Linear Algebra and Analysis

Masterclasses

By

Rajendra Bhatia

Published by
Indian Academy of Sciences

Foreword

The Masterclass series of eBooks bring together pedagogical articles on single broad topics taken from *Resonance*, the Journal of Science Education that has been published monthly by the Indian Academy of Sciences since January 1996. Primarily directed at students and teachers at the undergraduate level, the journal has brought out a wide spectrum of articles in a range of scientific disciplines. Articles in the journal are written in a style that makes them accessible to readers from diverse backgrounds, and in addition, they provide a useful source of instruction that is not always available in textbooks.

The second book in the series, *Linear Algebra and Analysis Masterclasses*, is by Prof. Rajendra Bhatia. A celebrated mathematician, Prof. Bhatia's career has largely been at the Indian Statistical Institute, New Delhi where he has been for over three decades and is currently a Distinguished Scientist. He has also contributed pedagogical articles regularly to *Resonance*, and these comprise the bulk of the present book. Only two of the ten articles in the book have not appeared earlier in *Resonance*.

Professor Bhatia's work has made significant inroads in a variety of areas, including mathematical physics, computer science, numerical analysis, and statistics. The book, which will be available in digital format and will be housed as always on the Academy website, will be valuable to both students and experts as a useful handbook on *Linear Algebra and Analysis*.

T. N. Guru Row
Editor of Publications
Indian Academy of Sciences
August 2016

About the Author

Rajendra Bhatia has spent the major part of his professional life at the Indian Statistical Institute, Delhi, where he now holds the position of Distinguished Scientist. He was earlier at the Tata Institute and at the University of Bombay, and has held visiting positions in several universities, starting with the University of California, Berkeley in 1979, the latest being Shanghai University in 2015.

Bhatia is the Founding Editor of the book series "*Texts and Readings in Mathematics*" or TRIM, which has published over 70 books, as well as the series "*Culture and History of Mathematics*". He is a Fellow of the Indian National Science Academy, the Indian Academy of Sciences and TWAS, The World Academy of Sciences. He is a recipient of the Indian National Science Academy Medal for Young Scientists, the Shanti Swarup Bhatnagar Award, the Hans Schneider Prize in Linear Algebra, and the J. C. Bose National Fellowship.

His research work in Matrix Analysis is cited equally often by mathematicians, statisticians, physicists and computer scientists. This is largely due to the fact that his work on matrix analysis (perturbation of spectra, matrix inequalities and equations, positivity, means) combines ideas and methods from Fourier analysis and differential geometry. It has stimulated much research and has been used in mathematical physics, computer science, numerical analysis, and statistics.

In 2005 Bhatia gave a definition of "geometric mean" of more than two positive definite matrices (a definition that has since become standard) and demonstrated that it had the right properties demanded by various subjects (operator theory, elasticity, diffusion tensor imaging etc). This has led to interesting theorems spanning analysis and differential geometry and has found applications in diverse areas such as image processing, smoothing of radar data, machine learning, and brain-computer interface.

Bhatia is a master of exposition. He is the author of several books, some of which are now the definitive treatises on their subjects. It is very timely that a collection of his shorter essays and didactic articles should now be made available in a convenient format.

T R Ramadas
Chennai Mathematical Institute
Chennai

Contents

The Work of the Fields Medallists: 1998 – William T Gowers*

Rajendra Bhatia
Indian Statistical Institute, New Delhi 110 016, India.

The subject Functional Analysis started around the beginning of this century, inspired by a desire to have a unified framework in which the two notions of *continuity* and *linearity* that arise in diverse contexts could be discussed abstractly. The basic objects of study in this subject are Banach spaces and the spaces of bounded (continuous) linear operators on them; the space $C[a, b]$ of continuous functions on an interval $[a, b]$ with the supremum norm, the L^p spaces arising in the theory of integration, the sequence spaces l_p, the Sobolev spaces arising in differential equations, are some of the well-known examples of Banach spaces. Thus there are many concrete examples of the spaces, enabling application of the theory to a variety of problems.

It is generally agreed that finite-dimensional spaces are well understood and thus the main interest lies in infinite-dimensional spaces. A Banach space is *separable* if it has a countable dense subset in it. From now on we will talk only of separable Banach spaces; the non-separable Banach spaces are too unwieldy.

The simplest examples of infinite-dimensional Banach spaces are the sequence spaces $l_p, 1 \leq p < \infty$ consisting of sequences $x = (x_1, x_2, \ldots)$ for which the sum $\sum_{i=1}^{\infty} |x_i|^p$ is finite; the pth root of the latter is taken as the norm of x. These spaces are separable. The space of all bounded sequences, equipped with the supremum norm, is called l_∞. It is not separable, but contains in it the space c_0 consisting of all convergent sequences, which is separable. The following was an open question for a long time: does every Banach space contain in it a subspace that is isomorphic to either c_0 or some $l_p, 1 \leq p < \infty$? It was answered in the negative by *B*. Tsirelson in 1974.

It may be recalled that in the theory of finite-dimensional vector spaces, bases play an important role. A *Schauder basis* (or a *topological basis)* for a Banach space X is a sequence $\{e_n\}$ in X such that every vector in X has a unique expansion where the infinite series is understood to converge in norm. Unlike in the finite-dimensional case, in general this notion depends on the order in which $\{e_n\}$ is enumerated. We say a Schauder basis $\{e_n\}$ is an *unconditional* basis if $\{e_{p(n)}\}$ is a Schauder basis for every permutation p of natural numbers.

It is easy to see that if a Banach space has a Schauder basis, then it is separable. There was a famous problem as to whether every separable Banach space has a Schauder basis. P Enflo showed in 1973 that the answer is no. It had been shown quite early by S Mazur that every (infinite-dimensional) Banach space has an (infinite-dimensional) subspace with a Schauder basis. (The spaces $l_p, 1 \leq p < \infty$ and c_0 do have Schauder bases.)

*Reproduced from *Resonance*, Vol. 4, No. 4, pp. 85–87, April 1999. (Research News)

One of the major results proved by W T Gowers, and independently by B Maurey, in 1991 is that there exist Banach spaces that do not have any infinite-dimensional subspace with an unconditional basis.

In many contexts the interest lies more in operators on a Banach space than the space itself. Many of the everyday examples of Banach spaces do have lots of interesting operators defined on them. But it is not clear whether every Banach space has nontrivial operators acting on it. If the Banach space has a Schauder basis one can construct examples of operators by defining their action on the basis vectors. Shift operators that act by shifting the basis vectors to the left or the right have a very rich structure. Another interesting family of operators is the projections. In a Hilbert space every subspace has an orthogonal complement. So, there are lots of orthogonal decompositions and lots of projections that have infinite rank and corank. In an arbitrary Banach space it is not necessary that any infinite-dimensional subspace must have a complementary subspace. Thus one is not able to construct nontrivial projections in an obvious way.

The construction of Gowers and Maurey was later modified to show that there exists a Banach space X in which every continuous projection has finite rank or corank, and further every subspace of X has the same property. This is equivalent to saying that *no* subspace Y of X can be written as a direct sum $W \oplus Z$ of two infinite-dimensional subspaces. A space with this property is called *hereditarily indecomposable*. In 1993 Gowers and Maurey showed that such a space cannot be isomorphic to *any* of its proper subspaces. This is in striking contrast to the fact that an infinite-dimensional Hilbert space is isomorphic to *each* of its infinite-dimensional subspaces (all of them are isomorphic to l_2). A Banach space with this latter property is called *homogeneous*.

In 1996 Gowers proved a dichotomy theorem showing that every Banach space X contains either a subspace with an unconditional basis or a hereditarily indecomposable subspace. A corollary of this is that every homogeneous space must have an unconditional basis. Combined with another recent result of R Komorowsky and N Tomczak-Jaegermann this leads to another remarkable result: every homogeneous space is isomorphic to l_2.

Another natural question to which Gowers has found a surprising answer is the Schroeder-Bernstein problem for Banach spaces. If X and Y are two Banach spaces, and each is isomorphic to a subspace of the other, then must they be isomorphic? The answer to this question has long been known to be no. A stronger condition on X and Y would be that each is a *complemented* subspace of the other. (A subspace is complemented if there is a continuous projection onto it; we noted earlier that not every subspace has this property.) Gowers has shown that even under this condition, X and Y need not be isomorphic. Furthermore, he showed this by constructing a space Z that is isomorphic to $Z \oplus Z \oplus Z$ but not to $Z \oplus Z$.

All these arcane constructions are not easy to describe. In fact, the norms for these Banach spaces are not given by any explicit formula, they are defined by indirect inductive procedures. All this suggests a potential new development in Functional Analysis. The concept of a Banach space has encompassed many interesting concrete spaces mentioned at the beginning. However, it might be *too* general since it also admits such strange objects. It is being wondered now

whether there is a new theory of spaces whose norms are easy to describe. These spaces may have a richer operator theory that general Banach spaces are unable to carry.

In his work Gowers has used techniques from many areas, specially from combinatorics whose methods and concerns are generally far away from those of Functional Analysis. For example, one of his proofs uses the idea of two-person games involving sequences of vectors and Ramsey Theory. Not just that, he has also made several important contributions to combinatorial analysis. We end this summary with an example of such a contribution.

A famous theorem of E. Szemeredi, which solved an old problem of P Erdos and P Turan, states the following. For every natural number k and for $0 < \delta < 1$, there exists a natural number $N(\delta, k)$ such that if $n > N(\delta, k)$, then every subset of $\{1, 2, \ldots, n\}$ of size δN contains an arithmetic progression of length k. Gowers has found a new proof of this theorem based on Fourier analysis. This proof gives additional important information that the original proof, and some others that followed, could not. It leads to interesting bounds for $N(\delta, k)$ in terms of k and δ.

David Hilbert*

Rajendra Bhatia
Indian Statistical Institute, New Delhi 110 016, India.

It will be difficult to find a twentieth century mathematician working in an area that was not touched by David Hilbert. There is Hilbert space, Hilbert scheme, Hilbert polynomial, Hilbert matrix, Hilbert inequality, Hilbert invariant integral, Hilbert norm-residue symbol, Hilbert transform, Hilbert class-field, Hilbert basis theorem, Hilbert irreducibility theorem, Hilbert nullstellensatz.

Hilbert also changed the way mathematicians think about their subject. The axiomatic spirit in which modern mathematics is done owes much to him.

In an address to the International Congress of Mathematicians in 1900, he proposed a list of 23 problems that, in his opinion, should be the principal targets for mathematicians in this century. This famous list, now called Hilbert's Problems, has directed the work of several leading mathematicians.

David Hilbert was born on January 23, 1862 near Königsberg, then the capital of East Prussia, now renamed as Kaliningrad in Russia. The seven bridges on the river Pregel flowing through this town are associated with one of the most famous problems in mathematics. The solution of this problem by Euler became the first theorem in graph theory. The famous philosopher Kant lived here and the great mathematician Jacobi taught at the university of this town.

David's parents were Maria and Otto Hilbert. David's father was a judge. Hilbert's teachers at Königsberg, then a leading university of Germany included H Weber and A Hurwitz. Among his fellow students was H Minkowski. Hilbert, Hurwitz and Minkowski began here a life-long friendship that nourished them in their scientific and personal lives.

Hilbert's research began with the *theory of invariants*, a subject with roots in geometry and number theory. The theory had begun with the work of A Cayley and was developed further by J J Sylvester, R Clebsch and P Gordan. Hilbert changed the face of the subject in two ways. First he broadened the scope of the theory by introducing the notion of invariants for general groups. Second, he proved the existence of a finite basis for the ring of invariants, not by explicit computations as others before had done, but by a general *existential* argument. Such an argument, now so commonly used, proceeds by showing that an object must exist, because if it did not, a contradiction would follow.

Gordan, then considered the 'King of Invariants', on seeing Hilbert's proof remarked "This is not Mathematics. It is Theology". It is somewhat ironic that Hilbert got a crucial idea for his theorem on invariants by studying the work of L Kronecker who was a staunch opponent of such non-constructive proofs.

*Reproduced from *Resonance*, Vol. 4, No. 8, pp. 3–5, August 1999. (Article-in-a-Box)

To meet such criticisms, and to show the way out of certain paradoxes that had arisen in the theory of sets, Hilbert advanced the doctrine of *formalism* as opposed to *logicism* of B Russell and *intuitionism* of L E J Brouwer. At issue was the very nature of mathematical proof. Today, most mathematicians have accepted the formalist viewpoint.

Hilbert's work on invariants became the cornerstone of modern algebra. He went on to do equally fundamental work in geometry, number theory, analysis, differential and integral equations, calculus of variations, and mathematical physics. In his *Zahlbericht* (1897), a monumental report written at the invitation of the German Mathematical Society, he presented a unification of the known results on algebraic number fields as "an edifice of rare beauty and harmony". In his book *Grundlagen der Geometrie* (1899) he laid down a list of complete axioms of Euclidean geometry. He examined the logical relations between these axioms and showed their *independence* by constructing *models* in which all but one of the axioms are satisfied. He went on to show that this axiomatic system is as *consistent* as the theory of real numbers.

Hilbert spent most of his professional life at Göttingen, for a long time regarded as the mathematics capital of the world. Among his predecessors here had been C F Gauss and B Riemann; among his contemporaries were F Klein, E Landau, H Weyl and Emmy Noether. The physicist Max Born began his scientific career as Hilbert's assistant. Born's first two assistants, when he later established an institute for physics at Göttingen, were W Pauli and W Heisenberg.

Hilbert's work on integral equations and eigenvalue problems was inspired by the important papers of E Fredholm. Just as he had done in other subjects, Hilbert laid emphasis on the fundamental principles of the subject. This laid the foundation for the theory of Hilbert spaces developed by J von Neumann and others. The classic book *Methods of Mathematical Physics* by Courant and Hilbert was also an outcome of this work. Here, several problems of differential and integral equations were formulated as problems in infinite-dimensional linear algebra. In this book physicists found many mathematical tools they needed to develop the new quantum mechanics. It is most remarkable that the word *spectrum* Hilbert had used to describe some quantities associated with linear operators later turned out to be exactly the spectrum associated with atomic emissions.

The first approach to quantum mechanics was the *matrix mechanics* of Heisenberg, developed further by Born and Jordan. When they approached Hilbert for advice, he replied that he did not know much about matrices except that he had thought of them in connection with some differential equations and perhaps they should look for such equations associated with their matrices. His suggestion was ignored as being a shot in the dark. However, soon E Schrödinger proposed an alternative approach to quantum mechanics called *wave mechanics*. This used differential equations and was very different from *matrix* mechanics. Soon however, the two theories were shown to be equivalent, just as Hilbert had anticipated.

Of course, Hilbert could also be wrong in his judgements. In a lecture in 1919, he gave some examples of problems in number theory that are simple to state but extremely hard to solve. He mentioned the Riemann hypothesis, Fermat's Last Theorem, and the conjecture

that $2^{\sqrt{2}}$ is a transcendental number (Hilbert's seventh problem in the famous list). He then added that he might see the proof of the Riemann hypothesis in his life time, that the youngest members of the audience might live to see Fermat's Last Theorem proved, but no one present in the hall would live to see a proof of transcendence of $2^{\sqrt{2}}$. Things did not go the way Hilbert had predicted. The transcendence of $2^{\sqrt{2}}$ was established by A Gel'fond in 1934 when Hilbert was alive; Fermat's Last Theorem was proved by Andrew Wiles in 1994 when perhaps all the members of Hilbert's audience in 1919 were dead; the Riemann hypothesis is yet to be proved. Incidentally, among Hilbert's first works in number theory is a new and simple proof of the transcendence of the number e (first established by Hermite) and of the number π (first established by Lindemann, Hilbert's teacher at Königsberg).

As a person, Hilbert was fair, firm and bold. In 1914, when the German government publicised a declaration in defence of its war actions signed by its most famous scientists, Hilbert's name was missing. The declaration included several statements beginning "It is not true that..." Hilbert refused to sign it on the ground that he could not ascertain whether these statements were true. In 1917 he wrote and published a tribute to the French mathematician G Darboux on his death. This tribute to an 'enemy' outraged some students who demonstrated at Hilbert's home demanding repudiation from him and the destruction of all copies of the publication. Hilbert refused and then insisted on getting an apology from the university. When the conservative professors of the university opposed the appointment of Emmy Noether, a mathematician of the highest calibre, because she was a woman, Hilbert retorted that the University Senate was not a bathhouse where women could not enter. He was outraged by, and was incredulous at, the dismissal of his Jewish colleagues by the Nazis.

