcal{H}_{V,\sigma_V}$ at the point $H(S)$ belong to $\operatorname{Ran} H$.

Without loss, say $\mathcal{H}_{V,\sigma_V}$ does not yield the minimal possible sum for the eigenvalues in $\sigma(V)$. Consider the two dimensional plane spanned by the orthonormal eigenvectors v_{in} and v_{out} associated to the eigenvalues λ_{in} and λ_{out} of S. We perform conjugations $R(-\theta)SR(\theta) \in \mathcal{S}_\Lambda$ by Jacobi rotations in this plane of an angle θ, as in Exercise 18. We are interested in

$$\alpha(t) = \langle\, e_V\,,\, H(\, e^{-tA}\, S\, e^{tA}\,)\,\rangle\,, \quad \text{where} \quad A = v_{in} \wedge v_{out}\,.$$

Expand the curve of matrices near $t = 0$:

$$e^{-tA}\, S\, e^{tA} = S + t\,[\,S\,,\,A\,] + t^2\big(\,\frac{A^2}{2}\,S - A\,S\,A + S\,\frac{A^2}{2}\,\big) + O(t^3)\,,$$

so that, in particular,

$$\frac{d}{dt}\alpha(0) = \langle\, e_V\,,\, \operatorname{diag}[\,S\,,\,A\,]\,\rangle = \operatorname{tr} E_V\,[\,S\,,\,A\,] = \operatorname{tr}[\,E_V\;S\,]\,A\,] = 0\,,$$

since E_V and S commute — indeed, the image of the chunk $\mathcal{C}_{V,\sigma_V}$ containing S is sent to the hyperplane $\mathcal{H}_{V,\sigma_V}$, and $\alpha(t)$ can only deviate quadratically from it. We now consider the quadratic term of the expansion of α at $t = 0$,

$$Q = \langle\, e_V, \operatorname{diag}\big(\,\frac{A^2}{2}\,S - A\,S\,A + S\,\frac{A^2}{2}\big)\rangle = \operatorname{tr} E_V\,\big(\,\frac{A^2}{2}\,S - A\,S\,A + S\,\frac{A^2}{2}\,\big)$$

which simplifies considerably if we apply the following algebraic facts:

$$E_V^2 = E_V,\; E_V\,S = S\,E_V,\; E_V\,v_{in} = v_{in},\; E_V\,v_{out} = 0\,.$$

In particular, $E_V A = v_{in} v_{out}^T$ and $AE_V = -v_{out} v_{in}^T$) and

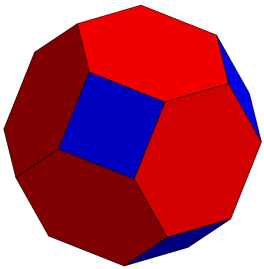
$$Q = \frac{1}{2} \operatorname{tr} v_{in} v_{in}^T S + \operatorname{tr} v_{in} v_{out}^T S\, v_{out}\, v_{in}^T - \frac{1}{2} \operatorname{tr} S\, v_{out} v_{out}^T$$

$$= \frac{1}{2}(\lambda_{in} - \lambda_{out}) > 0\,.$$

Thus, for small t we have $\alpha(t) > \alpha(0)$: take $S_+ = S(t) = e^{-tA} S\, e^{tA}$.

To get S_-, choose $\lambda_{in} \in \sigma_V$ and $\lambda_{out} \notin \sigma_V$ with $\lambda_{in} < \lambda_{out}$ and proceed exactly as above.

Exercise 20. The argument above computes second derivatives of $\alpha(t) = \langle\, e_V\,,\, H(\,S(t)\,)\,\rangle$ for special curves $S(t) = e^{-tA} S\, e^{tA}$. Show that this is sufficient to compute the full Hessian of α at $t = 0$ (hint: the polarization identity). More, the computations above obtain the eigenvalues of the Hessian. It is not an invertible matrix and thus α is not a Morse function, but a *Morse-Bott* function.



Step 4. Rounding up.

The points in $\partial \operatorname{Ran} H$ belong to extreme faces (which correspond to the balanced equalities presented after the statement of the theorem) and possibly by the points in $\mathcal{D}$, which consist of a thin set which does not disconnect $\operatorname{Ran} H$ — by compactness of $\mathcal{S}_\Lambda$, $\mathcal{D} \subset \operatorname{Ran} H$ and we are done. For completeness, the untiring reader might show that the extreme hyperplanes indeed generate nontrivial faces. □

In the picture, the polytope associated to the case $n = 4$.

There are eight hexagons corresponding to the extreme cases V spanned by one of the four canonical vectors e_i and σ_V equal to λ_1 or λ_4. The six squares correspond to the six possible two dimensional subspaces V and $\sigma_V = \{\lambda_1, \lambda_2\}$.

4.2 Mutations, the high and low roads

The Schur-Horn theorem is so interesting that it deserves additional contemplation. Kostant ([30]) obtained a first generalization in a Lie algebraic context. In 1982, both Atiyah ([2]) and Guillemin and Sternberg ([23]) obtained beautiful results related to the convexity of the image of a moment map of a Hamiltonian action of a torus: they imply the *complex* version of the Schur-Horn theorem, in which $\mathcal{S}_\Lambda$ is replaced by its complex counterpart $\{S = U^* \Lambda U \, , U \in SU(n)\}$.

Duistermaat ([17]) later obtained a real counterpart of such results from which the real Schur-Horn theorem follows immediately. These theorems rely on basic facts of symplectic geometry, namely equivariant versions of the so called Darboux theorem. The counterpart of the symplectic (complex) arguments to the statement that most walls are fake is trivial. In a sense, the real case of the Schur-Horn-theorem (which by the way implies the complex case) is harder.

There are variations of the Schur-Horn theorem for orthogonal and skew-symmetric matrices. The counterpart for singular values instead of eigenvalues is the Sing-Johnson theorem ([52],[56]). The results admit proofs with different levels of sophistication, which led Thompson ([57]) to comment on high and low levels in linear algebra.

We present now a different kind of convexity result. For a real diagonal matrix with simple spectrum Λ, let $\mathcal{J}_\Lambda$ be the set of Jacobi matrices (real, symmetric, tridiagonal matrices with strictly positive entries in coordinates $(i, i-1)$) with spectrum Λ.

Theorem 17. *$\overline{\mathcal{J}_\Lambda} \simeq \mathcal{P}_\Lambda$, in the sense that there is a homeomorphism between both spaces which is a diffeomorphism between interiors.*

The result was originally proved in [59]. Later, Bloch, Flaschka aand Ratiu ([5]) managed to phrase it in terms of a moment map, and obtained an explicit diffeomorphism. A low road argument was provided by Leite, Richa and Tomei ([35]).

Theorem 18. *Let $T \in \overline{\mathcal{J}_\Lambda}$ and consider any spectral decomposition $T = Q^T \Lambda Q, Q \in SO(n, \mathbb{R})$. Then the map*

$$BFR : \overline{\mathcal{J}_\Lambda} \to \mathcal{P}_\Lambda \quad T \mapsto \operatorname{diag} Q \Lambda Q$$

realizes explicitly the identification in the previous theorem.

4.3 Interlacing and more

4.3.1 Rank one perturbations

Say S is a real $n \times n$ symmetric matrix, with simple eigenvalues

$$\sigma(S) = \{\, \lambda_1 < \lambda_2 < \ldots < \lambda_n \,\}\,.$$

What may happen to the spectrum when we add a (real, symmetric) matrix P of rank one? The answer is very satisfactory.

We introduce notation. Without loss, we may suppose

$$S = \Lambda = \mathrm{diag}(\, \lambda_1 < \lambda_2 < \ldots < \lambda_n \,)$$

and $P = t\, v \otimes v = t\, v\, v^T$, for $||v|| = 1$, $t > 0$: we are interested in the eigenvalues of $\Lambda + t\, v \otimes v$.

It is clear that removing the signs of v has no effect in the problem. Indeed, if E is a diagonal sign matrix (i.e., a diagonal matrix with ± 1 along its diagonal entries), then the matrices

$$\Lambda + t\, v \otimes v \quad \text{and} \quad E(\Lambda + t\, v \otimes v)E^{-1} = \Lambda + t\,(Ev) \otimes (Ev)$$

have the same spectrum. Define

$$\overline{Q_+^n} = \{v \in \mathbb{R}^n \,|\, ||v|| = 1\,, \quad v_k \geq 0\,\}\,, \quad \lambda = (\lambda_1\, \ldots\, , \lambda_n) \in \mathbb{R}^n\,,$$

$$\mathcal{B} = [\,\lambda_1, \lambda_2\,] \,\times\, [\,\lambda_2, \lambda_3\,] \,\times\, \ldots \,\times\, [\lambda_n, \infty)\,,$$

$$\mathbb{R}^+ = \{t > 0\}\,, \quad \mathbb{R}_o^n = \{x \in \mathbb{R}^n \,|\, x_1 \leq x_2 \leq \ldots \leq x_n\}\,.$$

For $(v, t) \in \overline{Q_+^n} \,\times\, \mathbb{R}^+$, let $(\mu_1\, \ldots\, , \mu_n\,) \in \mathbb{R}_o^n$ be the ordered eigenvalues of the matrix $\Lambda + t\, v \otimes v$.