He lived to see the tragic destruction of his great centre of mathematics amidst the bigger tragedy of his country. He died on February 14, 1943. The times were such that only about ten persons attended his funeral service, and the news of his death reached the outside world several months later.

Algebraic Geometry Solves an Old Matrix Problem*

Rajendra Bhatia
Indian Statistical Institute, New Delhi 110 016, India.

Let A, B be $n \times n$ Hermitian matrices, and let $C = A + B$. Let $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$, $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$, and $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_n$ be the eigenvalues of $A, B,$ and C, respectively. Mathematicians, physicists, and numerical analysts have long been interested in knowing all possible relations between the n-tuples $\{\alpha_j\}, \{\beta_j\}$ and $\{\gamma_j\}$.

Since tr C = tr A + tr B, where tr stands for the trace of a matrix, we have

$$\sum_{i=1}^{n} \gamma_i = \sum_{i=1}^{n} (\alpha_i + \beta_i). \tag{1}$$

H Weyl (1912) was the first to discover several non-trivial relations between these numbers; these are the inequalities

$$\gamma_{i+j-1} \leq \alpha_i + \beta_j \quad \text{for} \quad i + j - 1 \leq n. \tag{2}$$

(See [1, Chapter 3]) for a proof and discussion of this and some of the other results described below.)

When $n = 2$, this yields three inequalities

$$\gamma_1 \leq \alpha_1 + \beta_1, \quad \gamma_2 \leq \alpha_1 + \beta_2, \quad \gamma_2 \leq \alpha_2 + \beta_1. \tag{3}$$

It turns out that, together with the equality (1), these three inequalities are sufficient to characterise the possible eigenvalues of $A, B,$ and C; i.e., if three pairs of real numbers $\{\alpha_1, \alpha_2\}$, $\{\beta_1, \beta_2\}$, $\{\gamma_1, \gamma_2\}$, each ordered decreasingly ($\alpha_1 \geq \alpha_2$, etc.), satisfy these relations, then there exist 2×2 Hermitian matrices A and B such that these pairs are the eigenvalues of A, B and $A+B$.

When $n \geq 3$, more relations exist. The first one due to Ky Fan (1949) says

$$\sum_{j=1}^{k} \gamma_j \leq \sum_{j=1}^{k} \alpha_j + \sum_{j=1}^{k} \beta_j, \quad \text{for } 1 \leq k \leq n. \tag{4}$$

When $k = n$, the two sides of (4) are equal; that is just the equality (1). A substantial generalisation of this was obtained by V B Lidskii (1950). For brevity, let $[1, n]$ denote the set $\{1, 2, \ldots, n\}$. Lidskii's theorem says that for every subset $I \subset [1, n]$ with cardinality $|I| = k$, we have

$$\sum_{i \in I} \gamma_i \leq \sum_{i \in I} \alpha_i + \sum_{j \leq k} \beta_j. \tag{5}$$

*Reproduced from *Resonance*, Vol. 4, No. 12, pp. 101–105, December 1999. (Research News).

Note that these inequalities include (4) as a special case – choose $I = [1, k]$.

Lidskii's theorem has an interesting history. It was first proved by F Berezin and I M Gel'fand in connection with their work on Lie groups. On their suggestion Lidskii provided an elementary proof. Others had difficulty following this proof. It was H W Wielandt (1955) who supplied a proof that was understood by others. Now several proofs of this theorem are known; see [1].

When $n = 3$, we get six relations from Weyl's inequalities :

$$\begin{aligned} &\gamma_1 \le \alpha_1 + \beta_1, \qquad \gamma_2 \le \alpha_1 + \beta_2, \qquad \gamma_2 \le \alpha_2 + \beta_1, \\ &\gamma_3 \le \alpha_1 + \beta_3, \qquad \gamma_3 \le \alpha_3 + \beta_1, \qquad \gamma_3 \le \alpha_2 + \beta_2. \end{aligned} \tag{6}$$

Five more follow from the inequalities (5):

$$\begin{aligned} \gamma_1 + \gamma_2 &\le \alpha_1 + \alpha_2 + \beta_1 + \beta_2, \\ \gamma_1 + \gamma_3 &\le \alpha_1 + \alpha_3 + \beta_1 + \beta_2, \\ \gamma_2 + \gamma_3 &\le \alpha_2 + \alpha_3 + \beta_1 + \beta_2, \\ \gamma_1 + \gamma_3 &\le \alpha_1 + \alpha_2 + \beta_1 + \beta_3, \\ \gamma_2 + \gamma_3 &\le \alpha_1 + \alpha_2 + \beta_2 + \beta_3. \end{aligned} \tag{7}$$

(use the symmetry in A, B). It turns out that one more relation

$$\gamma_2 + \gamma_3 \le \alpha_1 + \alpha_3 + \beta_1 + \beta_3, \tag{8}$$

is valid. Further, the relations (1), (6), (7) and (8) are sufficient to characterise the possible eignevalues of A, B and C.

The Lidskii–Wielandt theorem aroused much interest, and several more inequalities were discovered. They all have the form

$$\sum_{k\in K} \gamma_k \le \sum_{i\in I} \alpha_i + \sum_{j\in J} \beta_j, \tag{9}$$

where I, J, K are certain subsets of $[1, n]$ all having the same cardinality. Note that the inequalities (2), (4) and (5) all have this form.

This leads to the following questions. What are all the triples (I, J, K) of subsets of $[1, n]$ for which the inequalities (9) are true? Are these inequalities, together with (1), sufficient to characterise the $\alpha, \beta,$ and γ that can be eigenvalues of Hermitian matrices A, B and $A + B$?

In a fundamental paper in 1962, Alfred Horn made a conjecture that asserted that these inequalities, together with (1), are sufficient and that the set T_r^n of triples (I, J, K) of cardinality r in $[1, n]$ can be described by induction on r as follows. Let us write $I = \{i_1 < i_2 < \cdots < i_r\}$ and likewise for J and K. Then, for $r = 1$, (I, J, K) is in T_1^n if $k_1 = i_1 + j_1 - 1$. For $r > 1$, $(I, J, K) \in T_r^n$ if

$$\sum_{i\in I} i + \sum_{j\in J} j = \sum_{k\in K} k + \begin{pmatrix} r+1 \\ 2 \end{pmatrix} \tag{10}$$

and, for all $1 \le p \le r-1$ and all $(U,V,W) \in T_p^r$

$$\sum_{u\in U} i_u + \sum_{v\in V} j_v \le \sum_{w\in W} k_w + \begin{pmatrix} p+1 \\ 2 \end{pmatrix}. \tag{11}$$

Horn proved his conjecture for $n = 3$ and 4. Note that when $n = 2$, these conditions just reduce to the three inequalities given by (3). When $n = 3$, they reduce to the twelve inequalities (6)–(8). When $n = 7$, there are 2062 inequalities given by these conditions.

Horn's conjecture has finally been proved by A Klyachko (1998) and A Knutson and T Tao (1999) (see [2], [3]).

It turns out that this problem has some remarkable connections with problems in algebraic geometry and the representation theory of Lie groups. Let us indicate briefly the connection with algebraic geometry.

The classical minimax principle of Courant, Fischer, and Weyl says that the eigenvalues α_j of the Hermitian matrix A are characterised by extremal relations

$$\alpha_j = \max_{\dim V = j} \min_{x\in V,\ \|x\|=1} \operatorname{tr}(Axx^*) \tag{12}$$

Here, dim V stands for the dimension of a subspace V of $\mathbb{C}^n$. Note that xx^* is just the orthogonal projection operator on the 1-dimensional subspace spanned by x. Note also that tr Axx^* is just the number $x^*Ax = \langle x, Ax\rangle$.

The complex Grassmann manifold $G_k(\mathbb{C}^n)$ is the set of all k-dimensional linear subspaces of $\mathbb{C}^n$. For $k = 1$, this is just the complex projective space $\mathbb{C}\boldsymbol{P}^{n-1}$, the set of all complex lines through the origin in the space $\mathbb{C}^n$. Each k-dimensional subspace L of $\mathbb{C}^n$ is completely characterised by the orthogonal projection P_L with range L.

Given any Hermitian operator A on $\mathbb{C}^n$, let $A_L = P_L A P_L$. Note that tr $A_L =$ tr $P_L A P_L =$ tr AP_L. To prove the inequality (5), Wielandt invented a remarkable minimax principle. This says that for any $1 \le i_1 < \cdots < i_k \le n$

$$\sum_{j=1}^{k} \alpha_{i_j} = \max_{\substack{V_1\subset\cdots\subset V_k \\ \dim V_j = i_j}} \min_{\substack{L\in G_k(\mathbf{C}^n) \\ \dim (L\cap V_j)\ge j}} \operatorname{tr} A_L. \tag{13}$$

Note for $k = 1$, this reduces to (12).

Another such principle was discovered by Hersch and Zwahlen. Let v_j be the eigenvectors of the Hermitian matrix A corresponding to its eigenvalues α_j. For $m = 1,\ldots,n$, let V_m be the linear span of $v_1,\ldots,v_m$. Then, for any $1 \le i_1 < \cdots < i_k \le n$,

$$\sum_{j=1}^{k} \alpha_{i_j} = \min_{L\in G_k(\mathbf{C}^n)} \left\{\operatorname{tr} A_L :\ \dim (L\cap V_{i_j}) \ge j,\ j = 1,\ldots,k\right\}. \tag{14}$$

The Grassmannian $G_k(\mathbb{C}^n)$ is a smooth compact manifold of real dimension $2k(n-k)$. There is a famous embedding called Plücker embedding via which $G_k(\mathbb{C}^n)$ is realised as a projective variety in the space $\mathbb{C}\boldsymbol{P}^N$, where $N = \binom{n}{k} - 1$.

A sequence of nested subspaces $\{0\} \subset V_1 \subset V_2 \subset \ldots \subset V_n = \mathbb{C}^n$, where $\dim V_j = j$, is called a *flag*. Given a flag $\mathcal{F}$ and a set of indices $1 \leq i_1 < \cdots < i_k \leq n$ the subset

$$\{W \in G_k(\mathbb{C}^n) : \ \dim (W \cap V_{i_j}) \geq j, \quad j = 1, \ldots, k\}$$

of the Grassmanian is called a *Schubert variety*.

The principle (14) thus says that the sum $\sum \alpha_{i_j}$ is characterised as the minimal value of $\operatorname{tr} A_L$ evaluated on the Schubert variety corresponding to the flag constructed from the eigenvectors of A.

This suggests that inequalities like the ones conjectured by Horn could be related to Schubert calculus, a component of algebraic geometry dealing with intersection properties of flags. This line was pursued vigorously by R. C. Thompson beginning in the early seventies. Finally, the problem has now been solved by the efforts of several others using Schubert calculus.

There are other ways to look at Horn's inequalities. The matrices X and Y are said to be *unitarily equivalent* if there exists a unitary matrix U such that $X = UYU^*$. Two Hermitian matrices are unitarily equivalent if and only if they have the same eigenvalues. It is easy to see that Horn's conjecture (now proved) amounts to the following. Given Hermitian matrices A, B, consider the collection of all n-tuples that arise as eigenvalues of $A + UBU^*$ as U varies over all unitary matrices (with the convention that the eigenvalues of a Hermitian matrix are counted in decreasing order). Horn's inequalities assert that this is a convex polytope in $\mathbb{R}^n$ whose faces are characterised by the conditions (1), (10) and (11).

Postscript A fuller exposition of Horn's problem and its solution appeared in the article: R Bhatia, Linear Algebra to Quantum Cohomology: The Story of Alfred Horn's Inequalities, *American Mathematical Monthly,* Vol. 108, pp. 289–318, 2001.

Orthogonalisation of Vectors*

Matrix Decompositions and Approximation Problems

Rajendra Bhatia
Indian Statistical Institute, New Delhi 110 016, India.

1. The Gram-Schmidt Process

The Gram-Schmidt process is one of the first things one learns in a course on vectors or matrices. Let us recall it briefly.

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a vector with n coordinates x_j, each of which is a complex number. The collection of all such vectors is the vector space $\mathbb{C}^n$. It helps to think of $\boldsymbol{x}$ as a column vector and write $\boldsymbol{x}^*$ for the row vector with coordinates $\bar{x}_j$. The *inner product* (or the *scalar product*) between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is the number

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^* \boldsymbol{y} = \sum_{j=1}^{n} \bar{x}_j y_j.$$

The *norm* of $\boldsymbol{x}$ is defined as

$$\|\boldsymbol{x}\| = (\boldsymbol{x}^* \boldsymbol{x})^{\frac{1}{2}} = \left(\sum_{j=1}^{n} |x_j|^2 \right)^{\frac{1}{2}}.$$

If we are given n linearly independent vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$, the Gram-Schmidt process constructs an orthonormal basis out of them as follows. We put $\boldsymbol{q}_1 = \boldsymbol{a}_1 / \|\boldsymbol{a}_1\|$. This vector has norm 1. We now put $\boldsymbol{v}_2 = \boldsymbol{a}_2 - \langle \boldsymbol{q}_1, \boldsymbol{a}_2 \rangle \boldsymbol{q}_1$; and $\boldsymbol{q}_2 = \boldsymbol{v}_2 / \|\boldsymbol{v}_2\|$. Then $\boldsymbol{q}_2$ is orthogonal to $\boldsymbol{q}_1$ and has norm 1. At the next stage, we put $\boldsymbol{v}_3 = \boldsymbol{a}_3 - \langle \boldsymbol{q}_1, \boldsymbol{a}_3 \rangle \boldsymbol{q}_1 - \langle \boldsymbol{q}_2, \boldsymbol{a}_3 \rangle \boldsymbol{q}_2$; and $\boldsymbol{q}_3 = \boldsymbol{v}_3 / \|\boldsymbol{v}_3\|$. Continuing this way we obtain an orthonormal basis $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$. Note that for each $1 \leq k \leq n$, the linear spans of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ and $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k$ are equal.

How close are the vectors $\{\boldsymbol{q}_j\}$ to the original vectors $\{\boldsymbol{a}_j\}$? To make this precise let us define the distance between two ordered sets $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ and $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$ of vectors in $\mathbb{C}^n$ as

$$\left(\sum_{j=1}^{k} \|\boldsymbol{x}_j - \boldsymbol{y}_j\|^2 \right)^{\frac{1}{2}}. \tag{1}$$

Note that each $\boldsymbol{x}_j$ is an n-vector. If we write it as $\boldsymbol{x}_j = (\boldsymbol{x}_{j1}, \ldots, \boldsymbol{x}_{jn})$, then the quantity in (1) is

$$\left(\sum_{j=1}^{k} \sum_{r=1}^{n} |x_{jr} - y_{jr}|^2 \right)^{\frac{1}{2}}. \tag{2}$$

*Reproduced from *Resonance*, Vol. 5, No. 3, pp. 52–59, March 2000. (General Article)

Let us consider a very simple example in the space $\mathbb{C}^2$. Let $\boldsymbol{a}_1 = (1, 0), \boldsymbol{a}_2 = (\frac{4}{5}, \frac{3}{5})$. The vectors $\boldsymbol{a}_1, \boldsymbol{a}_2$ are linearly independent and each of them has norm 1. However, they are not orthogonal to each other. The Gram-Schmidt process applied to them gives the vectors $\boldsymbol{q}_1 = (1, 0), \boldsymbol{q}_2 = (0, 1)$. The distance between the pair $\{\boldsymbol{a}_1, \boldsymbol{a}_2\}$ and the pair $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ is $(\frac{4}{5})^{\frac{1}{2}}$. Can we find another pair of orthonormal vectors that is closer to $\{\boldsymbol{a}_1, \boldsymbol{a}_2\}$. If we try the obvious possibilities that the form of $\boldsymbol{a}_1, \boldsymbol{a}_2$ suggests, we soon find that the pair $\boldsymbol{y}_1 = (\frac{4}{5}, -\frac{3}{5}), \boldsymbol{y}_2 = (\frac{3}{5}, \frac{4}{5})$ is at distance $(\frac{12}{25})^{\frac{1}{2}}$ from $\{\boldsymbol{a}_1, \boldsymbol{a}_2\}$. Thus the Gram-Schmidt process while constructing an orthonormal basis can take us far away from the original set of vectors.

Another pair that is even closer to $\{\boldsymbol{a}_1, \boldsymbol{a}_2\}$ is the pair $\boldsymbol{u}_1 = (\frac{2}{\sqrt{5}}, -\frac{1}{\sqrt{5}}), \boldsymbol{u}_2 = (\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})$. One can see that the distance of this pair from $\{\boldsymbol{a}_1, \boldsymbol{a}_2\}$ is $(4 - \frac{8}{\sqrt{5}})^{\frac{1}{2}}$. Thus the three pairs $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$, $\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$ and $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ are at distance .8944, .6928 and .6498, respectively from the given pair $\{\boldsymbol{a}_1, \boldsymbol{a}_2\}$.