Theorem 19. *The map*

$$F :\, \mathcal{D} = \overline{Q_+^n} \,\times\, \mathbb{R}^+ \to \mathbb{R}_o^n\,, \quad (\, v\,, t\,) \mapsto\ (\mu_1\, \ldots\, , \mu_n\,)$$

induces a homeomorphism between $\mathcal{D}$ and $\mathcal{B} - \lambda$ which restricts to a diffeomorphism between the interior of both sets, $\mathcal{D}^o$ and $\mathcal{B}^o$.

The proof follows the argument of Theorem 16, and is simpler in some aspects. For $n = 2$, consider the map taking $(v, t) \in \mathcal{D}$ to $\mathbb{R}^2_o$, for the parametrization $v = (c, s)$, $c = \cos\theta$, $s = \sin\theta$, $\theta \in [0, \pi/2]$:

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} + t \begin{pmatrix} c \\ s \end{pmatrix} \begin{pmatrix} c & s \end{pmatrix} = \begin{pmatrix} \lambda_1 + t\,c^2 & c\,s \\ c\,s & \lambda_2 + t\,s^2 \end{pmatrix} \mapsto (\mu_1\,, \mu_2)\,,$$

The domain $\mathcal{D}$ is the rectangle $[0, \pi/2] \times (0, \infty)$. We are especially interested in the behavior of F at $\partial\mathcal{D} = \{0, \pi/2\} \times (0, \infty)$, the part of the boundary of $\mathcal{D}$ in $\mathcal{D}$. For $\theta = 0$, the eigenvalues $\mu_1 \leq \mu_2$ are $\lambda_1 \leq \lambda_2 + t$. But for $\theta = \pi/2$, things are slightly more complicated: the eigenvalues are $\lambda_1 + t$ and λ_2, so that

$$\mu_1 = \lambda_1 + t\,, \quad \mu_2 = \lambda_2\,, \quad \text{for} \quad t \in (0, \lambda_2 - \lambda_1]\,,$$

$$\mu_1 = \lambda_2\,, \quad \mu_2 = \lambda_1 + t\,, \quad \text{for} \quad t \in [\lambda_2 - \lambda_1, \infty)\,.$$

The picture clarifies matters: one boundary is sent to one face of $\mathcal{B} = [\lambda_1, \lambda_2] \times [\lambda_2, \infty)$, while the other occupies two. The picture also suggests (and a computation confirms) that as $t \to 0$, the image of $F(v, t)$ goes to the point $\lambda = (\lambda_1, \lambda_2)$. In the same fashion, as $t \to \infty$, $F(v, t)$ goes to $[\lambda_1, \lambda_2] \times \{\infty\}$, or, more simply, to ∞ — said differently, F is a proper map (inverse of compact sets of $\mathbb{R}^2_o$ are compact sets of $\mathcal{D}$) and this allows us to handle F as if it were a function between compact spaces in some arguments.

From a simple computation (which will be done for the general case), F is differentiable in $\mathcal{D}^o$ and has no critical points. Because of properness, we can apply now (essentially) the same fact that we used in the proof of the Schur-Horn theorem 16:

The boundary of the image of F is the union of three sets: the image of the boundary of its domain, the image of the critical set $\mathcal{C} \subset \mathcal{D}^o$ of F (which is empty) and the image of the set of nondifferentiable points of F (which lies in the first set anyway).

Once the identification of these three sets is accomplished, the proof follows as in Theorem 16 with minor modifications. Because of properness, the image of F has to stay in the box $\mathcal{B}$. Also, F takes $(\partial\mathcal{D}) - \lambda$ to $\partial\mathcal{B} - \lambda$ *injectively*: to see this, we only have to show that F takes $\mathcal{D}^o$ to $\mathcal{B}^o$ bijectively, and global injectivity follows from

degree theory. Alternatively, take a few more steps. By properness, once a point in $\mathcal{B}^o$ is attained, the full interior is, since the image of F must be open and closed — this settles surjectivity to $\mathcal{B}-\lambda$. As for injectivity, use a *monodromy argument*: say $x, y \in \mathcal{D}^o$ are taken to the same point $p \in \mathcal{B}^o$. A path $\gamma \subset \mathcal{D}^o$ joining x to y gives rise to a closed curve $F(\gamma) \subset \mathcal{B}^o$ with endpoints at p. Now deform $F(\gamma)$ in $\mathcal{B}^o$ to the constant path p and show that the deformation can be pulled back to the domain, clearly a contradiction.

In the n-dimensional case, the box $\mathcal{B}$ has $2n-1$ faces: to describe a face, fix an endpoint of one interval and use the others. In the last interval we do not consider the endpoint ∞. Thus each coordinate (but the last) gives rise to a *bottom* and a *top* face.

We are ready for the proof.

Proof. The function F is differentiable when all the eigenvalues μ_i are distinct — we start showing that this is the case for points in $\mathcal{D}^o$.

- $\sigma(\Lambda + t\, v \otimes v)$ is simple if $v \notin \partial\, \overline{Q^n_+}$.

  A double eigenvalue μ has an eigenvector z for which $z_1 = 0$. Since

$$\Lambda\, z \;+\; t\, v\, \langle\, v\,, z\,\rangle = \mu z\,,$$

we have $v_1\, \langle\, v\,, z\,\rangle = 0$. If $v_1 = 0$, then clearly $v \in \partial \overline{Q^n_+}$. If $\langle\, v\,, z\,\rangle = 0$, then $\Lambda\, z = \mu\, z$, so that μ is some eigenvalue λ_i of Λ and $z = e_i$, a canonical vector (recall that Λ has simple spectrum). But then

$$\Lambda\, e_i \;+\; t\, v\, \langle\, v\,, e_i\,\rangle = \lambda_i e_i$$

and some coordinate of v must be zero — again, $v \in \partial \overline{Q^n_+}$.

  In particular, F is differentiable in $\mathcal{D}^o$.

- There are no critical points in $\mathcal{D}^o$.

We compute the derivative of F, or better, of an equivalent function G. Take $u = \sqrt{t} v$ for $(v, t) \in \mathcal{D}^o$, and consider the matrix $\Lambda + u \otimes u$ with distinct eigenvalues μ_k and (normalized) eigenvectors w_k,

$$G(u) = (\mu_1(u)\,, \mu_2(u)\,, \ldots\,, \mu_n(u)) \;\in \mathbb{R}^n_o\,, \quad \Lambda\, w_k \;+\; u \otimes u\, w_k = \mu_k\, w_k\,.$$

We take directional derivatives $\partial_{w_j}\mu_j$ along w_k. From Section 3.2.1,

$$\partial_{w_j}\mu_k = \langle \partial_{w_j}\big(\Lambda + u\otimes u\big)\, w_k\,, w_k\,\rangle$$

$$= \langle \big(\, w_j\otimes u + u\otimes w_j\,\big)\, w_k\,, w_k\,\rangle = 2\,\langle\, w_j\,, w_k\,\rangle\,\langle\, u\,, w_k\,\rangle\,.$$

Thus, the differential of G at a point u is the diagonal matrix

$$DG(u) = \mathrm{diag}(\,\langle\, u\,, w_1\,\rangle\,, \langle\, u\,, w_2\,\rangle\,, \ldots,\, \langle\, u\,, w_n\,\rangle\,)\,,$$

which is invertible provided $\langle\, u\,, w_k\,\rangle \neq 0$ for all k. If not, from the eigenvector equation for some k,

$$\Lambda w_k + u\otimes u\, w_k = \mu_k\, w_k$$

so that $\Lambda w_k = \mu_k w_k$ and w_k equals some canonical vector e_i. Since $\langle\, u\,, e_i\,\rangle = 0$, the i-th coordinate of u is zero, a contradiction: $u\in\mathcal{D}^o$.

- The map F is proper.

  If λ_k are the eigenvalues of a real, symmetric matrix M, then

$$\sum_k \lambda_k^2 = \mathrm{tr}\,\langle M\,, M\,\rangle = \mathrm{tr}\ M^2\,.$$

(Without loss take S diagonal: this is obvious). For $S = \Lambda + t\, v\otimes v$,

$$\sum_k \mu_k^2 = \mathrm{tr}(\Lambda + tv\otimes v)^2 = \mathrm{tr}\,\Lambda^2 + 2t\langle\, v\,, \Lambda\, v\,\rangle + t^2\,.$$

Thus a sequence $F(v_n, t_n)$ in the image converges to λ if and only if $t_n\to 0$ and converges to ∞ if and only if $t_n\to\infty$.