One can see, using Lagrange multipliers, that among all pairs of orthonormal vectors, the pair $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ is the closest to $\{\boldsymbol{a}_1, \boldsymbol{a}_2\}$. We will soon see this by another argument.

The problem of finding the orthonormal basis closest to a given set of linearly independent vectors is of interest in quantum chemistry. In many models of atomic phenomena some of the quantities of interest are represented by orthonormal vectors. Experimental observations to measure these quantities are inaccurate and thus give us vectors that are not orthonormal. We might want to stay as close to the experimental data as possible when converting these vectors to orthonormal ones demanded by the model. The process of finding the closest orthonormal basis is called the *Löwdin Orthogonalisation* after the Swedish chemist P O Löwdin who introduced it. This is related to one of the basic theorems in linear algebra as we will see.

2. Matrix Approximation Problems

Let A be an $n \times n$ matrix with entries a_{ij}. Let A^* be the conjugate transpose of A-the matrix whose i, j entry is $\bar{a}_{ji}$. Let tr A stand for the trace of A. The *Hilbert-Schmidt norm* (or the *Frobenius norm*) of A is defined as

$$\|A\|_2 = \left(\sum_{i,j} |a_{ij}|^2 \right)^{1/2} = (\text{tr } A^*A)^{1/2}. \tag{3}$$

This norm is *unitarily invariant*: if U, V are unitary matrices, then

$$\|A\|_2 = \|UAV\|_2. \tag{4}$$

This is so because

$$\text{tr } (UAV)^*(UAV) = \text{tr } V^*A^*AV = \text{tr } A^*A. \tag{5}$$

Note that if $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}$ are elements of $\mathbb{C}^n$ and if we write the $n \times n$ matrix A whose

columns are $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ as $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$, then

$$\|A\|_2^2 = \sum_j \|\boldsymbol{a}_j\|^2.$$

The matrix A is invertible if and only if its columns are linearly independent as vectors, and it is unitary if and only if they are orthonormal. Thus the problem of finding the orthonormal basis closest to a given set of n linearly independent vectors is the same as the problem of finding the unitary matrix closest to a given invertible matrix. Here the closest matrix is one whose distance in the Hilbert-Schmidt norm from the given matrix is minimal.

This is a typical example of a matrix approximation problem.

3. The QR and the Polar Decompositions

The Gram-Schmidt process can be represented as an interesting matrix factoring theorem:

Every invertible matrix A can be factored as $A = QR$, where Q is unitary and R is upper triangular. We can choose R so that all its diagonal entries are positive. With this restriction Q and R are unique.

It is not difficult to see how this theorem follows from the Gram-Schmidt process. The columns of Q are orthonormal vectors constructed from the columns of A. The fact that $\{a_1, \ldots, a_k\}$ span the same linear space as $\{q_1, \ldots, q_k\}$ is reflected in the upper triangular form of R. The vectors Q are unique upto a multiplication by a complex number of modulus one. So, the restriction that the diagonal entries of R be positive imposes uniqueness.

The decomposition $A = QR$ is called the *QR decomposition*. If A is singular, it still has a QR decomposition. Now some of the rows of R are zero.

There is another factoring of an invertible matrix into two factors one of which is unitary. This is the *polar decomposition*:

Every invertible matrix A can be factored uniquely as $A = UP$, where U is unitary and P is positive definite.

The factor P is the unique positive definite square root of the positive definite matrix A^*A. If one puts $U = AP^{-1}$, then $U^*U = UU^* = I$. If A is singular, it still has a polar decomposition $A = UP$. Now the factor U is not unique, but P is.

The polar decomposition has an interesting extremal characterisation:

Theorem. *Among all unitary matrices the one closest to A is the matrix U in the polar decomposition A = UP.*

Proof. Let W be any unitary matrix. Then

$$\|A - W\|_2 = \|UP - W\|_2 = \|P - U^*W\|_2,$$

by the unitary invariance property (4). Thus to find the unitary matrix closest to A it suffices to find the one closest to P. If we show that the unitary matrix closest to P is the identity matrix I it will follow that the unitary matrix closest to UP is U.

For every unitary matrix V

$$\|P - V\|_2^2 = \text{tr}\,(P - V^*)(P - V) = \text{tr}\,(P^2 + I - PV - V^*P).$$

This quantity is minimum when

$$\text{tr}\,(PV + V^*P) = \text{tr}\,P(V + V^*) \tag{6}$$

is maximum. The trace is not affected if we apply a unitary similarity (i.e., $\text{tr}\,X = \text{tr}\,WXW^*$, for all X and unitary W). The spectral theorem tells us that we can apply such a similarity to bring V to the diagonal form. Thus we may assume that V is diagonal with entries $e^{i\theta_j}, 1 \le j \le n$ down its diagonal. So, the quantity in (6) is

$$\text{tr}\,P(V + V^*) = 2\sum_j p_{jj}\cos\theta_j.$$

Since $p_{jj} \ge 0$, clearly this is maximised when $\cos\theta_j = 1$. This translates to the condition $V = I$. ■

Thus the polar decomposition provides the basis for the Löwdin Orthogonalisation. The orthonormal basis closest to a set of linearly independent vectors $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}$ is obtained by writing the matrix $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$, then finding its polar decomposition $A = UP$, and reading the columns of $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$ to get the desired orthonormal basis $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$.

This explains the example discussed in Section 1. We have the polar decomposition

$$\begin{bmatrix} 1 & \frac{4}{5} \\ 0 & \frac{3}{5} \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix}.$$

Since $P = WSW^*$, where W is unitary and S diagonal with positive entries, we can write $A = UP = UWSW^* = VSW^*$, where V is unitary. This is called the *singular value decomposition* of A. To find the factors here, we have to diagonalise P. This involves a more elaborate calculation than the one for the Gram-Schmidt process.

4. The closest Hermitian matrix

The problem of finding the closest Hermitian matrix to a given matrix is motivated by the same considerations as that of finding the closest unitary matrix. It is simpler to solve this.

If $A = B + iC$, where B and C are Hermitian, then

$$\|A\|_2^2 = \text{tr}\,A^*A = \text{tr}\,(B - iC)(B + iC) = \text{tr}\,(B^2 + C^2) = \|B\|_2^2 + \|C\|_2^2.$$

Every matrix has a decomposition of this kind:

If we put $B = \frac{1}{2}(A+A^*)$ and $X = \frac{1}{2i}(A-A^*)$, then B, C are Hermitian and $A = B+iC$. This is analogous to the decomposition $z = x+iy$ of a complex number into its real and imaginary parts. For this reason B and C are called the *real* and *imaginary parts* of A and the decomposition $A = B + iC$ is called the *Cartesian decomposition*.

Now, if H is any Hermitian matrix, then

$$||A - H||_2^2 = ||H - B||_2^2 + ||C||_2^2.$$

Clearly, the choice $H = B$ minimises this quantity. Thus the Hermitian matrix closest to A is the real part of A.

The polar decomposition $A = UP$ can be thought of as the analogue of the polar representation $z = e^{i\theta}r$ of a complex number. Thus the statements about the closest unitary and Hermitian matrices proved above are analogues of the facts about the point on the unit circle and the point on the real line closest to a given complex number.

A matrix is said to be *normal* if $AA^* = A^*A$. This is equivalent to the condition that the factors U and P in the polar decomposition of A commute. Evidently Hermitian matrices and unitary matrices are normal.

The set of all Hermitian matrices is a real vector space; the set of all unitary matrices is a differentiable manifold. The set of all normal matrices does not have any nice geometric structure. This is one reason why the problem of finding the closest normal matrix to a given matrix turns out to be much harder than the problems we have considered. This problem is not yet solved completely. See [2] for a discussion, and also for examples of other problems where the solution for normal matrices is much harder than that for Hermitian or unitary matrices.

5. Approximation in other norms

The Hilbert-Schmidt norm is the simplest norm on matrices from the point of view of approximation problems. This is because it is like the Euclidean norm on vectors. There are other norms that are of interest. For example, if we think of A as a linear operator on $\mathbb{C}^n$, then the *operator norm* of A is defined as

$$||A|| = \max\{||Ax|| : x \in \mathbb{C}^n, ||x|| = 1\}.$$

Like the Hilbert-Schmidt norm, this norm is also unitarily invariant. There are several other norms on matrices that are unitarily invariant.

The answer to a minimisation problem often changes with the norm. That is natural, because the functions being minimised are different.

It is, therefore, interesting to know that for *every* unitarily invariant norm $|||\cdot|||$ on the space of matrices, the minimum of $|||A - W|||$ over unitary matrices is attained when W is the unitary factor in the polar decomposition of A; and the minimum of $|||A - H|||$ over Hermitian matrices is attained when $H = \frac{1}{2}(A + A^*)$.

Box 1.

Let A^* be the matrix obtained from A by taking the transpose of A and then replacing each entry by its complex conjugate. A matrix A is called *Hermitian* if $A = A^*$. A Hermitian matrix all whose eigenvalues are positive is called *positive definite*. An invertible matrix A is called *unitary* if $A^{-1} = A^*$. A is called *normal* if $AA^* = A^*A$. Hermitian matrices and unitary matrices are special kinds of normal matrices.

The *Spectral Theorem* says that every normal matrix A can be *diagonalised* by a *unitary conjugation*; i.e., there exists a unitary matrix U and a diagonal matrix D such that $A = UDU^*$. The diagonal entries of D are complex numbers. They are real if A is Hermitian, positive if A is positive definite, and complex numbers of modulus one if A is unitary.

Suggested Reading

[1] A detailed discussion of the polar and the *QR* deompositions may be found in H Helson, *Linear Algebra,* TRIM 4, Hindustan Book Agency, 1994.

[2] A more advanced treatment of matrix approximation problems may be found in R Bhatia, *Marix Analysis,* Springer-Verlag, 1997.

[3] The relevance of matrix approximation problems to quantum chemistry is explained in the article J A Goldstein and M Levy, Linear algebra and quantum chemistry, *American Math. Monthly,* **78**, 710–718, 1991.

[4] The Löwdin Orthogonalisation was proposed by P O Löwdin, On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals, *J. Chem. Phys.,* **18**, 365–374, 1950.

[5] Algorithms for finding the QR and the Singular Value Decompositions are discussed in G Golub and C Van Loan, *Matrix Computations,* The Johns Hopkins University Press, 1983.

Triangularization of a Matrix*

Rajendra Bhatia and Radha Mohan**
Indian Statistical Institute, New Delhi 110 016
E-mail: rbh@isid.ac.in
***Centre for Mathematical Sciences, St. Stephen's College, Delhi 110 007,*
E-mail: mohan_radha@hotmail.com

Two ideas that pervade all of mathematics are *equivalence*, and the related notion of *reduction*. If an object in a given class can be carried into another by a transformation of a special kind, we say the two objects are equivalent. Reduction means the transformation of the object into an equivalent one with a special form as simple as possible.

The group of transformations varies with the problem under study. In linear algebra, we consider arbitrary non-singular linear transformations while studying algebraic questions. In problems of geometry and analysis, where distances are preserved, unitary (orthogonal) transformations alone are admitted. In several problems of crystallography and number theory, the interest is in linear transformation with integral coefficients and determinant one.

In this article we restrict ourselves to $n \times n$ complex matrices. Two such matrices A and B are said to be *similar* if there exists a non-singular (invertible) matrix S such that $B = S^{-1}AS$. If this S can be chosen to be unitary ($S^{-1} = S^\star$) we say that A and B are *unitarily similar*. Similar matrices are representations of the same linear operator on $\mathbb{C}^n$ in two different bases. Unitarily similar matrices represent the same linear operator but in two different *orthonormal* bases. Similarity and unitary similarity are equivalence relations.

Similarity preserves (does not change) the rank, determinant, trace and eigenvalues of a matrix. Unitary similarity preserves all these and more. For example if A is Hermitian ($A = A^\star$), then every matrix unitarily similar to it is Hermitian too. If we define the *norm* of any matrix A as

$$\| A \|_2 = \left(\sum_{i,j} |a_{ij}|^2 \right)^{1/2},$$

then every matrix unitarily similar to A has the same norm. The simplest way to see this is to note that

$$\| A \|_2 = (\mathrm{tr} A^\star A)^{1/2} = \| U^\star A U \|_2,$$

where tr stands for the trace of a matrix.

It is generally agreed that the more zero entries a matrix has, the simpler it is. Much of linear algebra is devoted to reducing a matrix (via similarity or unitary similarity) to another that has lots of zeros.

The simplest such theorem is the *Schur Triangularization Theorem*. This says that *every matrix is unitarily similar to an upper triangular matrix.*

*Reproduced from *Resonance*, Vol. 5, No. 6, pp. 40–48, June 2000. (General Article)

Our aim here is to show that though it is very easy to prove it, this theorem has many interesting consequences.

Proof of Schur's Theorem

We want to show that given an $n \times n$ matrix A, there exists a unitary matrix U and an upper triangular matrix T such that $A = UTU^\star$. This is equivalent to saying that there exists an orthonormal basis for $\mathbb{C}^n$ with respect to which the matrix of the linear operator A is upper triangular. In other words, there exists an orthonormal basis $v_1, \ldots, v_n$ such that for each $k = 1, 2, \ldots n$, the vector Av_k is a linear combination of $v_1, \ldots, v_k$.

This can be proved by induction on n. Let λ_1 be an eigenvalue of A and v_1 an eigenvector of norm one corresponding to it. Let M be the one-dimensional subspace of $\mathbb{C}^n$ spanned by v_1, and let N be its orthogonal complement. Let P_N be the orthogonal projection with range N. For $y \in N$, let $A_N y = P_N A y$. Then A_N is a linear operator on the $(n-1)$-dimensional space N. By the induction hypothesis, there exists an orthonormal basis $v_2, \ldots, v_n$ of N such that the vector $A_N v_k$ for $k = 2, \ldots n$ is a linear combination of $v_2, \ldots, v_k$. The set $v_1, \ldots, v_n$ is an orthonormal basis for $\mathbb{C}^n$ and each $Av_k, 1 \leq k \leq n$, is a linear combination of $v_1, \ldots, v_k$. This proves the theorem. The basis $v_1, \ldots, v_n$ is called a *Schur basis* for A.

Notice that we started our argument by choosing an eigenvalue and eigenvector of A. Here we have used the fact that we are considering complex matrices only. The diagonal entries of the upper triangular matrix T are the eigenvalues of A. Hence, they are uniquely specified up to permutation. The entries of T above the diagonal are not unique. Since,

$$\sum_{i,j} |t_{ij}|^2 = \sum_{i,j} |a_{ij}|^2,$$

they can not be too large. The reader should construct two 3×3 upper triangular matrices which are unitarily similar.

The Spectral Theorem

A matrix A is said to be *normal* if $AA^\star = A^\star A$. Hermitian and unitary matrices are normal.

The Spectral Theorem says that a *normal matrix is unitarily similar to a diagonal matrix*.

This is an easy consequence of Schur's theorem: Note that the property of being normal is preserved under unitary similarity, and check that an upper triangular matrix is normal if and only if it is diagonal.

The Schur basis for a normal matrix A is thus a basis consisting of eigenvectors of A. Normal matrices are, therefore, matrices whose eigenvectors form an orthonormal basis for $\mathbb{C}^n$.

Some Density Theorems

A subset Y of a metric space X is said to be *dense* if every neighbourhood of a point in X contains a point of Y. This is equivalent to saying that every point in X is the limit of a sequence of points in Y. (The set of rational numbers and the set of irrational numbers are dense in $\mathbb{R}$.)

The space $\mathbb{M}(n)$ consisting of $n \times n$ matrices is a metric space if we define for every pair A, B the distance between them as $d(A, B) = \| A - B \|_2$. We will show that certain subsets are dense in $\mathbb{M}(n)$. The argument in each case will have some common ingredients. The property that characterizes the subset Y in question will be one that does not change under unitary similarity. So, if $A = UTU^\star$ and we show the existence of an element of Y in an ϵ-neighbourhood of an upper triangular T, then we would have also shown the existence of an element of Y in an ϵ-neighbourhood of A.