We now compute the image of the boundary of the domain.

- If $v_i = 0$, then either μ_{i-1} or μ_i equals λ_i.

  We consider the eigenvalues of $S = F(v) = \Lambda + tv\otimes v$,

$$S = \begin{pmatrix} * & 0 & * \\ 0 & \lambda_i & 0 \\ * & 0 & * \end{pmatrix}\,.$$

The juxtaposition of the four blocks $(*)$ gives rise to a $(n-1)\times(n-1)$ matrix $\tilde{S}=\tilde{\Lambda}+\tilde{v}\otimes\tilde{v}$, where $\tilde{\Lambda}$ is obtained from Λ by removal of the i-th row and column and $\tilde{v}$ from v by removal of $v_i=0$.

Clearly, $\lambda_i \in \sigma(F(u))=\{\mu_k\}$. By induction, the ordered eigenvalues $\tilde{\mu}=(\tilde{\mu}_1,\dots,\tilde{\mu}_{n-1})$ of $\tilde{S}$ lie in the box

$$[\lambda_1\,,\lambda_2]\times[\lambda_2\,,\lambda_3]\times\ldots\times[\lambda_{i-1}\,,\lambda_{i+1}]\times\ldots\times[\lambda_{n-1}\,,\lambda_n]\times[\lambda_n\,,\infty]\,.$$

The ordered eigenvalues μ of S are obtained from $\tilde{\mu}$ of $\tilde{S}$ by inserting $\mu_k=\lambda_i$ among the coordinates of $\tilde{\mu}$. Clearly,

$$\mu_k\in[\lambda_{i-1}\,,\lambda_{i+1}]=[\lambda_{i-1}\,,\lambda_i]\cup[\lambda_i\,,\lambda_{i+1}]\,,$$

so that $F(v)$ lies necessarily on the top face of coordinate $i-1$ or on the bottom face of the i-th coordinate.

We leave to the reader the inductive proof of the next statement.

• The restriction $F:\partial\mathcal{D}\to\partial\mathcal{B}\setminus\lambda$ is a bijection.

Now it's a matter of filling up, i.e., of showing that $F:\mathcal{D}\to\mathcal{B}\setminus\lambda$ is a bijection — this is goes exactly as in the 2×2 case. From Section 3.2.2, F extends continuosly, hence homeomorphically. □

A few remarks are in order. Clearly, the eigenvalues λ_k and μ_k interlace. When $t>0$, eigenvalue μ_k usually trespasses λ_k: informally, eigenvalues are pushed to the right. When $t<0$ interlacing still holds and the eigenvalues are pushed to the left.

There is nothing sacred about the positive quadrant $\overline{\mathbb{Q}^n_+}$ — the theorem holds for each quadrant, so given two strictly interlacing spectra λ and μ, there are actually 2^n rank one perturbations $v\otimes v$ for which $\sigma(\Lambda+tv\otimes v=\mu$, Λ being simple.

An interlacing theorem of this form is also true sometimes in infinite dimensions. Say, for example, that instead of Λ one has a self-adjoint operator with spectrum which is bounded from below, possibly starting with some isolated eigenvalues. The operator obtained by adding a symmetric rank one perturbation is still self-adjoint and there is interlacing of spectra until something nasty happens (i.e., essential spectrum). The proof uses min-max.

4.3.2 The sum of two Hermitian matrices

Now fix t in the previous section: consider the possible values of $\sigma(S + v \otimes v)$, where $||v||^2 = c$, for example. The ordered spectra have to lie in $\mathcal{B}$ defined in Theorem 19, but there is one more restriction:

$$\operatorname{tr}(S + v \otimes v) = \operatorname{tr} S + c\,,$$

which is a hyperplane $\mathcal{H}$ intersecting $\mathcal{B}$ in a convex set which may combinatorially more complicated then a box (for example: a plane can intersect a cube along a hexagon). It turns out that lying in $\mathcal{H} \cap \cap B$ is not only necessary but also a sufficient condition.

We phrase the question differently: what are the possible spectra of the sum of two real, symmetric matrices, one with eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$, the other with eigenvalues $\{c, 0, 0, \ldots,\ 0\}$?

In 1912, Hermann Weyl asked, what are the possible spectra of the sum $S + T$ of two real, symmetric matrices, if we fix $\sigma(S)$ and $\sigma(T)$? Some partial results were obtained until Horn ([26]) suggested a complete list of linear inequalities: the answer to this problem is again a convex polytope! His conjecture is now a theorem, following from very interesting work by a sequence of authors, among them Helmke, Klyachko, Knutson, Lidskii, Rosenthal and Tao ([31]).

4.3.3 Weinstein-Aronsjan, Sherman-Morrison

If $Ax = b$ is easy to solve, this should be used to solve $\tilde{A}x = b$ for $\tilde{A}$ near A. Thus, for example, if $\tilde{A}$ is metrically near A, one might consider a recursive algorithm. Here by proximity we mean

$$\tilde{A} = A + u \otimes v = A + u\, v^T\,.$$

The trick is simple: suppose A is invertible and write

$$\tilde{A}\, x = b \iff (A + u \otimes v\,)\, x = A\, x + u\, \langle v\,, x\,\rangle = b\,,$$

so that

$$x + A^{-1}\, u\, \langle v\,, x\,\rangle = A^{-1}\, b$$

and, by taking inner products,

$$\langle v\,, x\,\rangle + \langle v\,, A^{-1}\, u\,\rangle\, \langle v\,, x\,\rangle = \langle\, v\,,\ A^{-1}\, b\,\rangle$$

so that

$$\langle v\,,x\,\rangle = \frac{\langle\, v\,,A^{-1}\,b\,\rangle}{1+\langle\, v\,,A^{-1}\,u\,\rangle}\,,\quad x = A^{-1}\,b\,-\,A^{-1}\,u\,\langle v\,,x\,\rangle\,,$$

and $\tilde{A}$ is not invertible if and only if $1+\langle v\,,A^{-1}\,u\,\rangle = 0$. When A is an infinite dimensional operator, these equations are usually called the Weinstein-Aronsjan formulas ([28]). For matrices, extensions (essentially, the block form of such equations) are the Sherman-Morrison formulas, and are usually associated to matrix *tearing* ([43]).

Exercise 21. To set the record straight, the Weinstein-Aronsjan formulas relate the inverse of $\tilde{A}$ and A when they differ by a rank k perturbation, as opposed to the example above, where $\tilde{A}-A$ is of rank one. Obtain the formulas for this case.

Exercise 22. Try to solve the integro-differential equation

$$u'(x) - \int_0^1 u(t) = g(t)\,,\quad u(0) = u(1)\,.$$

Generically, the eigenvalues of A and $\tilde{A}$ are distinct: imitating Section 2.4.2, we show that the set of pairs (A,v) for which the spectra A and $\tilde{A}$ are disjoint is an open dense set of $\mathcal{M}(n)\times\mathbb{R}^n$. If the resultant

$$R\,(\,\det(A-\lambda\,I)\,,\,\det(A+v\otimes v-\lambda\,I\,)$$

is zero in a nontrivial ball of the product $\mathcal{S}\times\mathbb{R}^n\setminus\{0\}$, then it is identically zero. Let A be a diagonal matrix with distinct eigenvalues, and v a vector with nonzero entries: we now show that A and $\tilde{A}=A+t\,v\otimes v$ have distinct eigenvalues for small t. Indeed, the derivatives of the eigenvalues of $A+t\,v\times v$ are given by

$$\lambda_k'(t) = \langle\, v\times v\,e_k\,,e_k\,\rangle = \langle\, v\,,e_k\,\rangle^2\;\neq\;0\,,$$

and thus $\sigma(A+t\,v\times v)$ and $\sigma(A)$ are disjoint for small t.

We extend the formulas above. Subtract $\lambda\,I$ from A and $\tilde{A}$ by $\tilde{A}$:

$$\tilde{A}-\lambda\,I = A-\lambda\,I+u\otimes v = A-\lambda I+u\,v^T$$

and the solution of $(\tilde{A} - \lambda I) = b$ is then

$$\langle v\,, x\,\rangle = \frac{\langle\, v\,, (A-\lambda I)^{-1}\, b\,\rangle}{1 + \langle v\,, (A-\lambda I)^{-1}\, u\,\rangle}\,,$$

$$x = (A-\lambda I)^{-1}\, b \;-\; \langle v\,, x\,\rangle\, (A-\lambda I)^{-1}\, u\,.$$

Supposing that A and $\tilde{A}$ are no common eigenvalues,

$$g(\lambda) = 1 + \langle v\,, (A-\lambda I)^{-1}\, u\,\rangle = c\, \frac{\det(\tilde{A} - \lambda I)}{\det(A - \lambda I)}$$

and we obtain $c = 1$ by taking $\lambda \to \infty$.