Invertible matrices are dense. A matrix is invertible if and only if it does not have zero as an eigenvalue. This property is not affected by unitary similarity. We want to show that if A is any matrix then for every $\epsilon > 0$, there exists an invertible matrix B such that $\| A - B \|_2 < \epsilon$. Let $A = UTU^\star$, where T is upper triangular. If A is singular some of the diagonal entries of T are zero. Replace them by small non-zero numbers so that for the new upper triangular matrix T' obtained after these replacements we have $\| T - T' \|_2 < \epsilon$. Then T' is invertible and so is $A' = UT'U^\star$. Further,

$$\| A - A' \|_2 = \| U(T - T')U^\star \|_2 < \epsilon.$$

Matrices with distinct eigenvalues are dense. Use the same argument as above. If any two diagonal entries of T are equal, change one of them slightly.

Diagonalizable matrices are dense. A matrix is said to be *diagonalizable* if it is similar to a diagonal matrix; i.e. if it has n linearly independent eigenvectors. Since eigenvectors corresponding to distinct eigenvalues of any matrix are linearly independent, every matrix with distinct eigenvalues is diagonalizable. (The converse in not true). So the set of diagonalizable matrices includes a dense set (matrices with distinct eigenvalues) and hence is itself dense.

These density theorems are extremely useful. Often it is easy to prove a statement for invertible or diagonalizable matrices. Then one can extend it to all matrices by a limiting procedure. We give some examples of this argument.

The exponential of a matrix is defined as

$$e^A = I + A + \frac{A^2}{2!} + \cdots .$$

(The series is convergent.) We want to calculate the determinant $\det(e^A)$. It turns out that $\det(e^A) = e^{\mathrm{tr}(A)}$. This is obviously true if A is a diagonal matrix: if the diagonal entries of A are $\lambda_1, \ldots, \lambda_n$ then $\det(e^A) = e^{\lambda_1} \ldots e^{\lambda_n} = e^{\lambda_1 \ldots \lambda_n} = e^{\mathrm{tr}(A)}$. From this one can see that this equality is also true for diagonalizable matrices; just note that $e^{SAS^{-1}} = Se^AS^{-1}$. Finally, the equality carries over to all matrices since both sides are continuous functions of a matrix and every matrix is a limit of diagonalizable matrices.

Let A, B be any two matrices. We know that $\det(AB) = \det(BA)$, and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. More generally, it is true that AB and BA have the same characteristic polynomial and hence the same eigenvalues (including multiplicities). Recall that the k-th coefficient in the characteristic polynomial of A is (up to a sign) the sum of $k \times k$ principal minors of A. These are polynomial functions of the entries of A, and hence depend continuously on A. Thus, to prove that AB and BA have the same characteristic polynomial, it is enough to prove this when B belongs to a dense subset of $\mathbb{M}(n)$. The set of invertible matrices is such a set. But if B is invertible, then $B(AB)B^{-1} = BA$, i.e. AB and BA are similar. Hence, they have the same characteristic polynomial.

This theorem, in turn is very useful in several contexts. Let A and B be two positive semidefinite matrices. Then all their eigenvalues are non-negative. The product AB is not Hermitian (unless A and B commute), so *a priori* it is not even clear whether AB has real eigenvalues. We can, in fact, prove that it has non-negative real eigenvalues. Let $B^{1/2}$ be the unique positive square root of B. Then $AB = (AB^{1/2})B^{1/2}$ and this has the same eigenvalues as $B^{1/2}AB^{1/2}$. This matrix is positive semidefinite, and hence has non-negative eigenvalues.

The *Cayley Hamilton Theorem* says that every matrix satisfies its characteristic equation; i.e. if $\chi(z)$ is the polynomial in the variable z obtained by expanding $\det(zI - A)$, and $\chi(A)$ is the matrix obtained from this polynomial on replacing z by A, then $\chi(A) = 0$. The reader is invited to write a proof for this using the above ideas; the proof is easy for diagonal matrices.

A Bound for Eigenvalues

In many problems it is of interest to calculate the eigenvalues of a matrix A. This is not always easy. Sometimes, it helps to know the eigenvalues approximately, or at least that they lie (or do not lie) in some region of the complex plane. From Schur's Theorem, it is clear that, if λ_i are the eigenvalues of A, then

$$\sum_{i=1}^{n} |\lambda_i|^2 \leq \sum_{i,j} |a_{ij}|^2.$$

The two sides are equal if and only if A is normal.

This leads to an amusing (but not the easiest) proof of the arithmetic-geometric mean inequality. Let $a_1, \ldots, a_n$ be non-negative numbers. The eigenvalues of the matrix

$$A = \begin{pmatrix} 0 & a_1 & 0 & \ldots & 0 \\ 0 & \ddots & a_2 & \ldots & 0 \\ 0 & 0 & \ddots & \ddots & \\ 0 & 0 & & \ddots & a_{n-1} \\ a_n & 0 & & \ldots & 0 \end{pmatrix}$$

are the n-th roots of $a_1 a_2 \ldots a_n$. Hence by the above inequality

$$n(a_1 a_2 \ldots a_n)^{2/n} \leq {a_1}^2 + \cdots + {a_n}^2.$$

Changing ${a_i}^2$ to a_i, we get the inequality

$$(a_1 a_2 \dots a_n)^{1/n} \leq \frac{a_1 + \dots + a_n}{n}$$

between the geometric mean and the arithmetic mean. We even get the condition for equality; just note that A is normal if and only if $a_1 = a_2 = \dots = a_n$.

Here is a more serious and powerful application of these ideas.

Theorem. *If A, B are normal matrices such that AB is normal, then BA is also normal.*

Proof. Let $\lambda_i(AB), 1 \leq i \leq n$, be the eigenvalues of AB. Since AB is normal

$$\sum_{i=1}^{n} |\lambda_i(AB)|^2 = \| AB \|_2{}^2.$$

To prove that BA is normal, we have to show that this is true when AB is replaced by BA. We have seen that $\lambda_i(AB) = \lambda_i(BA)$. So, we have to show that

$$\| AB \|_2{}^2 = \| BA \|_2{}^2,$$

i.e.,

$$\mathrm{tr}(B^\star A^\star AB) = \mathrm{tr}(A^\star B^\star BA).$$

Using the fact that $\mathrm{tr}(XY) = \mathrm{tr}(YX)$ for all matrices X, Y, and the normality of A, B, the two sides of this desired equality are seen to be equal to $\mathrm{tr}(AA^\star BB^\star)$. This proves the theorem. □

The reader might try to find another proof of this theorem. (If the reader is unable to find such a proof from the mere definition of normality, she should not be surprised. The statement is false in infinite-dimensional Hilbert spaces. It is, however, true if one of the operators A or B is compact.)

Commuting Matrices

Let A and B be two matrices. Schur's Theorem tells us that there exist unitary matrices U, V and upper triangular matrices R, T such that $A = URU^\star, B = VTV^\star$. It turns out that if A and B commute ($AB = BA$), then we can choose $U = V$. In other words, if A and B commute, they have a common Schur basis.

To prove this, we first show that A, B have a common eigenvector. Let λ be an eigenvalue of A, and let $W = \{x : Ax = \lambda x\}$ be the associated eigenspace. If $x \in W$, then

$$ABx = B(Ax) = B(\lambda x) = \lambda(Bx).$$

Thus, $Bx \in W$. This says that the space W is *invariant* under B. So, there exists $y \in W$ such that $By = \mu y$. This y is a common eigenvector for A and B.

The rest of the proof is similar to the one we gave earlier for Schur's Theorem.

The same argument shows that if $\{A_\alpha\}$ is any family of pairwise commuting matrices, then all A_α have a common Schur basis.

Distance between Eigenvalues

Let A and B be commuting matrices with eigenvalues $\lambda_1, \ldots, \lambda_n$ and $\mu_1, \ldots, \mu_n$ respectively. We have seen that there exists a unitary matrix U such that $A = UTU^\star, B = UT'U^\star$. The diagonal entries of T and T' are the numbers $\lambda_1, \ldots, \lambda_n$ and $\mu_1, \ldots, \mu_n$ (in some order). Hence,

$$\left(\sum_{i=1}^{n} |\lambda_i - \mu_i|^2 \right)^{1/2} \leq \| T - T' \|_2 \leq \| A - B \|_2 \, .$$

Thus, it is possible to enumerate the n-tuples $\{\lambda_j\}$ and $\{\mu_j\}$ so that the *distance between them is smaller than the distance between A and B* (in the sense made precise by this inequality).

This is no longer true if A and B do not commute. For example, consider

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

A famous theorem of Hoffman and Wielandt says that if A and B both are normal, then the above inequality is true even when A, B do not commute.

This article is based on a talk given by the first author at a Refresher Course for College Teachers organized by CPDHE in April 1999.

Suggested Reading

[1] S Axler, *Linear Algebra Done Right*, Springer Verlag, New York, 1997.

[2] H Helson, *Linear Algebra*, TRIM, Hindustan Book Agency, 1994.

[3] I N Herstein and D J Winter, *Matrix Theory and Linear Algebra*, Macmillan, New York, 1989.

[4] R A Horn and C R Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1991.

Eigenvalues of AB and BA*

Rajendra Bhatia
Indian Statistical Institute, New Delhi 110 016, India.

Let A, B be $n \times n$ matrices with complex entries. Given below are several proofs of the fact that AB and BA have the same eigenvalues. Each proof brings out a different viewpoint and may be presented at the appropriate time in a linear algebra course.

Let $\mathrm{tr}(T)$ stand for the trace of T, and $\det(T)$ for the determinant of T. The relations

$$\mathrm{tr}(AB) = \mathrm{tr}(BA) \quad \text{and} \quad \det(AB) = \det(BA). \tag{1}$$

are usually proved early in linear algebra courses.

Let

$$\lambda^n - c_1(T)\lambda^{n-1} + \cdots + (-1)^n c_n(T) \tag{2}$$

be the characteristic polynomial of T, and let $\lambda_1(T), \lambda_2(T), \ldots, \lambda_n(T)$ be its n roots, counted with multiplicities and in any order. These are the eigenvalues of T. We know that $c_k(T)$ is the kth elementary symmetric polynomial in these n numbers. Thus

$$\begin{aligned} c_1(T) &= \sum_{j=1}^{n} \lambda_j(T) = \mathrm{tr}\,(T) \\ c_2(T) &= \sum_{i<j} \lambda_i(T)\lambda_j(T) \\ &\vdots \\ c_n(T) &= \prod_{j=1}^{n} \lambda_j(T) = \det(T). \end{aligned}$$

To say that AB and BA have the same eigenvalues amounts to saying that

$$c_k(AB) = c_k(BA) \quad \text{for} \quad 1 \le k \le n. \tag{3}$$

We know that this is true when $k = 1$, or n; and want to prove it for other values of k.

Proof 1. It suffices to prove that, for $1 \le m \le n$,

$$\lambda_1^m(AB) + \cdots + \lambda_n^m(AB) = \lambda_1^m(BA) + \cdots + \lambda_n^m(BA). \tag{4}$$

(Recall Newton's identities by which the n elementary symmetric polynomials in n variables are expressed in terms of the n sums of powers.) Note that the eigenvalues of T^m are the mth

*Reproduced from *Resonance*, Vol. 7, No. 1, pp. 88–93, January 2002. (Classroom)

powers of the eigenvalues of T. So, $\sum \lambda_j^m(T) = \sum \lambda_j(T^m) = \text{tr}\,(T^m)$. Thus the statement (4) is equivalent to

$$\text{tr}\,[(AB)^m] = \text{tr}\,[(BA)^m].$$

But this follows from the first equation in (1) :

$$\text{tr}\,[(AB)^m] = \text{tr}\,(ABAB\cdots AB) = \text{tr}(BABA\ldots BA) = \text{tr}\,[(BA)^m].$$

Proof 2. One can prove the relations (3) directly. The coefficient $c_k(T)$ is the sum of all the $k \times k$ principal minors of T. A direct computation (the Binet-Cauchy formula) leads to the equations (3). A more sophisticated version of this argument involves the antisymmetric tensor product $\wedge^k(T)$. This is a matrix of order $\binom{n}{k}$ whose entries are the $k \times k$ minors of T. So

$$c_k(T) = \text{tr}\ \wedge^k(T),\ 1 \le k \le n.$$

Among the pleasant properties of $\wedge^k$ is multiplicativity: $\wedge^k(AB) = \wedge^k(A)\,\wedge^k(B)$. So

$$\begin{aligned} c_k(AB) &= \text{tr}\,[\wedge^k(AB)] = \text{tr}\,[\wedge^k(A)\,\wedge^k(B)] \\ &= \text{tr}\,[\wedge^k(B)\,\wedge^k(A)] = \text{tr}\ \wedge^k(BA) = c_k(BA). \end{aligned}$$

Proof 3. This proof invokes a continuity argument that is useful in many contexts. Suppose A is invertible (nonsingular). Then $AB = A(BA)A^{-1}$. So AB and BA are similar, and hence have the same eigenvalues. Thus the equalities (3) are valid when A is invertible. Two facts are needed to get to the general case from here. (i) if A is singular, we can choose a sequence A_m of nonsingular matrices such that $A_m \to A$. (Singular matrices are characterised by the condition $\det(A) = 0$. Since det is a polynomial function in the entries of A, the set of its zeros is small. See also the discussion in *Resonance*, June 2000, page 43). (ii) The functions $c_k(T)$ are polynomials in the entries of T and hence, are continuous. So, if A is singular we choose a sequence A_m of nonsingular matrices converging to A and note

$$c_k(AB) = \lim_{m\to\infty} c_k(A_mB) = \lim_{m\to\infty} c_k(BA_m) = c_k(BA).$$

Proof 4. This proof uses 2×2 block matrices. Consider the $(2n) \times (2n)$ matrix $\begin{bmatrix} X & Z \\ O & Y \end{bmatrix}$ in which the four entries are $n \times n$ matrices, and O is the null matrix. The eigenvalues of this matrix are the n eigenvalues of X together with the eigenvalues of Y. (The determinant of this matrix is $\det(X)\det(Y)$.) Given any $n \times n$ matrix A, the $(2n) \times (2n)$ matrix $\begin{bmatrix} I & A \\ O & I \end{bmatrix}$ is invertible, and its inverse is $\begin{bmatrix} I & -A \\ O & I \end{bmatrix}$. Use this to see that

$$\begin{bmatrix} I & A \\ O & I \end{bmatrix}^{-1} \begin{bmatrix} AB & O \\ B & O \end{bmatrix} \begin{bmatrix} I & A \\ O & I \end{bmatrix} = \begin{bmatrix} O & O \\ B & BA \end{bmatrix}$$

Thus the matrices $\begin{bmatrix} AB & O \\ B & O \end{bmatrix}$ and $\begin{bmatrix} O & O \\ B & BA \end{bmatrix}$ are similar and hence, have the same eigenvalues. So, AB and BA have the same eigenvalues.

Proof 5. Another proof based on block matrices goes as follows. Let $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ be a block matrix. If A_{11} is nonsingular, then multiplying A on the right by $\begin{bmatrix} I & A_{11}^{-1}A_{12} \\ O & I \end{bmatrix}$ we get the matrix $\begin{bmatrix} A_{11} & O \\ A_{21} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$. Hence,

$$\det(A) = \det(A_{11})\det(A_{22} - A_{21}A_{11}^{-1}A_{12}).$$

[The matrix $A_{22} - A_{21}A_{11}^{-1}A_{12}$ is called the *Schur complement* of A_{11} in A. This determinant identity is one of the several places where it shows up.] In the same way, if A_{22} is invertible, then $\det(A) = \det(A_{22})\det(A_{11} - A_{12}A_{22}^{-1}A_{21})$. So, if A_{11} commutes with A_{21}, then $\det(A) = \det(A_{11}A_{22} - A_{21}A_{12})$; and if A_{22} commutes with A_{12}, then $\det(A) = \det(A_{22}A_{11} - A_{12}A_{21})$. Now let A, B be any two $n \times n$ matrices, and consider the block matrix $\begin{bmatrix} \lambda I & A \\ B & \lambda I \end{bmatrix}$. This is a very special kind of block matrix satisfying all conditions in the preceding lines. So $\det(\lambda^2 I - AB) = \det(\lambda^2 I - BA)$. This is true for all complex numbers λ. So, AB and BA have the same characteristic polynomial.

Proof 6. Let A be an idempotent matrix, i.e., $A^2 = A$. Then A represents a projection operator (not necessarily an orthogonal projection). So, in some basis (not necessarily orthonormal) A can be written as $A = \begin{bmatrix} I & O \\ O & O \end{bmatrix}$. In this basis let $B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$. Then $AB = \begin{bmatrix} B_{11} & B_{12} \\ O & O \end{bmatrix}$, $BA = \begin{bmatrix} B_{11} & O \\ B_{21} & O \end{bmatrix}$. So, AB and BA have the same eigenvalues. Now let A be any matrix. Then there exists an invertible matrix G such that $AGA = A$. (The two sides are equal as operators on the null space of A. On the complement of this space, A can be inverted. Set G to be the identity on the null space of A.) Note that GA is idempotent and apply the special case to GA and BG^{-1} in place of A and B. This shows $GABG^{-1}$ and $BG^{-1}GA$ have the same eigenvalues. In other words AB and BA have the same eigenvalues.