We are very close to another proof of the interlacing theorem. More specifically, take A real symmetric with simple spectrum and $u = v$ and suppose first the generic hypothesis that A and $A + v \otimes v$ have no common eigenvalues — we want to show that the roots and poles of $g(\lambda)$ alternate: simply compute $g'(\lambda)$ for $\lambda \in \mathbb{R}$,

$$g'(\lambda) = \langle v\,, (A-\lambda I)^{-1} (A-\lambda I)^{-1}\, v\,\rangle \;\geq\; 0\,,$$

from which interlacing follows. Take limits to get rid of the generic hypothesis using the continuity of the eigenvalues (Section 3.2.2). The result in Section 4.3.1 is clearly more precise.

Exercise 23. Choose one of the two approaches above to handle another interlacing situation. Take S real symmetric and let $\hat{S}$ be obtained from S by removing the last row and column. Show that the spectra of S and $\hat{S}$ interlace.

Chapter 5

The spectral theorem

The spectral theorem is about generalizing the finite dimensional diagonalization process. Indeed, in one of its forms, it conjugates a self-adjoint operator to a normal form, a multiplication operator, which is pretty similar to a diagonal matrix. But this is only partially right.

The spectral theorem interpreted as a *functional calculus* is at least as relevant. As in Section 3.1, let $\mathcal{B}(X)$ be the algebra of bounded linear transformations with the operator norm on a Banach space X. For an operator $T \in \mathcal{B}(X)$, polynomials $p(T)$ and entire functions like e^T make sense. To go further we need to take limits, which require finer estimates: enter complex variables.

5.1 The Dunford-Schwartz calculus

Recall (as if one could forget) Cauchy's theorem from the basic complex variable course, presented without any effort towards generality.

Theorem 20. *Let γ be a smooth, positively oriented simple curve bounding an open set $\Omega \subset \mathbb{C}$. Let $f : \Omega \to \mathbb{C}$ be an analytic function which extends continuously to the closure $\overline{\Omega}$. Then, for $z \in \Omega$,*

$$f(z) \,=\, \frac{1}{2\pi i} \int_\gamma \frac{f(w)}{z-w}\, dw\,.$$

The *Dunford-Schwartz calculus* gives meaning to such expression when replacing z by an operator T. More, the calculus applies to (unbounded) closed operators, as presented in [19]. And again, no symmetry is needed. As usual, we are limited to an outline of the concepts. Lorch's little book is highly recommended ([40]), as well as Bueno's for the matrix functional calculus ([6]).

First, we need to integrate continuous curves from, say, $\mathbb{C}$ to a Banach space Y (which in our case is $\mathcal{B}(X)$, but we emphasize the generality of the construction). In a nutshell, such integrals are limits of Riemann sums, which in turn only require the vector space structure of Y. Limits, of course, are defined by the norm of Y. These two sentences should convince the reader of a fundamental fact: the integral of a curve of matrices, for example, is just the integral of each matrix entry along this curve — recall the end of Section 2.1.

This naive idea should make the reader comfortable when integrating operators in infinite dimensions: if H is a Hilbert space and $T \in \mathcal{B}(H)$, one can think of the many expressions $\langle u, Tv\rangle$ and, in particular, the integral of a curve $t \in I \to T(t) \in \mathcal{B}(H)$ gives rise to an operator whose 'entry' associated to u and v is simply

$$\langle\, u,\, \Big(\, \int_I T(t)\, dt \,\Big)\, v \,\rangle = \int_I \langle\, u,\, T(t)\, v \,\rangle\, .$$

Diagonalizable matrices will lead the way. For $M = PDP^{-1}$,

$$M^2 \,=\, (P\, D\, P^{-1})\, (P\, D\, P^{-1}) = P\, D^2\, P^{-1},$$

and, more generally, for any polynomial p,

$$p(\, M\,) \,=\, P\, p(\, D\,)\, P^{-1},$$

where $p(D)$ is the diagonal matrix with entries $p(D_{ii})$. By the way, the constant term c in p has to be replaced by the matrix cI (why?). As in Section 3.1, this must be true for entire functions, which are uniform limits of polynomials on compact sets, and possibly more: the reader should have no difficulty in showing, for example, that

$$M^{-1} \,=\, P\, D^{-1}\, P^{-1}\, .$$

We replace z by M in the integrand of Cauchy's formula,

$$p(w)\, (M - wI)^{-1} \,=\, p(w)\, P\, (D - wI)^{-1}\, P^{-1} \,=\, P\, p(w)\, (D - wI)^{-1}\, P^{-1}\, .$$

Notice that $p(w)$ is a *scalar*, and thus it commutes with any matrix. Integration handles the constant matrices P and P^{-1} as *constants*, and they are taken out of the integral: this is the next exercise.

Exercise 24. Consider a curve of matrices $M(t) \in \mathcal{M}\,, t \in I$ and fixed matrices $A, B \in \mathcal{M}$. Show that

$$\int_I A\, M(t)\, B\, dt \;=\; A\,\Big(\int_I M(t)\, dt\Big)\, B\,.$$

Clearly, the result holds also for $T \in \mathcal{B}(X)$, X Banach.

Then, along a curve γ,

$$\int_\gamma p(w)\,(M - wI)^{-1}\, dz \;=\; P\,\Big(\int_\gamma p(w)\,(D - wI)^{-1}\, dt\Big)\, P\,.$$

The last integral is a diagonal matrix obtained by integrating diagonal entries (again, integrate entry by entry!). Thus for each position (i, i),

$$p(D_{ii}) \;=\; \frac{1}{2\pi i}\int_\gamma \frac{p(w)}{d_{ii} - w}\; dw$$

and this happens when γ is a simple positively oriented curve surrounding all possible numbers $d_{ii} \in \mathbb{C}$ — γ should surround $\sigma(M)$!

This should be the fundamental formula of the Dunford-Schwartz calculus ([19],[40]). Say $\Omega \subset \mathbb{C}$ and $f : \Omega \to \mathbb{C}$ satisfy the usual hypotheses of Cauchy's formula. For a Banach space B and $T \in \mathcal{B}$,

$$f(T) \;=\; \frac{1}{2\pi i}\int_\gamma f(w)\,(T - wI)^{-1}\; dw$$

for a simple, positively oriented curve surrounding $\sigma(T)$.

Let us see a first example: we prove the Cayley-Hamilton theorem for matrices — given a matrix M, the evaluation of its characteristic polynomial $p(\lambda) = \det(M - \lambda I)$ for $\lambda = M$ equals zero (why can't you just replace $\lambda = M$ in the formula for p?). By the calculus,

$$p(M) \;=\; \frac{1}{2\pi i}\int_\gamma p(w)\,(M - wI)^{-1}\; dw \;=\; \frac{1}{2\pi i}\int_\gamma p(w)\,\frac{(M - wI)^c}{p(w)}\; dw\,.$$

where we used Cramer's rule, the formula for the inverse of a matrix in terms of its *cofactor* matrix M^c — all that we have to know about it is that the cofactor is a polynomial in the entries of the matrix. We are thus integrating n^2 polynomials in w, one on each entry (the $p(w)$'s cancel), so by Cauchy's theorem again, the integral is *zero*.

Exercise 25. This is an open ended problem. The Cayley-Hamilton theorem is true for matrices with entries in arbitrary commutative rings with an identity. Have we proved it in this generality or not? After all, its statement, even in $\mathcal{M}(n, \mathbb{C})$, is of an arithmetic nature: the n^2 polynomials corresponding to the entries of $p(M - \lambda I)$ are all equal to zero. How much does the complex result says about the general case? I am reminded of some geometry problems in which auxiliary lines are convenient to the solution. In this case, we throw in a bunch of auxiliary *axioms*, those defining the complex numbers. The appropriate context for this question is close to universal algebra, possibly something in logic.

As another automatic application of the Dunford-Schwartz calculus, we show that for an operator $T \in \mathcal{B}$ whose spectrum is strictly contained in the open unit disk in $\mathbb{C}$, we must have $T^n \to 0$. Indeed,

$$T^n = \frac{1}{2\pi i} \int_\gamma w^n \, (M - wI)^{-1} \, dw$$

where we take for γ a circle centered at $0 \in \mathbb{C}$ with radius slightly less than 1 (recall that since $\sigma(T)$ is compact, the hypothesis allows for such a γ surrounding $\sigma(T)$). Now take absolute values and use that $w^n \to 0$, that simple (the norm of the denominator is clearly bounded away from zero).