Proof 7. Since $\det AB = \det BA, 0$ is an eigenvalue of AB if and only if it is an eigenvalue of BA. Suppose a nonzero number λ is an eigenvalue of AB. Then there exists a (nonzero) vector v such that $ABv = \lambda v$. Applying B to the two sides of this equation we see that Bv is an eigenvector of BA corresponding to eigenvalue λ. Thus every eigenvalue of AB is an eigenvalue of BA. This argument gives no information about the (algebraic) multiplicities of the eigenvalues that the earlier six proofs did. However, following the same argument one sees that if $v_1, \ldots, v_k$ are linearly independent eigenvectors for AB corresponding to a nonzero eigenvalue λ, then $Bv_1, \ldots, Bv_k$ are linearly independent eigenvectors of BA corresponding to

the eigenvalue λ. Thus a nonzero eigenvalue of AB has the same *geometric multiplicity* as it has as an eigenvalue of BA. This may not be true for a zero eigenvalue. For example, if $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, then $AB = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $BA = O$. Both AB and BA have zero as their only eigenvalue. Its geometric multiplicity is one in the first case and two in the second case.

Proof 8. We want to show that a complex number z is an eigenvalue of AB if and only if it is an eigenvalue of BA. In other words, $(zI - AB)$ is invertible if and only if $(zI - BA)$ is invertible. This is certainly true if $z = 0$. If $z \neq 0$ we can divide A by z. So, we need to show that $(I - AB)$ is invertible if and only if $(I - BA)$ is invertible. Suppose $I - AB$ is invertible and let $X = (I - AB)^{-1}$. Then note that

$$\begin{aligned} (I - BA)(I + BXA) &= I - BA + BXA - BABXA \\ &= I - BA + B(I - AB)XA \\ &= I - BA + BA = I \end{aligned}$$

Thus $(I - BA)$ is invertible and its inverse is $I + BXA$.

This calculation seems mysterious. How did we guess that $I + BXA$ works as the inverse for $I - BA$? Here is a key to the mystery. Suppose a, b are numbers and $|ab| < 1$. Then

$$\begin{aligned} (1 - ab)^{-1} &= 1 + ab + abab + ababab + \cdots \\ (1 - ba)^{-1} &= 1 + ba + baba + bababa + \cdots \end{aligned}$$

If the first quantity is x, then the second one is $1 + bxa$. This suggests to us what to try in the matrix case.

This proof gives no information about multiplicities of eigenvalues — algebraic or geometric — since it does not involve either the characteristic polynomial or eigenvectors. This apparent weakness turns into a strength when we discuss operators on infinite dimensional spaces.

Let $\mathcal{H}$ be the Hilbert space l_2 consisting of sequences $x = (x_1, x_2, \ldots)$ for which $\sum_{j=1}^{\infty} \|x_j\|^2 < \infty$. Let A be a bounded linear operator on $\mathcal{H}$. The *spectrum* of $\sigma(A)$ is the complement of the set of all complex numbers λ such that $(A - \lambda I)^{-1}$ exists and is a bounded linear operator. The *point spectrum* of A is the set $\sigma_{\text{pt}}(A)$ consisting of all complex numbers λ for which there exists a nonzero vector v such that $Av = \lambda v$. In this case λ is called an eigenvalue of A and v an eigenvector. The set $\sigma(A)$ is a nonempty compact set while the set σ_{pt} can be empty. In other words, A need not have any eigenvalues, and if it does the spectrum may contain points other than the eigenvalues (Unlike in finite-dimensional vector spaces, a one-to-one linear operator need not be onto.)

Now let A, B be two bounded linear operators on $\mathcal{H}$. Proof 8 tells us that the sets $\sigma(AB)$ and $\sigma(BA)$ have the same elements with the possible exception of zero. Proof 7 tells us the same thing about $\sigma_{\text{pt}}(AB)$ and $\sigma_{\text{pt}}(BA)$. It also tells us that the geometric multiplicity of

each nonzero eigenvalue is the same for AB and BA. (There is no notion of determinant, characteristic polynomial and algebraic multiplicity in this case.)

The point zero can behave differently now. Let A, B be the operators that send the vector $(x_1, x_2, \ldots)$ to $(0, x_1, x_2, \ldots)$ and $(x_2, x_3, \ldots)$ respectively. Then BA is the identity operator while AB is the orthogonal projection onto the space spanned by vectors whose first coordinate is zero. Thus the sets $\sigma(AB)$ and $\sigma_{\text{pt}}(AB)$ consist of two points 0 and 1, while the corresponding sets for BA consist of the single point 1.

A final comment on rectangular matrices A, B. If both products AB and BA make sense, then the nonzero eigenvalues of AB and BA are the same. Which of the proofs shows this most clearly?

(*This is a corrected version of a note that appeared in Resonance, January 2002.*)

The Unexpected Appearance of Pi in Diverse Problems*

Rajendra Bhatia
Indian Statistical Institute, New Delhi 110 016, India.

There is a famous essay titled *The Unreasonable Effectiveness of Mathematics in the Natural Sciences* by the renowned physicist Eugene P Wigner. The essay opens with the paragraph:

There is a story about two friends, who were classmates in high school, talking about their jobs. One of them became a statistician and was working on population trends. He showed a reprint to his former classmate. The reprint started, as usual, with the Gaussian distribution and the statistician explained to his former classmate the meaning of the symbols for the actual population, for the average population, and so on. His classmate was a bit incredulous and was not quite sure whether the statistician was pulling his leg. "How can you know that?" was his query. "And what is this symbol here?" "Oh," said the statistician, "this is π" "What is that?" "The ratio of the circumference of the circle to its diameter." "Well, now you are pushing your joke too far," said the classmate, "surely the population has nothing to do with the circumference of the circle."

Wigner then goes on to discuss the surprisingly powerful role mathematics plays in the study of nature. I have quoted this para for making a small point. The number π, *the ratio of the circumference of the circle to its diameter*, appears in many contexts that seem to have no connection with diameters, areas, or volumes. One such problem that I discuss here concerns properties of natural numbers.

Every student of calculus learns the Wallis product formula

$$\frac{\pi}{2} = \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{5}\frac{6}{5}\frac{6}{7}\frac{8}{7}\frac{8}{9}\cdots \tag{1}$$

On the right hand side there is an infinite product and this is to be interpreted as

$$\lim_{n\to\infty} \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{3}\cdots\frac{2n}{2n-1}\frac{2n}{2n+1}. \tag{2}$$

This formula attributed to John Wallis (1616–1703) is remarkable for several reasons. It is, perhaps, the first occurence of an infinite product in mathematics. And it connects π with natural numbers. The formula has a simple proof. Let

$$I_n = \int_0^{\pi/2} (\sin x)^n dx.$$

Integrate by parts to get the recurrence formula

$$I_n = \frac{n-1}{n} I_{n-2}.$$

*Reproduced from *Resonance*, Vol. 8, No. 6, pp. 34–43, June 2003. (General Article)

The sequence I_n is a monotonically decreasing sequence of positive numbers. This and the recurrence formula show that

$$I < \frac{I_n}{I_{n+1}} < 1 + \frac{1}{n}.$$

So I_n/I_{n+1} tends to 1 as $n \to \infty$. Note that $I_0 = \pi/2$ and $I_1 = 1$. The recurrence formula can be used to get

$$\frac{I_{2n+1}}{I_{2n}} = \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{5}\cdots\frac{2n}{2n-1}\frac{2n}{2n+1}\frac{2}{\pi}.$$

Taking the limit as $n \to \infty$ we get (1).

Many *infinite sums* involving natural numbers lead to π. One that we need for our discussion is a famous formula due to Leonhard Euler (1707–1783)

$$\frac{\pi^2}{6} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots \tag{3}$$

A (natural) number is said to be *square-free* if in its prime factoring no factor occurs more than once. Thus $70 = 2 \times 5 \times 7$ is a square-free number while $12 = 2 \times 2 \times 3$ is not.

Many problems in number theory are questions about the distribution of various special kinds of numbers among all numbers. Thus we may ask:

What is the proportion of square-free numbers among all numbers?

Or

If a number is picked at random what is the probability that it is square-free?

Now, randomness is a tricky notion and this question needs more careful formulation. However, let us ignore that for the time being. It is reasonable to believe that if we pick a number at random it is as likely to be odd as it is even. This is because in the list

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, \ldots$$

every alternate number is even. In the same way every third number is a multiple of 3, every fourth number is a multiple of 4, and so on. Thus the probability that a randomly picked number is a multiple of k is $1/k$, and the probability that it is *not* a multiple of k is $1 - 1/k$:

Let $p_1, p_2, p_3 \ldots$ be the sequence of prime numbers. Let n be a randomly chosen number. For each prime p_j the probability that p_j^2 is not a factor of n is $1 - 1/p_j^2$: Given two primes p_j and p_k, what is the probability that neither p_j^2 nor p_k^2 is a factor of n? Again from probabilistic reasoning we know that the probability of the simultaneous occurence of two *independent* events is the product of their individual probabilities. (Thus the probability of getting two consecutive heads when a coin is tossed twice is 1/4.) Whether n has a factor p_j^2 has no bearing on its having p_k^2 as a factor. Thus the probability that neither p_j^2 nor p_k^2 is a factor of n is

$(1 - 1/p_j^2)\,(1 - 1/p_k^2)$. Extending this reasoning one sees that the probability of n being square free is the infinite product

$$\prod_{j=1}^{\infty}\left(1 - \frac{1}{p_j^2}\right) \tag{4}$$

There is a connection between this product and the series in (3). It is convenient to introduce here a famous object called the *Riemann zeta function*. This is defined by the series

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}. \tag{5}$$

This series surely converges for all real numbers $s > 1$. Let us restrict ourselves to these values of s, though the zeta function can be defined meaningfully for other complex numbers. The formula (3) can be written as

$$\zeta(2) = \frac{\pi^2}{6}. \tag{6}$$

The zeta function and prime numbers come together in the following theorem of Euler.

Theorem. *For all* $s > 1$

$$\zeta(s) = \prod_{n=1}^{\infty} \frac{1}{1 - p_n^{-s}}. \tag{7}$$

Proof. Fix an N, and use the geometric series expansion of $\frac{1}{1-x}$ to get

$$\prod_{n=1}^{N} \frac{1}{1 - p_n^{-s}} = \prod_{n=1}^{N} \sum_{m=0}^{\infty} p_n^{-ms} \tag{8}$$

The last expression is equal to

$$\sum_{j=1}^{\infty} \frac{1}{n_j^s},$$

where $n_1, n_2 \ldots$ is an enumeration of those numbers that have $p_1, p_2, \ldots, p_N$ as their only prime factors. As $n \to \infty$, the sequence $\{n_j\}$ expands to include all natural numbers. This proves the theorem.

As a consequence the product (4) has the value $6/\pi^2$. This is the probability that a number picked at random is square-free.

This is one more situation where the number π has made an appearance quite unexpectedly. Our main point has been made; several interesting side-lines remain.

First note that our argument shows that if we pick a number n at random, then the probability that it has no prime factor with multiplicity k is $1/\zeta(k)$.

With a little thinking one can see that the *probability that two numbers picked at random are coprime is* $6/\pi^2$. (This problem is equivalent to the one we have been discussing.)

There is another interesting way of looking at this problem. Let $\mathbb{Z}^2$ be the collection of all points in the plane whose coordinates are integers. This is called the *integer lattice*. If the line segment joining the origin $(0,0)$ to a point (m,n) does not pass through any other lattice point we say that the point (m,n) can be seen from the origin. For example, the point $(1,-1)$ can be seen from the origin but the point $(2,-2)$ can not be seen. Among all lattice points what is the proportion of those that can be seen from the origin? The answer, again, is $6/\pi^2$. The proof of this is left to the reader.

The argument used in proving the Theorem above can be modified to give a proof of the fact that there are infinitely many prime numbers. The probability that a randomly picked number from the set $\{1,2,\dots,N\}$ is 1 goes to zero as N becomes large. So the product $\Pi_p(1-1/p)$ where p varies over all primes is smaller than any positive number. This would not be possible if there were only a finitely many factors in the product.

The number π entered the picture via the formula (3). How does one prove it? Several proofs are known. The daring 'proof' first given by Euler goes as follows.

Let $\alpha_1, \alpha_2, \dots$ be the roots of the polynomial equation $a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m = 0$. Then

$$\sum \frac{1}{\alpha_i} = \frac{-a_1}{a_0}.$$

We can write

$$\cos\sqrt{x} = 1 - \frac{x}{2} + \frac{x^2}{24} + \cdots .$$

This is a 'polynomial of infinite degree', and the roots of $\cos\sqrt{x} = 0$ are

$$\frac{(2n+1)^2\pi^2}{4}, \quad n = 0, 1, 2, \dots$$

Hence,

$$\sum_{n=0}^{\infty} \frac{1}{(2n+1)^2} = \frac{\pi^2}{8}. \tag{9}$$

The formula (3) follows from this easily.

Surely this argument has flaws. They can all be removed! With the notions of uniform convergence and $\epsilon - \delta$ arguments, we can prove formulas like

$$\frac{\sin x}{x} = \prod_{n=1}^{\infty} \left(1 - \frac{x^2}{n^2 x^2}\right), \tag{10}$$

from which the formulas (1) and (3) can be derived by simple manipulations. Finding the sum of the series (3) was one of the early major triumphs of Euler. He was aware that the argument we have described above is open to several criticisms. So he gave another proof that goes as follows.

$$\frac{\pi^2}{8} \quad = \quad \frac{(\arcsin 1)^2}{2} = \int_0^1 \frac{\arcsin x}{\sqrt{1-x^2}} dx$$

$$
\begin{aligned}
&= \int_0^1 \frac{1}{\sqrt{1-x^2}} \left[x + \sum_{n=1}^{\infty} \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n} \frac{x^{2n+1}}{2n+1} \right] dx \\
&= 1 + \sum_{n=1}^{\infty} \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n(2n+1)} \frac{2n(2n-2) \cdots 2}{(2n+1)(2n-1) \cdots 3} \\
&= \sum_{n=0}^{\infty} \frac{1}{(2n+1)^2}.
\end{aligned}
$$

Following the ideas of his *first* proof Euler showed that $\zeta(2m)$ is π^{2m} multiplied by a rational number. Thus for example,

$$\zeta(4) = \frac{\pi^4}{90}, \quad \zeta(6) = \frac{\pi^4}{945}. \tag{11}$$

Neither Euler, nor anyone else in three centuries after him, has found much about the values of $\zeta(k)$ when k is an odd integer. In 1978 R Apéry showed that $\zeta(3)$ is an irrational number. Even this much is not known about $\zeta(5)$.

Another general method for finding sums like (3) and (11) goes via Fourier series. If f is a continuous function on $[-\pi, \pi]$ and $f(x) = \sum_{n=-\infty}^{\infty} a_n e^{inx}$ its Fourier expansion, then

$$\sum_{n=-\infty}^{\infty} |a_n|^2 = \int_{-\pi}^{\pi} |f(x)|^2 dx. \tag{12}$$

The method depends on recognising the summands of a particular series as coefficient of the Fourier series of a particular function f and then computing the integral in (12).

Having seen expression like (10) and (12) one is no longer surprised that $\zeta(2m)$ involves π in some way.

Finally, let us briefly discuss some issues related to 'picking a natural number at random'.

Two standard examples of completely random phenomena are tossing of a coin and throwing of a dice. In the first case we have two, and in the second case six, equally likely outcomes. The 'sample space' in the first case is the set $\{1, 2\}$ (representing the two outcomes head and tail) and in the second case it is the set $\{1, 2, \ldots, 6\}$. One can imagine an experiment with N equally likely outcomes $\{1, 2, \ldots, N\}$.

The *uniform probability distribution* on the set $X = \{1, 2, \ldots, N\}$ is the function that assigns to each subset E of X values according to the following rules

$$\mu(\{j\}) = \mu(\{k\}) \quad \text{for all} \quad j, k, \tag{13}$$

$$\mu(E) = \sum_{j \in E} \mu(\{j\}), \tag{14}$$

$$\mu(X) = 1. \tag{15}$$

Note that these three conditions imply that $\mu(\{j\}) = 1/N$ for all j. This is a model for a random phenomenon (like in some games of chance) with N equally likely outcomes.

It is clear that if X is replaced by the set $\mathbb{N}$ of all natural numbers, then no function satisfying the three conditions (13)–(15) exists. So, if 'picking an element of $\mathbb{N}$ at random' means assigning each of its elements j an equal 'probability' we run into a problem. However, there is a way to get around this.

Let $X = \{1, 2, \ldots, N\}$ and let E be the set of even numbers in X. If N is even, then $\mu(E) = 1/2$. But if $N = 2m + 1$ is odd, then $\mu(E) = m/(2m + 1)$. This is less than $1/2$, but gets very close to $1/2$ for large N. In this sense a number picked at random is as likely to be even as odd.

In the same spirit we can prove the following.

For every $\varepsilon > 0$, there exists a number N, such that if μ is the uniform probability distribution on the set $X = \{1, 2, \ldots, N\}$ and E is the set of square-free numbers in X, then

$$\frac{6}{\pi^2} < \mu(E) < \frac{6}{\pi^2} + \varepsilon.$$

The reader may prove this using the following observations. We know that

$$\prod_{j=1}^{\infty} \left(1 - \frac{1}{p_j^2}\right) = \frac{6}{\pi^2}.$$

The factors in this product are smaller than 1. So, the sequence

$$\prod_{j=1}^{M} \left(1 - \frac{1}{p_j^2}\right), \quad M = 1, 2, \ldots$$

decreases to its limit. Choose an M such that

$$\frac{6}{\pi^2} < \prod_{j=1}^{M} \left(1 - \frac{1}{p_j^2}\right) < \frac{6}{\pi^2} + \varepsilon$$

and let $N = \prod_{j=1}^{M} p_j^2$.

A (non-uniform) *probability distribution* on X is a function μ that satisfies the conditions (14)–(15) but not (necessarily) the condition (13). There is nothing that prevents the existence of such a distribution on $\mathbb{N}$. Any series with non-negative terms and with sum 1 gives such a distribution. In particular if we set

$$\mu(\{j\}) = \frac{6}{\pi^2} \frac{1}{j^2}, \quad j = 1, 2, \ldots, \tag{16}$$

then μ is a probability distribution on $\mathbb{N}$. This assigns different probabilities to different elements of $\mathbb{N}$. The reader may like to interpret and prove the following statement.

The probability that two natural numbers picked at random have j as their greatest common divisor is $\mu(\{j\})$ as defined by (16).

Suggested Reading

[1] G H Hardy and E M Wright, *An Introduction to the Theory of Numbers*, Oxford University Press, 1959. See Chapter VIII, and in particular Theorems 332 and 333. The latter theorem attributed to Gegenbauer (1885) says that if $Q(x)$ is the number of square-free numbers not exceeding x, then

$$Q(x) = \frac{6x}{\pi^2} + O(\sqrt{x}).$$

Here $O(\sqrt{x})$ represents a function whose absolute value is bounded by $A\sqrt{x}$ for some constant A.

Use this formula, with a computer program for testing whether a number is square-free, to obtain the value of π up to the third decimal place.

[2] P J Davis and R Hersch, *The Mathematical Experience*, Birkhauser, 1981. We have borrowed our main argument from the discussion on page 366 here. This occurs in a chapter titled *The Riemann Hypothesis* where the authors present an argument showing that this most famous open problem in mathematics has an affirmative solution with probability one.

[3] M Kac, *Enigmas of Chance*, University of California Press, 1987. See Chapters 3,4, and in particular pages 89–91 of this beautiful autobiography for deeper connections between number theory and probability found by its author. See also his book *Statistical Independence in Probability, Analysis and Number Theory*, Mathematical Association of America, 1959.

[4] W Dunham, *Journey Through Genuis*, Penguin Books, 1991. See Chapter 9 titled *The Extraordinary Sums of Leonhard Euler* for an entertaining history of the formula (3).

[5] R Bhatia, *Fourier Series*, Hindustan Book Agency, Second Edition 2003. See Chapter 3 for several series and products that lead to π.

The Logarithmic Mean*

Rajendra Bhatia
Indian Statistical Institute, New Delhi 110 016, India.

The inequality between the arithmetic mean (AM) and geometric mean (GM) of two positive numbers is well known. This article introduces the logarithmic mean, shows how it leads to refinements of the AM–GM inequality. Some applications and properties of this mean are shown. Some other means and related inequalities are discussed.

One of the best known and most used inequalities in mathematics is the inequality between the harmonic, geometric, and arithmetic means. If a and b are positive numbers, these means are defined, respectively, as

$$H(a,b) = \left(\frac{a^{-1}+b^{-1}}{2}\right)^{-1}, \quad G(a,b) = \sqrt{ab}, \quad A(a,b) = \frac{a+b}{2}, \tag{1}$$

and the inequality says that

$$H(a,b) \le G(a,b) \le A(a,b). \tag{2}$$

Means other than the three "classical" ones defined in (1) are used in different problems. For example, the *root mean square*

$$B_2(a,b) = \left(\frac{a^2+b^2}{2}\right)^{1/2}, \tag{3}$$

is often used in various contexts. Following the mathematician's penchant for generalisation, the four means mentioned above can be subsumed in the family

$$B_p(a,b) = \left(\frac{a^p+b^p}{2}\right)^{1/p}, \quad -\infty < p < \infty, \tag{4}$$

variously known as *binomial means, power means,* or *Hölder means*. When $p = -1, 1$, and 2, respectively, $B_p(a,b)$ is the harmonic mean, the arithmetic mean, and the root mean square. If we understand $B_0(a,b)$ to mean

$$B_0(a,b) = \lim_{p\to 0} B_p(a,b),$$

then

$$B_0(a,b) = G(a,b). \tag{5}$$

*Reproduced from *Resonance*, Vol. 13, No. 6, pp. 583–594, June 2008. (General Article)

In a similar vein we can see that

$$B_\infty(a,b) := \lim_{p\to\infty}\left(\frac{a^p+b^p}{2}\right)^{1/p} = \max(a,b),$$
$$B_{-\infty}(a,b) := \lim_{p\to-\infty}\left(\frac{a^p+b^p}{2}\right)^{1/p} = \min(a,b).$$

A little calculation shows that

$$B_p(a,b) \le B_q(a,b) \qquad \text{if} \quad p \le q. \tag{6}$$

This is a strong generalization of the inequality (2). We may say that for $-1 \le p \le 1$ the family B_p *interpolates* between the three means in (1) as does the inequality (6) with respect to (2).

A substantial part of the mathematics classic *Inequalities* by G Hardy, J E Littlewood and G Pölya is devoted to the study of these means and their applications. The book has had quite a few successors, and yet new properties of these means continue to be discovered.

The purpose of this article is to introduce the reader to the *logarithmic mean,* some of its applications, and some very pretty mathematics around it.

The logarithmic mean of two positive numbers a and b is the number $L(a,b)$ defined as

$$L(a,b) = \frac{a-b}{\log a - \log b} \quad \text{for} \quad a \ne b, \tag{7}$$

with the understanding that

$$L(a,a) = \lim_{b\to a} L(a,b) = a.$$

There are other interesting representations for this object, and the reader should check the validity of these formulas:

$$L(a,b) = \int_0^1 a^t b^{1-t}dt, \tag{8}$$
$$\frac{1}{L(a,b)} = \int_0^1 \frac{dt}{ta+(1-t)b}, \tag{9}$$
$$\frac{1}{L(a,b)} = \int_0^\infty \frac{dt}{(t+a)(t+b)}. \tag{10}$$

The logarithmic mean always falls between the geometric and the arithmetic means; i.e.,

$$G(a,b) \le L(a,b) \le A(a,b). \tag{11}$$

We indicate three different proofs of this and invite the reader to find more.

When $a = b$, all the three means in (11) are equal to a. Suppose $a > b$, and put $w = a/b$. The first inequality in (11) is equivalent to saying

$$\sqrt{w} \le \frac{w-1}{\log w} \qquad \text{for} \quad w > 1.$$

Replacing w by u^2, this is the same as saying

$$2 \log u \;\leq\; \frac{u^2-1}{u} \quad \text{for} \quad u > 1. \tag{12}$$

The two functions $f(u) = 2 \log u$, and $g(u) = (u^2 - 1)/u$ are equal to 0 at $u = 1$, and a small calculation shows that $f'(u) < g'(u)$ for $u > 1$. This proves the desired inequality (12), and with it the first inequality in (11). In the same way, the second of the inequalities (11) can be reduced to

$$\frac{u-1}{u+1} \leq \frac{\log u}{2} \quad \text{for} \quad u \geq 1.$$

and proved by calculating derivatives.

A second proof goes as follows. Two applications of the arithmetic-geometric mean inequality show that

$$t^2 + 2t\sqrt{ab} + ab \;\leq\; t^2 + t(a+b) + ab \;\leq\; t^2 + t(a+b) + \left(\frac{a+b}{2}\right)^2$$

for all $t \geq 0$. Using this, one finds that

$$\int_0^\infty \frac{dt}{(t+\frac{a+b}{2})^2} \leq \int_0^\infty \frac{dt}{(t+a)(t+b)} \leq \int_0^\infty \frac{dt}{(t+\sqrt{ab})^2}.$$

Evaluation of the integrals shows that this is the same as the assertion in (11).

Since a and b are positive, we can find real numbers x and y such that $a = e^x$ and $b = e^y$. Then the first inequality in (11) is equivalent to the statement

$$e^{(x+y)/2} \leq \frac{e^x - e^y}{x-y},$$

or

$$1 \leq \frac{e^{(x-y)/2} - e^{(y-x)/2}}{x-y}.$$

This can be expressed also as

$$1 \leq \frac{\sinh{(x-y)/2}}{(x-y)/2}.$$

In this form we recognise it as one of the fundamental inequalities of analysis: $t \leq \sinh t$ for all $t \geq 0$. Very similar calculations show that the second inequality in (11) can be reduced to the familiar fact $\tanh t \leq t$ for all $t \geq 0$.

Each of our three proofs shows that if $a \neq b$, then $G(a,b) < L(a,b) < A(a,b)$. One of the reasons for the interest in (11) is that it provides a refinement of the fundamental inequality between the geometric and the arithmetic means.

The logarithmic mean plays an important role in the study of conduction of heat in liquids flowing in pipes. Let us explain this briefly. The flow of heat by steady unidirectional conduction is governed by *Newton's law of cooling*: if q is the rate of heat flow along the x-axis across an area A normal to this axis, then

$$q = k\,A\,\frac{dT}{dx}, \tag{13}$$

where dT/dx is the temperature gradient along the x direction and k is a constant called the thermal conductivity of the material. (See, for example, R Bhatia, *Fourier Series*, Mathematical Association of America, 2004, p.2). The cross-sectional area A may be constant, as for example in a cube. More often (as in the case of a fluid travelling in a pipe) the area A is a variable. In engineering calculations, it is then more convenient to replace (13) by

$$q = k\,A_m\,\frac{\triangle T}{\triangle x}, \tag{14}$$

where $\triangle T$ is the difference of temperatures at two points at distance $\triangle x$ along the x-axis, and A_m is the *mean cross section* of the body between these two points. For example, if the body has a uniformly tapering rectangular cross section, then A_m is the arithmetic mean of the two boundary areas A_1 and A_2.

Consider, heat flow in a long hollow cylinder where end effects are negligible. Then the heat flow can be taken to be essentially radial. (see, for example, J Crank: *The Mathematics of Diffusion*, Clarendon Press, 1975.) The cross sectional area in this case is proportional to the distance from the centre of the pipe. If L is the length of the pipe, the area of the cylindrical surface at distance x from the axis is $2\pi xL$. So, the total heat flow q across the section of the pipe bounded by two coaxial cylinders at distance x_1 and x_2 from the axis, using (13), is seen to satisfy the equation

$$q\int_{x_1}^{x_2}\frac{dx}{2\pi xL} = k\triangle T, \tag{15}$$

or,

$$q = \frac{k\,2\pi L\ \triangle T}{\log\ x_2 - \log\ x_1}.$$

If we wish to write this in the form (14) with $x_2 - x_1 = \triangle x$, then we must have

$$A_m = 2\pi L\,\frac{x_2 - x_1}{\log\ x_2 - \log\ x_1} = \frac{2\pi Lx_2 - 2\pi Lx_1}{\log\ 2\pi Lx_2 - \log\ 2\pi Lx_1}.$$

In other words,

$$A_m = \frac{A_2 - A_1}{\log\ A_2 - \log\ A_1},$$

the logarithmic mean of the two areas bounding the cylindrical section under consideration. In the engineering literature this is called the *logarithmic mean area.*

If instead of two coaxial cylinders we consider two concentric spheres, then the cross sectional area is proportional to the square of the distance from the centre. In this case we have, instead of (15),

$$q \int_{x_1}^{x_2} \frac{dx}{4\pi x^2} = k\triangle T.$$

A small calculation shows that in this case

$$A_m = \sqrt{A_1 A_2},$$

the geometric mean of the two areas bounding the annular section under consideration.

Thus the geometric and the logarithmic means are useful in calculations related to heat flow through spherical and cylindrical bodies, respectively. The latter relates to the more common phenomenon of flow through pipes.

Let us return to inequalities related to the logarithmic mean. Let t be any nonzero real number. In the equality (11) replace a and b by a^t and b^t, respectively. This gives

$$(ab)^{t/2} \leq \frac{a^t - b^t}{t(\log a - \log b)} \leq \frac{a^t + b^t}{2},$$

from which we get

$$t(ab)^{t/2} \frac{a-b}{a^t - b^t} \leq \frac{a-b}{\log a - \log b} \leq t\frac{a^t + b^t}{2} \frac{a-b}{a^t - b^t}.$$

The middle term in this equality is the logarithmic mean. Let G_t and A_t be defined as

$$\begin{aligned} G_t(a,b) &= t(ab)^{t/2} \frac{a-b}{a^t - b^t}, \\ A_t(a,b) &= t\frac{a^t + b^t}{2} \frac{a-b}{a^t - b^t}. \end{aligned}$$

We have assumed in these definitions that $t \neq 0$. If we define G_0 and A_0 as the limits

$$\begin{aligned} G_0(a,b) &= \lim_{t\to 0} G_t(a,b), \\ A_0(a,b) &= \lim_{t\to 0} A_t(a,b), \end{aligned}$$

then

$$G_0(a,b) = A_0(a,b) = L(a,b).$$

The reader can verify that

$$\begin{aligned} G_1(a,b) &= \sqrt{ab}, \quad A_1(a,b) = \frac{a+b}{2}, \\ G_{-t}(a,b) &= G_t(a,b), \quad A_{-t}(a,b) = A_t(a,b). \end{aligned}$$

For fixed a and b, $G_t(a,b)$ is a decreasing function of $|t|$, while $A_t(a,b)$ is an increasing function of $|t|$. (One proof of this can be obtained by making the substitution $a = e^x$, $b = e^y$.) The last inequality obtained above can be expressed as

$$G_t(a,b) \le L(a,b) \le A_t(a,b), \tag{16}$$

for all t. Thus we have an infinite family of inequalities that includes the arithmetic-geometric mean inequality, and other interesting inequalities. For example, choosing $t = 1$ and $1/2$, we see from the information obtained above that

$$\sqrt{ab} \;\le\; \frac{a^{3/4}b^{1/4} + a^{1/4}b^{3/4}}{2} \;\le\; L(a,b) \;\le \left(\frac{a^{1/2} + b^{1/2}}{2}\right)^2 \;\le\; \frac{a+b}{2}. \tag{17}$$

This is a refinement of the fundamental inequality (11). The second term on the right is the binomial mean $B_{1/2}(a,b)$. The second term on the left is one of another family of means called *Heinz means* defined as

$$H_\nu(a,b) = \frac{a^\nu b^{1-\nu} + a^{1-\nu}b^\nu}{2}, \quad 0 \le \nu \le 1. \tag{18}$$

Clearly

$$\begin{aligned} H_0(a,b) &= H_1(a,b) = \frac{a+b}{2}, \\ H_{1/2}(a,b) &= \sqrt{ab}, \\ H_{1-\nu}(a,b) &= H_\nu(a,b). \end{aligned}$$

Thus the family H_ν is yet another family that interpolates between the arithmetic and the geometric means. The reader can check that

$$H_{1/2}(a,b) \;\le\; H_\nu(a,b) \;\le\; H_0(a,b), \tag{19}$$

for $0 \le \nu \le 1$. This is another refinement of the arithmetic-geometric mean inequality.