The main result is the following theorem, whose proof is found in any presentation of the Dunford-Schwartz calculus. Let B be a complex Banach space, $\mathcal{B}$ be the algebra of linear continuous maps $T : B \to B$ endowed with the usual operator norm (Section 3.1). Let $U \subset \mathbb{C}$ be an open set containing $\sigma(T)$. Let $\gamma_k \subset U$ be a collection of curves enclosing open (topological) disks $D_k \subset U$ so that $\sigma(T) \subset \cup_k D_k$. Let $\mathcal{A}_\gamma$ be the algebra of continuous functions $f : \cup_k \overline{D}_k \to \mathbb{C}$ which are analytic in $\cup_k D_k$, with norm $||f|| = \sup_{z \in \cup_k \gamma_k} |f(z)|$.

Theorem 21. *There is a unique continuous algebra homomorphism* $\Phi : \mathcal{A}_\gamma \to \mathcal{B}$ *taking* $f(x) = 1$ *to* $\Phi(f) = I$ *and* $f(x) = x$ *to* $\Phi(f) = T$.

Say T is an $n \times n$ matrix with (possibly repeated) eigenvalues λ_k. Draw small (positively oriented) circles γ_k containing a unique distinct eigenvalue in each of the associated disks D_k. Let f_k be identically 1 in D_k and 0 in the other disks. Then

$$f_k(T)^2 = f_k(T)\,, \quad f_k(T)f_\ell(T) = 0 \text{ for } k \neq \ell\,, \quad \sum_k f_k(T) = I\,,$$

because the functions f_k satisfy these identities.

The projections $f_k(T)$ are the key ingredients in the Jordan decomposition of T. Their ranges are invariant subspaces and T on each subspace is of the form $\lambda_k I + N$, where N is nilpotent. Notice also that tr $f_k(T)$ is the multiplicity of the eigenvalue λ_k. The decomposition theorem follows from a normal form of a nilpotent operator.

And in infinite dimensions? A connected component $\sigma_k \subset \sigma(T)$ induces an invariant subspace $V_k =$ Ran $f_k(T)$ of T in the same fashion (γ does not intersect $\sigma(T)$). More, if tr $f_k(T) < \infty$, then the Jordan theorem applies to the restriction of T to V_k.

Exercise 26. Say λ_k is an isolated eigenvalue of $T_0 \in \mathcal{B}(X)$ of finite algebraic multiplicity, i.e., tr $f_k(T_0) = n < \infty$. For T near T_0, the functions

$$s_0(x) = f_k(x)\,,\ s_1(x) = x\, f_k(x)\,,\ s_2(x) = x^2\, f_k(x)\ \ldots$$

give rise to operators $s_j(T)$ of finite trace equal to the sum of the j-th power of the eigenvalues of T near λ_k (what does that mean?). In particular, tr $s_0(T) = n$. Clearly, these expressions are analytic in T. In a nutshell, even if the eigenvalues are hard to describe as continuous functions of T, polynomial symmetric functions, like the sum of the k-th powers, are as smooth as possible.

A standard proof of the spectral theorem for bounded self-adjoint operators proceeds by extending this algebra homomorphism. It turns out that for such T, we have $||f|| = ||T||$ and Φ extends naturally to continuous functions on $\sigma(T) \subset \mathbb{R}$. The subsequent step is the extension to (Borel) measurable functions ([45]).

For an entire function f and an $n \times n$ matrix $M = \lambda + N$, where N is nilpotent, it is clear that the computation of $f(T)$ through the power series gives rise to a polynomial, since $N^k = 0$ for some $k \in \mathbb{N}$. The upshot is that, in finite dimensions, say for $\sigma(T) \subset \mathbb{R}$, possible extensions of Φ are naturally limited to spaces C^k. In infinite dimensions, where the nilpotency indices become arbitrarily large, a space of analytic functions comes up naturally. Recall that a bound on the uniform norm of f in an open disk D surrounded by γ leads to uniform bounds of derivatives of f in closed disks in D, another gift from complex variable theory.

Exercise 27. Using Theorem 12, prove that

$$(S - \lambda I)^{-1} = \frac{1}{\lambda_1 - \lambda}\, v_1 \otimes v_1 + \frac{1}{\lambda_2 - \lambda}\, v_2 \otimes v_2 + \ldots + \frac{1}{\lambda_n - \lambda}\, v_n \otimes v_n\,,$$

where the v_k's are normalized eigenvectors of V. Thus the orthogonal projections associated to the invariant subspaces of S are the residues of the resolvent $R(\lambda) = (S - \lambda I)^{-1}$. For $g(\lambda) = \langle\, e_1\,, R(\lambda)\, e_1\,\rangle$,

$$g(\lambda) = \frac{c_1^2}{\lambda_1 - \lambda}\, v_1 \otimes v_1 + \frac{c_2^2}{\lambda_2 - \lambda}\, v_2 \otimes v_2 + \ldots + \frac{c_n^2}{\lambda_n - \lambda}\, v_n \otimes v_n\,,$$

where the c_k's are the first coordinates of the v_k's. In particular, if T is a Jacobi matrix, they are the norming constants of Section 2.3.1.

We finish with an example. Let $\mathbb{C}^+$ and $\mathbb{C}^-$ be the open right and left complex half planes. For $z \in \mathbb{C}^+ \cup \mathbb{C}^-$, the *sign function* $s(z)$ is

$$s(z) = \begin{cases} 1\,, & z \in \mathbb{C}^+\,, \\ -1\,, & z \in \mathbb{C}^-\,. \end{cases}$$

Given $T \in \mathcal{B}$ (in particular, a square matrix), one may compute $s(T)$ provided $\sigma(T)$ does not meet the imaginary axis. This function is used rather frequently by engineers and numerical analysts ([3]). Notice that $s(T)$ can be computed by the Dunford-Schwartz calculus, in principle: it is an *analytic* function in some open neighborhood of T. We present instead two computational alternatives, in order to convince the reader that functions of matrices allow for a lot of craftsmanship (there is much more in [24]).

Consider the two iterations

$$T_{k+1} = \frac{1}{2}\,(T_k + T_k^{-1})\,, \quad T_{k+1} = \frac{1}{2}\,T_k\,(\,3I - T_k^2\,)\,.$$

The first iteration, starting with $T_0 = T$, converges (quadratically!) to $s(T)$. This is easy to see for matrices: just check the effect of the iteration step on complex numbers in $\mathbb{C}^+ \cup \mathbb{C}^-$. For the second, if $\sigma(T) \subset (-\sqrt{3}\,, \sqrt{3})$, quadratic convergence is guaranteed provided $0 \notin \sigma(T)$ — again, check the effect of the iteration step on $\sigma(T)$. By the way, what about (infinite dimensional) operators?

Exercise 28. Use the sign function to count the number of eigenvalues of T in a quadrilateral of the complex plane. In a sense, this is an extension of Exercise 2.

5.2 Orthogonal polynomials

We simply can not honor one of the most interesting subjects in mathematics in such few pages — our intention is simply to show how a number of ideas in the previous sections combine. Alas, we will not provide examples and will almost trivialize the analytic context, but the reader will at least realize that we are at crossroads of different mathematical avenues. To go further, the only problem is the embarassment of riches — Szegö ([54]), Deift ([11]), Trefethen ([62]) are very different point of views, all of them very interesting.

Start with a real Hilbert space $H = L^2(I, d\mu)$, for some finite interval $I = [a, b] \subset \mathbb{R}$. We take μ to be a *probability measure*, i.e., a measure on the Borel sets of I with the property that

$$\int_I d\mu \,=\, 1\,.$$

Natural choices for μ are the Lebesgue measure, or a finite sum of deltas. Recall the inner product between real functions $u, v, \in H$,

$$\langle\, u\,,\, v\,\rangle = \int_I \, u(x)\; v(x)\; d\mu(x)\,.$$

Consider the multiplication operator

$$T : H \to H\,, \quad u(x) \;\mapsto\; x\; u(x)\,,$$

which is clearly bounded and symmetric, in the sense that

$$\langle u\,,\,T\,v\rangle = \langle T\,u\,,\,v\rangle\,,\quad \forall\,u\,,v\,\in H\,.$$

We now perform the construction described in Section 2.3. More precisely, consider the Krylov sequence of polynomials in H,

$$u_0 = 1\,,u_1 = T\,u_0 = x\,,u_2 = T^2\,u_0 = x^2\,,,\ldots$$

and their Gram-Schmidt orthonormalization (Lanczos) in H,

$$p_0 = u_0\,,p_1\,,p_2\,\ldots$$

The procedure works until the polynomial u_k becomes linear dependent from the previous ones, and then $p_0, p_1, \ldots, p_{n-1}$ span an n-dimensional invariant subspace $V \subset H$ of T. This is the case if μ is a sum of n deltas in I. If μ is Lebesgue measure, all vectors are independent. The open ended notation $\{p_0, p_1, \ldots\}$ indicates the largest set of such independent vectors, both for n finite or infinite.