If we choose $t = 2^{-n}$, for any natural number n, then we get from the first inequality in (16)

$$2^{-n}(ab)^{2^{-(n+1)}} \frac{a-b}{a^{2^{-n}} - b^{2^{-n}}} \;\le\; L(a,b).$$

Using the identity

$$a - b = \left(a^{2^{-n}} - b^{2^{-n}}\right)\left(a^{2^{-n}} + b^{2^{-n}}\right)\left(a^{2^{-n+1}} + b^{2^{-n+1}}\right)\cdots\left(a^{2^{-1}} + b^{2^{-1}}\right),$$

we get from the inequality above

$$(ab)^{2^{-(n+1)}} \prod_{m=1}^{n} \frac{a^{2^{-m}} + b^{2^{-m}}}{2} \;\le\; L(a,b). \tag{20}$$

Similarly, from the second inequality in (16) we get

$$L(a,b) \le \frac{a^{2^{-n}} + b^{2^{-n}}}{2} \prod_{m=1}^{n} \frac{a^{2^{-m}} + b^{2^{-m}}}{2}. \tag{21}$$

If we let $n \to \infty$ in the two formulas above, we obtain a beautiful product formula:

$$L(a,b) = \prod_{m=1}^{\infty} \frac{a^{2^{-m}} + b^{2^{-m}}}{2}. \tag{22}$$

This adds to our list of formulas (7)–(10) for the logarithmic mean.

Choosing $b = 1$ in (22) we get after a little manipulation the representation for the logarithm function

$$\log\, x = (x-1) \prod_{m=1}^{\infty} \frac{2}{1 + x^{2^{-m}}}, \tag{23}$$

for all $x > 0$.

We can turn this argument around. For all $x > 0$ we have

$$\log\, x = \lim_{n\to\infty} n\left(x^{1/n} - 1\right). \tag{24}$$

Replacing n by 2^n, a small calculation leads to (23) from (24). From this we can obtain (22) by another little calculation.

There are more analytical delights in store; the logarithmic mean even has a connection with the fabled Gauss arithmetic-geometric mean that arises in a totally different context. Given positive numbers a and b, inductively define two sequences as

$$\begin{aligned} a_0 &= a, & b_0 &= b \\ a_{n+1} &= \frac{a_n + b_n}{2}, & b_{n+1} &= \sqrt{a_n b_n}. \end{aligned}$$

Then $\{a_n\}$ is a decreasing, and $\{b_n\}$ an increasing, sequence. All a_n and b_n are between a and b. So both sequences converge. With a little work one can see that $a_{n+1} - b_{n+1} \le \frac{1}{2}(a_n - b_n)$, and hence the sequences $\{a_n\}$ and $\{b_n\}$ converge to a common limit. The limit $AG(a,b)$ is called the Gauss arithmetic-geometric mean. Gauss showed that

$$\begin{aligned} \frac{1}{AG(a,b)} &= \frac{2}{\pi} \int_0^{\infty} \frac{dx}{\sqrt{(a^2 + x^2)(b^2 + x^2)}} \\ &= \frac{2}{\pi} \int_0^{\pi/2} \frac{d\varphi}{\sqrt{a^2 \cos^2 \varphi + b^2 \sin^2 \varphi}}. \end{aligned} \tag{25}$$

These integrals called "elliptic integrals" are difficult ones to evaluate, and the formula above relates them to the mean value $AG(a,b)$. Clearly

$$G(a,b) \le AG(a,b) \le A(a,b). \tag{26}$$

Somewhat unexpectedly, the mean $L(a,b)$ can also be realised as the outcome of an iteration closely related to the Gauss iteration. Let A_t and G_t be the two families defined earlier. A small calculation, that we leave to the reader, shows that

$$\frac{A_t + G_t}{2} = A_{t/2}, \qquad \sqrt{A_{t/2}G_t} = G_{t/2}. \tag{27}$$

For $n = 1, 2, \ldots$, let $t = 2^{1-n}$, and define two sequences a'_n and b'_n as $a'_n = A_t$, $b'_n = G_t$; i.e.,

$$\begin{aligned} a'_1 &= A_1 = \frac{a+b}{2}, \qquad b'_1 = G_1 = \sqrt{ab}, \\ a'_2 &= A_{1/2} = \frac{a'_1 + b'_1}{2}, \qquad b'_2 = G_{1/2} = \sqrt{A_{1/2}G_1} = \sqrt{a'_2 b'_1}, \\ &\vdots \\ a'_{n+1} &= \frac{a'_n + b'_n}{2}, \qquad b'_{n+1} = \sqrt{a'_{n+1}b'_n}. \end{aligned}$$

We leave it to the reader to show that the two sequences $\{a'_n\}$ and $\{b'_n\}$ converge to a common limit, and that limit is equal to $L(a,b)$. This gives one more characterisation of the logarithmic mean. These considerations also bring home another interesting inequality

$$L(a,b) \le AG(a,b). \tag{28}$$

Finally, we indicate yet another use that has recently been found for the inequality (11) in differential geometry. Let $\|T\|_2$ be the Euclidean norm on the space of $n \times n$ complex matrices; i.e.

$$\|T\|_2^2 = \operatorname{tr} T^*T = \sum_{i,j=1}^{n} \left|t_{ij}\right|^2.$$

A matrix version of the inequality (11) says that for all positive definite matrices A and B and for all matrices X, we have

$$\left\|A^{1/2}XB^{1/2}\right\|_2 \le \left\|\int_0^1 A^t X B^{1-t}dt\right\|_2 \le \left\|\frac{AX+XB}{2}\right\|_2. \tag{29}$$

The space $\mathbb{H}_n$ of all $n \times n$ Hermitian matrices is a real vector space, and the exponential function maps this onto the space $\mathbb{P}_n$ consisting of all positive definite matrices. The latter is a Riemannian manifold. Let $\delta_2(A,B)$ be the natural Riemannian metric on $\mathbb{P}_n$. A very fundamental inequality called the *exponential metric increasing property* says that for all Hermitian matrices H and K

$$\delta_2\left(e^H, e^K\right) \ge \|H - K\|_2. \tag{30}$$

A short and simple proof of this can be based on the first of the inequalities in (29). The inequality (30) captures the important fact that the manifold $\mathbb{P}_n$ has nonpositive curvature. For more details see the Suggested Reading.

Suggested Reading

[1] G Hardy, J E Littlewood and G Pölya, *Inequalities*, Cambridge University Press, Second edition, 1952. (This is a well-known classic. Chapters II and III are devoted to "mean values".)

[2] P S Bullen, D S Mitrinovic, and P M Vasic, *Means and Their Inequalities*, D Reidel, 1998. (A specialised monograph devoted exclusively to various means.)

[3] W H McAdams, *Heat Transmission*, Third edition, McGraw Hill, 1954. (An engineering text in which the logarithmic mean is introduced in the context of fluid flow.)

[4] B C Carlson, The logarithmic mean, *American Mathematical Monthly*, Vol. 79, pp. 615–618, 1972. (A very interesting article from which we have taken some of the material presented here.)

[5] R Bhatia, Positive Definite Matrices, *Princeton Series in Applied Mathematics*, 2007, and also TRIM 44, Hindustan Book Agency, 2007. (Matrix versions of means, and inequalities for them, can be found here. The role of the logarithmic mean in this context is especially emphasized in Chapters 4–6.)

[6] R Bhatia and J Holbrook, Noncommutative geometric means, *Mathematical Intelligencer*, 28 (2006) 32–39. (A quick introduction to some problems related to matrix means, and to the differential geometric context in which they can be placed.)

[7] Tung-Po Lin, The Power Mean and the Logarithmic Mean, *American Mathematical Monthly*, Vol. 81, pp. 879–883, 1974.

[8] S. Chakraborty, A Short Note on the Versatile Power Mean, *Resonance*, Vol. 12, No. 9, pp. 76–79, September 2007.

Convolutions

Rajendra Bhatia
Indian Statistical Institute, New Delhi 110 016, India

I am expected to tell you, in 25 minutes, something that should interest you, excite you, pique your curiosity, and make you look for more. It is a tall order, but I will try. The word "interactive" is in fashion these days. So I will leave a few things for you to check.

Let f_1 and f_2 be two polynomials, say

$$f_1(x) = a_0 + a_1x + a_2x^2, \tag{1}$$

$$f_2(x) = b_0 + b_1x + b_2x^2 + b_3x^3. \tag{2}$$

(Here the coefficients a's and b's could be integers, rational, real, or complex numbers.) Their product f_1f_2 is the polynomial

$$\begin{aligned} f_1f_2(x) = {} & a_0b_0 + (a_0b_1 + a_1b_0)x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 \\ & + (a_0b_3 + a_1b_2 + a_2b_1)x^3 + (a_1b_3 + a_2b_2)x^4 \\ & + a_2b_3x^5. \end{aligned} \tag{3}$$

What pattern do you see in the coefficients of the product f_1f_2?

Let us consider the general situation. Suppose f_1 and f_2 are polynomials of degrees m and n, respectively:

$$f_1(x) = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m, \tag{4}$$

$$f_2(x) = b_0 + b_1x + b_2x^2 + \cdots + b_nx^n. \tag{5}$$

Their product f_1f_2 is a polynomial of degree $m + n$, and has the expression

$$f_1f_2(x) = c_0 + c_1x + c_2x^2 + \cdots + c_{n+m}x^{n+m}. \tag{6}$$

What is the "formula" for the coefficients c's in terms of the a's and b's? You can see that c_k is the sum of all a_jb_ℓ, where $j + \ell = k$. This can be written briefly as

$$c_k = \sum_{j+\ell=k} a_jb_\ell, \tag{7}$$

or as

$$c_k = \sum_{j=0}^{k} a_jb_{k-j}. \tag{8}$$

A little care is needed in interpreting the meaning of this formula. The indices k vary from 0 to $n + m$ but the j's do not go beyond m. So, what is the meaning of the summation in (8)

with j going up to k when k is bigger than m? If we agree to put $a_{m+1}, a_{m+2}, \ldots, a_{m+n}$, and $a_{n+1}, b_{n+2}, \ldots, b_{m+n}$ all equal to zero, then (8) is meaningful. This is a helpful device.

Let C_{00} be the collection of all sequences with only finitely many nonzero terms. Thus a typical element of C_{00} is a sequence

$$a = (a_0, a_1, \ldots, a_m, 0, 0, 0, \ldots). \tag{9}$$

If

$$b = (b_0, b_1, \ldots, b_n, 0, 0, 0, \ldots) \tag{10}$$

is another such sequence, then we define the *convolution* of a and b to be the sequence

$$c = (c_0, c_1, \ldots, c_{m+n}, 0, 0, 0, \ldots), \tag{11}$$

whose terms c_k are given by (8). We write this relation between a, b and c as $c = a * b$.

Let $\mathcal{P}$ be the collection of all polynomials (of any degree). Each polynomial is determined by its coefficients (i.e., there is exactly one polynomial $f_a(x)$ whose coefficients are $a = (a_0, a_1, \ldots, a_m)$. As I explained, it is convenient to think of this as the sequence $(a_0, a_1, \ldots, a_m, 0, 0, 0, \ldots)$. If we have two polynomials f_a and f_b of degree m and n, respectively, then their sum is a polynomial whose degree is $\max(m, n)$. The coefficients of this polynomial are the terms of the sequence

$$a + b = (a_0 + b_0, a_1 + b_1, \ldots).$$

The product $f_a f_b$ is a polynomial of degree $m+n$. Call this polynomial f_c. Then the coefficients of f_c are c_k where $c = a * b$.

You have learnt about binary operations. The operations $*$ is a binary operation on the set C_{00}. Here are some questions. Is this operation commutative? Is it associative? Does there exist an identity element for this operation? i.e., is there a sequence e in C_{00} such that $a * e = a$ for all a? If such an e exists, then we ask further whether every element a of C_{00} has an inverse; i.e., does there exist a sequence a' such that $a * a' = e$?

Let $s(a) = a_0 + a_1 + \cdots + a_m$, be the sum of the coefficients in (4), and define $s(b)$ and $s(c)$ in the same way. You can see that

$$s(c) = s(a)\ s(b). \tag{12}$$

(Please do the calculations!)

The idea of convolution occurs at several places. One of them is in the calculation of probabilities. Let $(a_1, \ldots, a_n)$ be nonnegative real numbers such that $a_1 + \cdots + a_n = 1$. Then $a = (a_1, \ldots, a_n)$ is called a "probability vector". (Think of an experiment with n possible outcomes with probabilities $a_1, \ldots, a_n$.) If a and b are two probability vectors, then their convolution $c = a * b$ is another probability vector. (Use the relation (12) to see this.) What is the meaning of this?

Think of a simple game of chance like throwing a dice. There are six possible outcomes, $1, 2, \ldots, 6$, each with probability $1/6$. The probability vector (or the probability distribution)

corresponding to this is $(1/6, 1/6, \ldots, 1/6)$, which for brevity I write as $\frac{1}{6}(1, 1, \ldots, 1)$. Suppose we throw the dice twice and observe the sum of the two numbers that turn up. The possible values for the sum are the numbers between 2 and 12. But they occur with different probabilities. The numbers 2 and 12 can occur in only one way: both the throws should result in 1, or both should result in 6. On the other hand the sum can be 5 in four different ways

$$5 = 1 + 4 = 2 + 3 = 3 + 2 = 4 + 1.$$

Thus the probability of the sum being 2 is $1/36$ while its being 5 is $4/36$. Let me write (a, b) to mean that in the first throw of the dice the number a showed up, and in the second b. Let $s = a + b$. Then the familiar laws of probability say that

$$\text{Prob}(s = 5) = \text{Prob}(1, 4) + \text{Prob}(2, 3) + \text{Prob}(3, 2) + \text{Prob}(4, 1).$$

That is because the probabilities add up when the events are mutually exclusive. The outcomes of the two throws are independent, and probabilities multiply when the events are independent. So we have

$$\begin{aligned}\text{Prob}(s = 5) &= \text{Prob}(1)\text{Prob}(4) + \text{Prob}(2)\text{Prob}(3) \\ &\quad + \text{Prob}(3)\text{Prob}(2) + \text{Prob}(4)\text{Prob}(1) \\ &= \frac{1}{6^2} + \frac{1}{6^2} + \frac{1}{6^2} + \frac{1}{6^2} \\ &= \frac{4}{36} = \frac{1}{9}.\end{aligned}$$

Here again you see convolution at work:

$$\text{Prob}(s = k) = \sum_{j=1}^{k-1} \text{Prob}(j)\text{Prob}(k - j). \tag{13}$$

If we represent the probability distribution corresponding to the throwing of a dice by $p_1 = \frac{1}{6}(1, 1, 1, 1, 1, 1)$, then the probability distribution corresponding the "sum of two throws of a dice" is

$$p_2 = p_1 * p_1 = \frac{1}{36}(1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1).$$

You should check by a calculation what

$$p_3 = p_1 * p_1 * p_1$$

is. (Now we are observing the sum of the outcomes of three throws of a dice. There are 16 possibilities ranging between 3 and 18.) Plot the points corresponding to p_1, p_2, p_3,. The plots look like the one in *Figures* 1, 2, and 3.

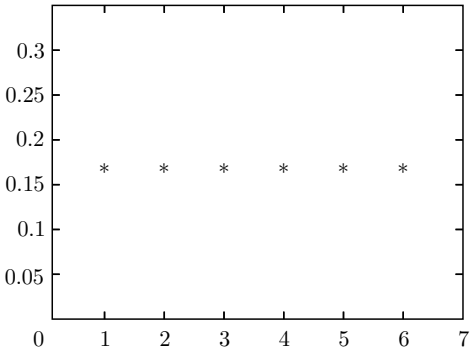

Figure 1

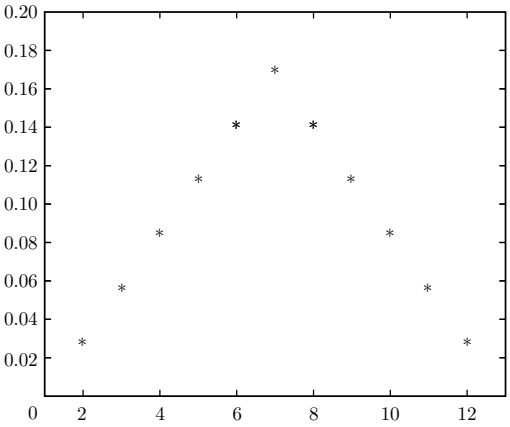

Figure 2

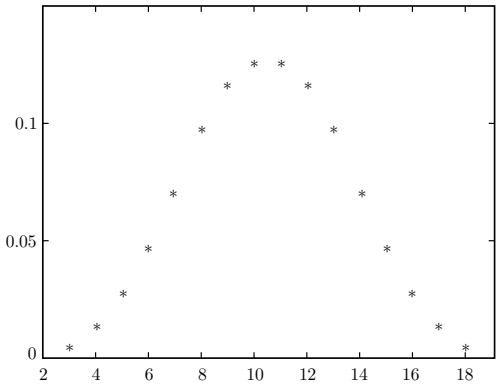

Figure 3

I will now discuss the "continuous version" of the same phenomenon. Let $p(x)$ be a function on the real line $(-\infty, \infty)$ satisfying two conditions

$$p(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} p(x)dx = 1.$$

Such a function is called a *probability density function.* This corresponds to a "random variable" F which can possibly take all real values, and the probability of F being in the interval $[a, b]$ is

$$\int_a^b p(x)\ dx.$$

If p_1 and p_2 are probability density functions, their convolution $p_1 * p_2$ is defined as

$$(p_1 * p_2)(x) = \int_{-\infty}^{\infty} p_1(t)\, p_2(x - t)\, dt. \tag{14}$$

Observe the similarity with the discrete convolution defined in (8). (The sum has now been replaced by an integral and the indices k and j by x and t, respectively.) The function $(p_1 * p_2)(x)$ is another probability distribution. If p_1 and p_2 correspond to random variables F_1 and F_2 then $p_1 * p_2$ corresponds to their sum $F_1 + F_2$. We saw this in the case of two throws of a dice. The general case involves a similar calculation with integrals.