The polynomials p_k have degree k — they are the *orthogonal polynomials* associated to the measure μ.

Exercise 29. Show that if V is infinite dimensional then polynomials are indeed a dense subset of H.

More, equipping V with the basis $\{p_0, p_1, \ldots\}$, the multiplication operator T is represented by a matrix J which is real, symmetric, tridiagonal with strictly positive entries $t_{i,i+1} = t_{i+1,i}$ — in a nutshell, J is a *Jacobi matrix*, as in Section 2.3. For convenience we index both rows and columns of J starting with zero and rename entries,

$$\begin{pmatrix} j_{00} & j_{01} & 0 & 0 & 0 & \ldots \\ j_{10} & j_{11} & j_{12} & 0 & 0 & \ldots \\ 0 & j_{21} & j_{22} & j_{23} & 0 & \ldots \\ 0 & 0 & j_{32} & j_{33} & j_{34} & \ldots \\ 0 & 0 & 0 & j_{43} & j_{44} & \ldots \\ & \ldots & & \ldots & & \ldots \end{pmatrix} = \begin{pmatrix} a_0 & b_0 & 0 & 0 & 0 & \ldots \\ b_0 & a_1 & b_1 & 0 & 0 & \ldots \\ 0 & b_1 & a_2 & b_2 & 0 & \ldots \\ 0 & 0 & b_2 & a_3 & b_3 & \ldots \\ 0 & 0 & 0 & b_3 & a_4 & \ldots \\ & \ldots & & \ldots & & \ldots \end{pmatrix}.$$

We have just obtained the *three terms recurrence* of the orthogonal polynomials $p_k(x)$: dropping the dependence on x,

$$T\,p_0 = x\;p_0 = a_0 p_0 \,+\, b_0 p_1\,,\quad T\,p_1 = x\;p_1 = b_0 p_0 \,+\, a_1 p_1 \,+\, b_1 p_2\,,$$

and, in general,

$$T\,p_k = x\;p_k = b_{k-1}p_{k-1} \;+\; a_k p_k \;+\; b_k p_{k+1} \quad k \geq 1\,.$$

Take inner products and use orthonormality to get the next result.

Proposition 7. *The entries a_k and b_k are given by*

$$a_0 = \langle\, p_0\,, T\,p_0\,\rangle\,, \quad a_k = \langle\, p_k\,, T\,p_k\,\rangle\,, \quad b_{k-1} = \langle\, p_{k-1}\,, T\,p_k\,\rangle \quad k \geq 1\,.$$

It is convenient to normalize the p_k's so as they become *monic*: set $c_k\tilde{p}_k = p_k$, so that the top coefficient of $\tilde{p}_k$ is equal to one. The recurrence relation for the $\tilde{p}_k$'s is (notice that $c_0 = 1$)

$$x\;\tilde{p}_0 = a_0\tilde{p}_0 \;+\; b_0\;c_1\;\tilde{p}_1\,,$$

$$x\,\tilde{p}_k = b_{k-1}\frac{c_{k-1}}{c_k}\;\tilde{p}_{k-1} \;+\; a_k\tilde{p}_k \;+\; b_k\frac{c_{k+1}}{c_k}\;\tilde{p}_{k+1} \quad k \geq 1\,.$$

Now, compare top coefficients to conclude that

$$b_k\frac{c_{k+1}}{c_k} = 1 \quad k \geq 0\,,$$

and since $p_0 = \tilde{p}_0 \equiv 1$,

$$x\;\tilde{p}_0 = a_0\tilde{p}_0 \;+\; \tilde{p}_1\,, \quad x\,\tilde{p}_k = b_{k-1}^2\;\tilde{p}_{k-1} \;+\; a_k\tilde{p}_k \;+\; \tilde{p}_{k+1} \quad k \geq 1$$

or

$$\tilde{p}_0 \equiv 1\,, \quad \tilde{p}_1 = (x - a_0)\,, \quad \tilde{p}_{k+1} = (x - \,a_k)\,\tilde{p}_k \;-\; b_{k-1}^2\;\tilde{p}_{k-1} \quad k \geq 1\,.$$

Let $D_k(\lambda)$ be the characteristic polynomial of the principal minor of dimension $k \times k$ of J. and set $D_0(\lambda) \equiv 1$. In particular,

$$D_0(\lambda) = 1\,, \quad D_1(\lambda) = a_0 - \lambda\,, \quad D_2(\lambda) = (a_0 - \lambda)(a_1 - \lambda) - b_0^2\,.$$

and, in general, expanding the determinant along the last row,

$$D_0 \equiv 1\,, \quad D_1 = a_0 - \lambda\,, \quad D_{k+1} = (a_k - \lambda)\,D_k \;-\; b_{k-1}^2\,D_{k-1}\,, \quad k \geq 1\,.$$

Comparing recursions, we get the following result.

Proposition 8. *The monic orthogonal polynomials are, up to sign, the determinants of the principal minors,*

$$\tilde{p}_{2k}(x) = D_{2k}(x) \qquad \text{and} \quad \tilde{p}_{2k+1}(x) = -D_{2k+1}(x)\,.$$

The next result combines a number of previous statements.

Theorem 22. *The polynomials $p_k(x)$ have simple roots. The roots of p_k and p_{k+1} interlace.*

Proof. The roots of p_k, being eigenvalues of a Jacobi matrix, are distinct, by Exercise 4. More, p_k and p_{k+1} are characteristic polynomials of two matrices satisfying the hypothesis of Exercise 23. □

A line of active research is the distribution of the zeros of orthogonal polynomials p_k for large values of the index k. They turn out to be surprisingly independent of the matrix μ under very mild hypothesis — said differently they display *universality properties* ([11]).

As an extra bonus, we compute the eigenvectors of the principal minors. Rewrite the three terms recurrence in matrix form:

$$\begin{pmatrix} x\,p_0(x) \\ x\,p_1(x) \\ x\,p_2(x) \\ x\,p_3(x) \\ x\,p_4(x) \\ \cdots \end{pmatrix} = \begin{pmatrix} a_0 & b_0 & 0 & 0 & 0 & \cdots \\ b_0 & a_1 & b_1 & 0 & 0 & \cdots \\ 0 & b_1 & a_2 & b_2 & 0 & \cdots \\ 0 & 0 & b_2 & a_3 & b_3 & \cdots \\ 0 & 0 & 0 & b_3 & a_4 & \cdots \\ & \cdots & & \cdots & & \cdots \end{pmatrix} \begin{pmatrix} p_0(x) \\ p_1(x) \\ p_2(x) \\ p_3(x) \\ p_4(x) \\ \cdots \end{pmatrix}.$$

Consider $n = 4$. Take for x a root r of the polynomial p_4:

$$r \begin{pmatrix} p_0(r) \\ p_1(r) \\ p_2(r) \\ p_3(r) \end{pmatrix} = \begin{pmatrix} a_0 & b_0 & 0 & 0 \\ b_0 & a_1 & b_1 & 0 \\ 0 & b_1 & a_2 & b_2 \\ 0 & 0 & b_2 & a_3 \end{pmatrix} \begin{pmatrix} p_0(r) \\ p_1(r) \\ p_2(r) \\ p_3(r) \end{pmatrix}.$$

Again, each root r of p_4 is an eigenvalue of the 4×4 principal minor of J associated to an explicit eigenvector.

5.3 A quadrature algorithm

Orthogonal polynomials have been presented as a special case of Lanczos's procedure. We now relate them to the Jacobi inverse variables from Section 2.3.1. From this association, we obtain a *quadrature algorithm*: consider the problem of approximating the integral

$$\int_I f \, d\mu \,,$$

for reasonable functions $f : \mathbb{R} \to \mathbb{R}$. Given $N = 2n - 1$, we obtain *interpolating points* λ_i and *weights* c_i, for $i = 0, \ldots, n-1$ for which

$$\int_I p \, d\mu = \sum_{i=0}^{n-1} c_i^2 \, p(\lambda_i) \,, \qquad (*)$$

for all polynomials of degree less than or equal to N.

The idea behind the algorithm is simple. Consider the (possibily infinite) orthogonal matrix J associated to the three term recursion of the orthogonal polynomials given by μ and the multiplication operator T. By Proposition 7, the entries a_k and b_k of J are given respectively by integrals of polynomials of degree $2k+1$ and $2k+2$. By Proposition 2, on the other hand, the common entries of J and its $n \times n$ principal minor J_{n-1} can be obtained in a different fashion, by making use of the *inverse variables* of J_{n-1}. We provide the details and set $n = 4$ to simplify notation.

In the previous section, orthogonal polynomials p_k associated to μ and T gave rise to eigenvalues and eigenvectors of principal minors of a(possibly infinite) Jacobi matrix J. Given the principal minor

$$J_3 = \begin{pmatrix} a_0 & b_0 & 0 & 0 \\ b_0 & a_1 & b_1 & 0 \\ 0 & b_1 & a_2 & b_2 \\ 0 & 0 & b_2 & a_3 \end{pmatrix} ,$$

we compute its (simple) spectrum $\lambda_1 > \ldots > \lambda_4$, which we arrange as $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_4)$. The entries of the eigenvectors are the values of the p_k's at points λ_i — such eigenvectors are not normal in $\mathbb{R}^4$:

we write $J_3 = W^T \Lambda W$ for some orthogonal matrix W^T, so that its columns are normalized eigenvectors of J_3,

$$W^T = \begin{pmatrix} p_0(\lambda_1) & p_0(\lambda_2) & p_0(\lambda_3) & p_0(\lambda_4) \\ p_1(\lambda_1) & p_1(\lambda_2) & p_1(\lambda_3) & p_1(\lambda_4) \\ p_2(\lambda_1) & p_2(\lambda_2) & p_2(\lambda_3) & p_2(\lambda_4) \\ p_3(\lambda_1) & p_3(\lambda_2) & p_3(\lambda_3) & p_3(\lambda_4) \end{pmatrix} \begin{pmatrix} c_0 & 0 & 0 & 0 \\ 0 & c_1 & 0 & 0 \\ 0 & 0 & c_2 & 0 \\ 0 & 0 & 0 & c_3 \end{pmatrix},$$

The numbers $c_k > 0$ are the first coordinates of the normalized eigenvalues of J_3: they are the *norming constants* of J_3, as defined in Section 2.3.1. Indeed, $p_0 \equiv 1$ and the first row of W is also an orthogonal vector, so that $\sum_k c_k^2 = 1$.

Theorem 23. *For $N = 2n-1$ and a measure μ supported in I, the quadrature equality $(*)$ above is true for the inverse variables (λ_i, c) of the principal minor J_{n-1} of the Jacobi operator J associated to the multiplication operator $T : L^2(I,\mu) \to L^2(I,\mu)$.*

Proof. For a polynomial of degree zero, the equation $(*)$ is true:

$$1 = \int_I d\mu = \sum_{i=0}^{n-1} c_i^2 = 1\,.$$

The entries a_k, $k = 0, \dots, n-1$ and b_k, $k = 0, \dots, n-2$ are common to J and J_{n-1}, so that, from Propositions 2 and 7,

$$a_0 = \langle\, p_0\,, T\,p_0\,\rangle = \langle\, v_0\,, \Lambda\, v_0\,\rangle\,, \quad a_k = \langle\, p_k\,, T\,p_k\,\rangle = \langle\, v_k\,, \Lambda\, v_k\,\rangle \quad k \geq 1\,,$$
$$b_{k-1} = \langle\, p_{k-1}\,, T\,p_k\,\rangle = \langle\, v_{k-1}\,, \Lambda\, v_k\,\rangle \quad k \geq 1\,.$$

We warn the reader: the vectors v_k, which are the rows of W^T, are *not* the eigenvalues of Λ. When such inner products are equal,

$$\int_I x\, p_k(x)\, p_\ell(x)\, d\mu = \langle\, p_k\,, T\,p_\ell\,\rangle = \langle\, v_k\,, \Lambda\, v_\ell\,\rangle = \sum_{i=1}^{n} c_i^2\, p_k(\lambda_i)\, p_\ell(\lambda_i)\,.$$

This is sufficient to prove equation $(*)$ for all polynomials p of degree less than or equal to N — start with $k = \ell = 0$ to show equality for polynomials of degree 1, then increase by one either index to extend to degree 2, and continue up to the equality associated to entry a_{n-1}: it yields the result for degree $N = 2n-1$. □

5.4 The spectral theorem — a sketch

Let us rephrase some facts from the previous sections. Take a probability measure and an operator

$$\mu = \sum_{k=1}^{n} c_k^2 \, \delta_{\lambda_k} \, , \quad T : L^2(\mathbb{R}, \mu) \to L^2(\mathbb{R}, \mu)$$

with a cyclic vector $p_0 \in L^2(\mathbb{R}, \mu)$. Denote by $(\mathbb{R}^n, \langle ., . \rangle)$ the usual inner product in $\mathbb{R}^n$. The correspondence between orthogonal bases

$$p_k \in L^2(\mathbb{R}, \mu) \mapsto w_k = c_k(v_k(\lambda_i)) \in \mathbb{R}^n$$

extends by linearity to an *isometry* $Q : L^2(\mathbb{R}, \mu) \to (\mathbb{R}^n, \langle ., . \rangle)$ which *diagonalizes* T: $T = Q^T \, \Lambda \, Q$. For the orthogonal matrix with rows given by the vectors w_k, $J = W^T \Lambda W$ is a Jacobi matrix.

We may have taken T to be the usual multiplication operator $Mf = x \, f(x)$, and we would have obtained an isometry conjugating J to M. Thus J admits two normal forms under orthogonal conjugation, Λ and M: the second form extends to infinite dimensions.

Starting with J, on the other hand, we obtain μ by computing poles and residues of $g(\lambda)$. Indeed, from Exercise 27, the function $g(\lambda) = \langle \, e_1 \, , (T - \lambda \, I)^{-1} \, e_1 \, \rangle$ is given by

$$g(\lambda) = \frac{c_1^2}{\lambda_1 - \lambda} \, v_1 \otimes v_1 + \frac{c_2^2}{\lambda_2 - \lambda} \, v_2 \otimes v_2 + \, \ldots \, + \frac{c_n^2}{\lambda_n - \lambda} \, v_n \otimes v_n \, ,$$

where the $c_k > 0$'s are the norming constants of T, i.e., the first coordinates of the normalized eigenvectors v_k's. In a more compact notation, preparing to jump to infinite dimensions,

$$g(\lambda) = \int_{\mathbb{R}} \frac{1}{x - \lambda} \, d \, \mu(x) \qquad (*),$$

where μ is the probability measure above. Notice a very special property of g: all poles are real and the residues are positive.

Now, let H be a separable Hilbert space. From Theorem 5, a general bounded self-adjoint operator $T : H \to H$ splits into a direct sum of Jacobi operators $J_\alpha : H_\alpha \to H_\alpha$, for appropriate subspaces

$H_\alpha \subset H$. To prove the spectral theorem for T, it suffices then to prove it for a Jacobi operator $J : H \to H$, where without loss we may take $H = \ell^2(\mathbb{N})$ (we consider the finite dimensional case as settled).

A proof of the spectral theorem along this lines is technically simpler: the issues related to spectral multiplicity are finessed. The theorem states: there is a measure μ supported in $I = [-||J||, ||J||]$ and an isometric bijection $Q^T : \ell^2(\mathbb{N}) \to L^2(I, \mu)$ for which

$$T = Q\, M\, Q^T\,,$$

where $M : L^2(I, \mu) \to L^2(I, \mu)$, $(Mf)(x) = xf(x)$ is multiplication by x. The proof sometimes is presented for this statement, as in [11], or in some variation, as in [40]. Both texts are beautiful.

Suppose we know already, from standard estimates, that the spectrum of J is real. The key technical object is a representation theorem of Herglotz, which ensures that the function

$$g(\lambda) = \langle\, e_1\,, (J - \lambda\, I)^{-1}\, e_1\,\rangle$$

is given by a probability measure μ supported in I, as in $(*)$. Indeed this is true for analytic functions which take the open upper half-plane to itself: this is the appropriate phrasing of the special properties of g outlined in the finite dimensional case. Asymptotic properties of g (it goes to zero at infinity) then yield the formula.

Once μ is available, everything follows as in the finite dimensional case: the conjugation Q, the multiplication operator T... Notice by the way the following alternative to retrieve J from μ. Expand

$$g(\lambda) = \frac{-1}{\lambda}\langle\, e_1\,, (I - \frac{J}{\lambda})^{-1}\, e_1\,\rangle = \frac{-1}{\lambda}\langle\, e_1\,, \big(I + \frac{J}{\lambda} + (\frac{J}{\lambda})^2 + \dots\big)\, e_1\,\rangle\,.$$

Thus, from $g(\lambda)$ we obtain the expressions $\langle\, e_1\,, J^n\, e_1\,\rangle$, from which the entries a_k and b_k are recursively computed.

As is well know, the spectral theorem for unbounded self-adjoint operators follows from the bounded case, using a trick by Von Neumann from the functional calculus. An alternative route closer to the techniques above leads to a question of independent interest — given a measure μ in $\mathbb{R}$, when are polynomials dense in $L^2(\mathbb{R}, \mu)$? The interested reader should consult [11] again.

Bibliography

[1] R.Abraham and J.Marsden, *Foundations of Mechanics*, second edition, Benjamin/Cummings, Reading, 1978.

[2] M. Atiyah, *Convexity and commuting Hamiltonians*, Bull. London Math. Soc., 14 1-15, 1982.

[3] G. Beylkin and M.J. Mohlenkamp, *Numerical operator calculus in higher dimensions*, PNAS, 99(16), 10246-10251, 2002.

[4] G. Blind and R. Blind, *The semiregular polytopes*, Comment. Math. Helvetici, 66, 150-154, 1991.

[5] A. Bloch, H. Flaschka and T. Ratiu, *A convexity theorem for isospectral manifolds of Jacobi matrices in a compact Lie algebra*, *Duke Math. J.* 61, 41-65, 1990.

[6] H. Bueno, *Funções de Matrizes*, I Bienal SBM, http://www.mat.ufmg.br/ hamilton/Minicursos/FunMatr.pdf

[7] F. Chung, *Spectral Graph Theory*, CBMS Regional Conf. Ser. Math. 92, Amer. Math. Soc., 1997.

[8] F. Chung and S.Sternberg, *Laplacian and vibrational spectra for homogeneous graphs*, J. Graph Th. 16, 605-627, 1992.

[9] H.S.M. Coxeter, *Regular polytopes*, Dover, New York, 1973.

[10] D. Cvetkovic, M. Doob, M. and H. Sachs, *Spectra of Graphs: Theory and Applications*, Wiley, New York, 1998.

[11] P. Deift, *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*, Courant Lecture Notes 3, New York, 2000.

[12] P. Deift, *Applications of a commutation formula*, Duke Math. J. 45, 267-310 (1978).

[13] P. Deift, T. Nanda and C. Tomei, *Differential equations for the symmetric eigenvalue problem*, *SIAM J. Num. Anal.* 20, 1-22, 1983.

[14] P. Deift, S. Rivera, C. Tomei and D. Watkins, *A monotonicity property for Toda-type flows*, *SIAM J. Matrix Anal. Appl.*, 12, 463-468, 1991.

[15] P. Deift and E. Trubowitz, *Inverse Scattering on the Line*, Comm. Pure Appl. Math 32, 121-251, 1979.

[16] J. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[17] J. Duistermaat, *The momentum maps*, Topics in Differential Geometry, vols. I e II, Colloq. Math. Soc. Janos Bolyai 46, 347-392, 1988.

[18] J. Duistermaat, J. C. Kolk, *Lie groups*, Universitext, Springer, New York, 2000.

[19] N. Dunford and J. Schwartz, *Linear Operators*, Wiley, Englewood Cliffs, 1988.

[20] L. Eldén, *Matrix methods in data mining and pattern recognition*, Fundamentals of Algorithms, SIAM, Philadelphia, 2007.

[21] H. Flaschka, *The Toda lattice, I*, *Phys. Rev. B* 9, 1924-1925, 1974.

[22] D. Fried, *The cohomology of an isospectral flow*, Proc. Amer. Math. Soc., 98, 363-368, 1986.

[23] V. Guillemin and S. Sternberg, *Convexity properties of the moment mapping, I*, Invent. Math. 67, 491-513, 1982.

[24] N. Higham, *Functions of matrices, theory and computation*, SIAM, Philadelphia, 2008.

[25] A. Horn, *Doubly stochastic matrices and the diagonal of a rotation matrix*, Amer. J. Math. 76, 620-630, 1954.

[26] A. Horn, *Eigenvalues of sums of Hermitian matrices*, Pacific J. Math. 12 225-241, 1962.

[27] R. Horn and C. Johnson, *Topics in matrix analysis*, CUP, Cambridge, 1999.

[28] T. Kato, *Perturbation theory for linear operators*, Classics in Mathematics, Springer, New York, 2013,

[29] T.G. Kolda and B.W. Bader, *Tensor decompositions and applications*, SIAM Review 51 (3), 455-500, 2008.

[30] B. Kostant, *On convexity, the Weyl group and the Iwasawa decomposition*, Ann. Sci. École Norm. Sup. 6, 413-455, 1973.

[31] A. Knutson and T. Tao, *Honeycombs and Sums of Hermitian Matrices*, Notices Amer. Math. Soc. 48, 175-186, 2001.

[32] S. Lang, *Algebra*, Graduate Texts in Mathematics 211, Springer, New York, 2002.

[33] P. Lax, *Linear algebra and its applications*, Wiley, Englewood Cliffs, 2007.

[34] P. Lax, *Functional Analysis*, Wiley, Englewood Cliffs, 2002.

[35] R. Leite, T. Richa and C. Tomei, *Geometric proofs of some theorems of Schur-Horn type*, Lin. Alg. Appl. 286, 149-173,1999.

[36] R.S. Leite, N.C. Saldanha and C. Tomei, An atlas for tridiagonal isospectral manifolds, *Lin. Alg. Appl.* 429, 387-402, 2008.

[37] R.S. Leite, N.C. Saldanha and C. Tomei, *The Asymptotics of Wilkinson's shift: Loss of Cubic Convergence*, *FoCM* 10 (1), 15-36, 2010

[38] R.S. Leite, N.C. Saldanha and C.Tomei, *Dynamics of the symmetric eigenvalue problem with shift strategies*, Int Math Res Notices 395, 63-77, 2012.

[39] P.H. Leslie, *The use of matrices in certain population mathematics*, Biometrika 33, 183-212, 1945.

[40] E. Lorch, *Spectral Theory*, University Texts in the Mathematical Sciences, Oxford University Press, New York, 1962.

[41] A.W. Marshall, I. Olkin and B. Arnold, *Inequalities: theory of majorization and its applications*, Springer Series in Statistics, Springer, New York, 2011.

[42] J. Moser, *Finitely many mass points on the line under the influence of an exponential potential*, In: *Dynamic systems theory and applications*, (ed. J. Moser) 467-497, New York, 1975.

[43] B. Noble and J. W. Daniel, *Applied Linear Algebra*, Prentice-Hall, Englewood Cliffs, 1987.

[44] B. Parlett, *The Symmetric Eigenvalue Problem*, Classics in Applied Mathematics 20, SIAM, 1997.

[45] M. Reed and B. Simon, *Functional Analysis*, Methods of Modern Mathematical Physics, volume I, Academic Press, New York, 1972.

[46] D. Rodriguez, *Tensor product algebra as a tool for VLSI implementation of the discrete Fourier transform*, Acoustics, Speech, and Signal Processing, 1991. ICASSP-91, 1991, IEEE

[47] N. Saldanha and C. Tomei, *Spectra of regular polytopes,* Disc. Comp. Geom. 7, 403-414, 1992.

[48] N. Saldanha and C. Tomei, *Spectra of semi-regular polytopes*, Bull. Bras. Math. Soc. 29, 25-51, 1998.

[49] J.P. Serre, *Linear Representations of Finite Groups,* GTM 42, Springer, New York, 1977.

[50] B. Simon, *Representations of Finite and Compact Groups*, graduate Studies in Mathematics 10, AMS, Providence, 1996.

[51] I. Schur, *Über eine Klasse von Mittelbildungen mit Anwendungen auf die Determinantentheorie*, Sitzungsber. Berl. Math. Ges. 22, 9-20, 1923.

[52] F.B. Sing, *Some results on matrices with prescribed diagonal elements and singular values*, Canad. Math. Bull. 19, 89-92, 1976.

[53] M. Spivak, *Calculus on manifolds*, Addison-Wesley, 1971.

[54] G. Szegö, *Orthogonal polynomials*, Colloquim Publications 23, AMS, Providence, 1939.

[55] T. Tao, *Topics in Random Matrix theory*, Grad. Studies Math. 132, Amer. Math. Soc., Providence, 2012.

[56] R.C. Thompson, *Singular values, diagonal elements, and convexity*, SIAM J. Appl. Math. 32, 39-63, 1977.

[57] R.C. Thompson, *High, Low, and Quantitative Roads in Linear Algebra*, Lin. Alg. Appl. 162, 23-64, 1992.

[58] M.Toda, *Wave propagation in anharmonic lattices*, *J. Phys. Soc. Japan*, 501-506, 1967.

[59] C. Tomei, *The Topology of Manifolds of Isospectral Tridiagonal Matrices*, Duke Math. J. 51, 981-996, 1984.

[60] C. Tomei, *The Toda lattice, old and new*, J. Geom. Mech. 5, 511-530, 2013.

[61] L.N. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[62] L.N. Trefethen, *Approximation Theory and Approximation Practice*, SIAM, Philadelphia, 2013.

[63] L.N. Trefethen and M. Embree, *Spectra and Pseudospectra: the Behavior of nonnormal Matrices and Operators*, Princeton, Princeton, 2005.

[64] J. Wilkinson, *The algebraic eigenvalue problem*, Oxford University Press, 1965.