As a simple example, let us consider

$$p_1(x) = \begin{cases} 1 & \text{if} \quad |x| \leq 1/2 \\ 0 & \text{if} \quad |x| > 1/2. \end{cases}$$

The graph of p is *Figure* 4.

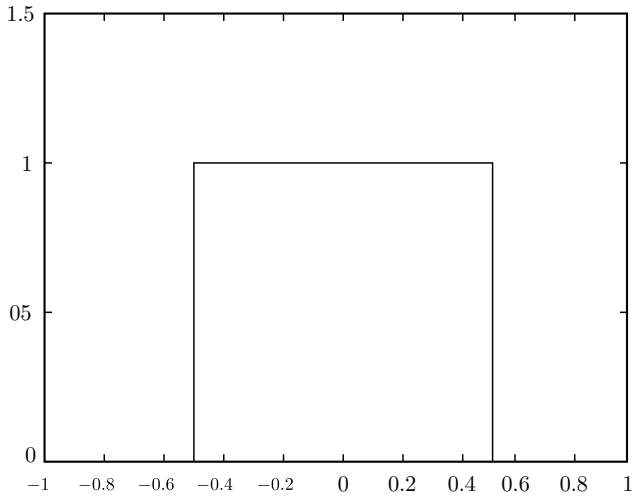

Figure 4

This is called a "rectangular distribution". You are invited to calclulate p_2 defined as

$$p_2(x) = (p_1 * p_1)(x) = \int_{-\infty}^{\infty} p_1(t)p_1(x-t)dt.$$

(It is a simple integration.)

You will see that

$$p_2(x) = \begin{cases} 1 - |x| & \text{if} \quad |x| \le 1 \\ 0 & \text{if} \quad |x| \ge 1. \end{cases}$$

The graph of p_2 is *Figure* 5.

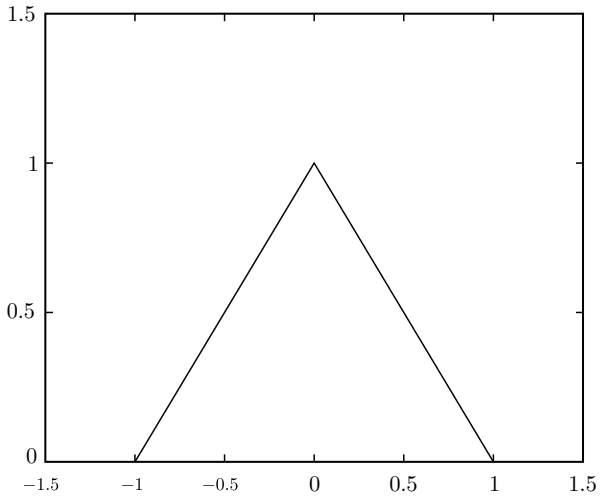

Figure 5

Let us persist a little more and calculate

$$p_3(x) = (p_1 * p_2)(x) = (p_1 * p_1 * p_1)(x).$$

The answer is

$$p_3(x) = \begin{cases} \frac{1}{8}(3 - 2|x|)^2 & \text{if} \quad \frac{1}{2} \le |x| \le \frac{3}{2} \\ \frac{3}{4} - x^2 & \text{if} \quad |x| \le \frac{1}{2} \\ 0 & \text{if} \quad |x| \ge \frac{3}{2}. \end{cases}$$

The graph of p_3 normalized so that $p_3(0) = 1$ is *Figure* 6.

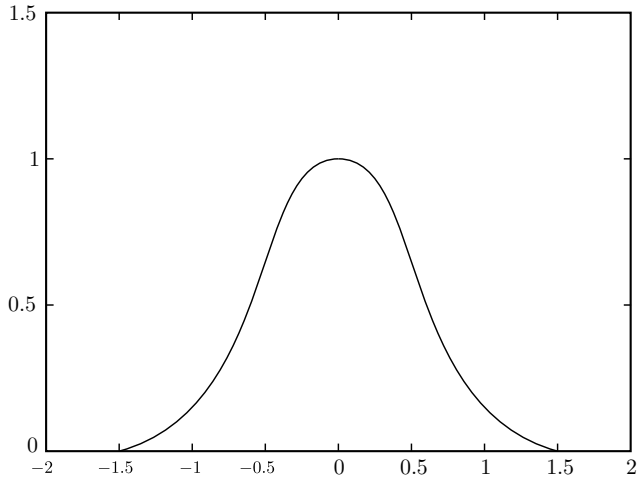

Figure 6

We can go on and calclulate $p_4(x)$. I asked a computer to do it for me and to show me the graph of p_4. It is *Figure* 7.

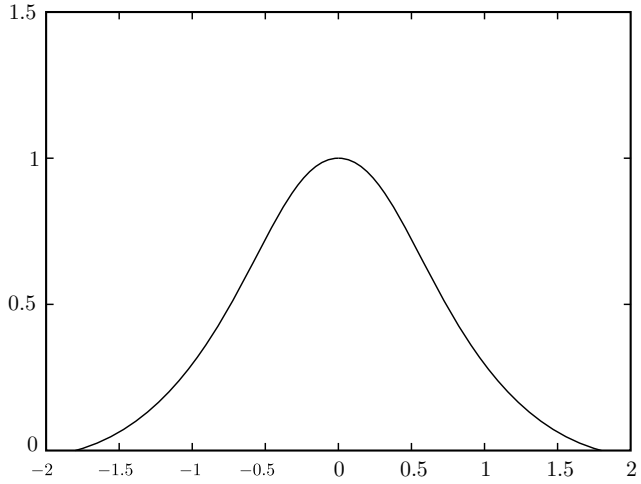

Figure 7

Do you see a pattern emerge? The graphs seem to look more and more like the "normal curve", the famous bell-shaped curve.

Was there something special about the rectangular distribution that led to this'? I start with another distribution

$$p_1(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & \text{if} \quad |x| \le 1 \\ 0 & \text{if} \quad |x| \ge 1. \end{cases}$$

This looks like *Figure* 8.

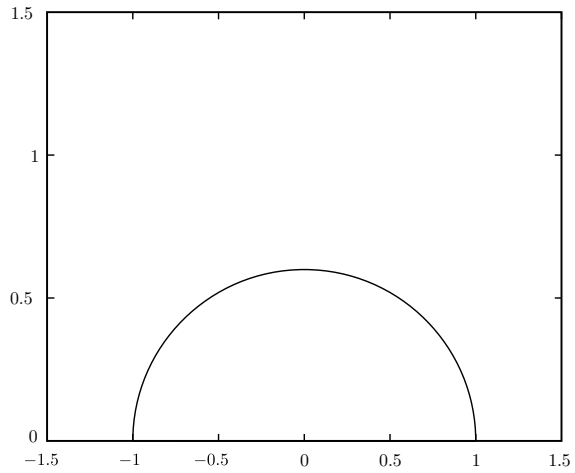

Figure 8

Successive convolutions of p_1 with itself p_2, p_3 and p_4 have graphs *Figures* 9, 10, 11.

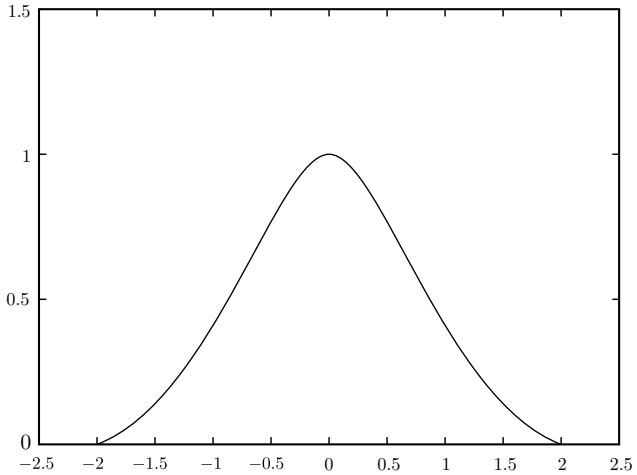

Figure 9

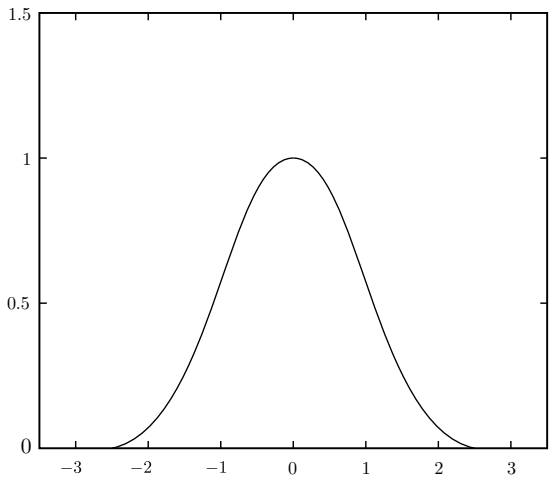

Figure 10

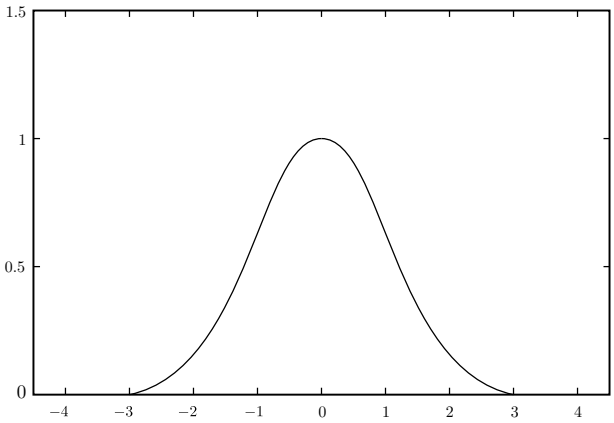

Figure 11

Here is yet another example in which the function is "random" (*Figure* 12). Again three successive convolutions are shown in the (*Figures* 12-15) that follow.

This seems to be a very striking phenomenon. Starting with different probability distributions we seem to get close to a normal distribution if we take repeated convolutions. Does this happen always? (The answer is: "with rare exceptions".) So the normal distribution occupies a very special position. One of the most important theorems in probability is the "Central Limit Theorem". That tells us more about this phenomenon. I hope you will find out about this soon. Another feature that stands out in these examples is that successive convolutions seem to make the functions smoother. This too is a general phenomenon, exploited by mathematicians and by design engineers.

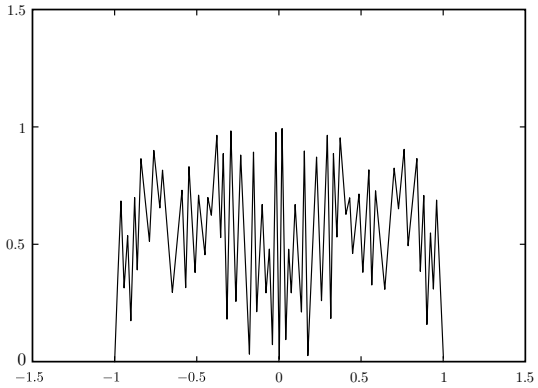

Figure 12

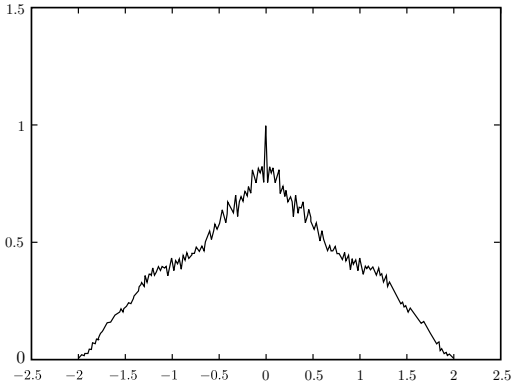

Figure 13

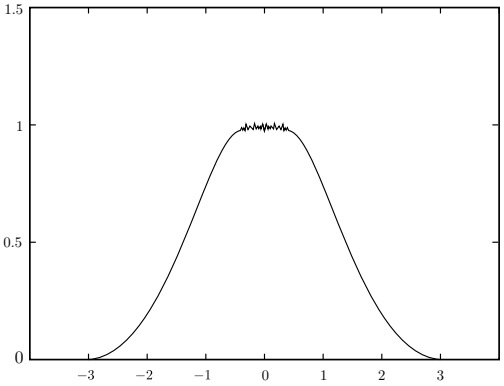

Figure 14

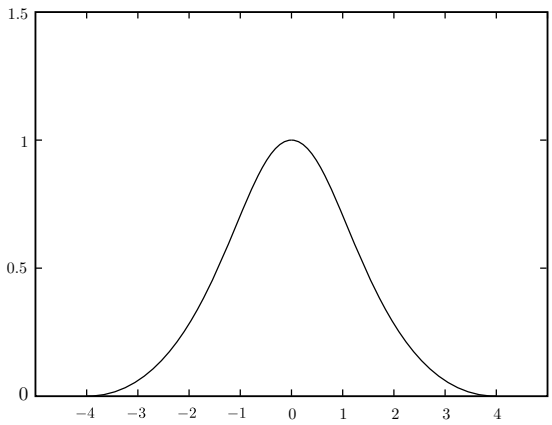

Figure 15

Finally I wish to point out that there is an analogy between multiplication of ordinary numbers and that of polynomials. Every number can be thought of as a polynomial with the "base" of the system acting as the "ideterminate" x. Thus, for example, in the decimal system

$$3769 = 9 + 6.10 + 7.10^2 + 3.10^3$$

Ordinary multiplication of numbers is, therefore akin to multiplication of polynomials. There is a famous algorithm called the Fast Fourier Transform that computes convolutions quickly and helps computers do arithmetic operations like multiplication much faster.

(I thank Mrs Srijanani Anurag Prasad for preparing the figures.)

Vibrations and Eigenvalues

Rajendra Bhatia
Indian Statistical Institute, New Delhi 110 016, India

President's Address at the 45th Annual Conference of the Association of Mathematics Teachers of India, Kolkata, December 27, 2010

Vibrations occur everywhere. My speech reaches you by a series of vibrations starting from my vocal chords and ending at your ear drums. We make music by causing strings, membranes, or air columns to vibrate. Engineers design safe structures by controlling vibrations.

I will describe to you a very simple vibrating system and the mathematics needed to analyse it. The ideas were born in the work of Joseph-Louis Lagarange(1736–1813), and I begin by quoting from the preface of his great book *Méchanique Analitique* published in 1788:

We already have various treatises on mechanics but the plan of this one is entirely new. I have set myself the problem of reducing this science [mechanics],and the art of solving the problems pertaining to it, to general formulae whose simple development gives all the equations necessary for the solutions of each problem ... No diagrams will be found in this work. The methods which I expound in it demand neither constructions nor geometrical or mechanical reasonings, but solely algebraic [analytic] operations subjected to a uniform and regular procedure. Those who like analysis will be pleased to see mechanics become a new branch of it, and will be obliged to me for having extended its domain.

Consider a long thin tight elastic string (like the wire of a *veena*) with fixed end points. If it is plucked slightly and released, the string vibrates. The problem is to find equations that describe these vibrations and to find solutions of these equations. The equations were first found by Jean d'Alembert, and two different forms of the solution were given by him and by Leonhard Euler.

Lagrange followed a different path: he *discretised* the problem. Imagine the string is of length $(n+1)d$, has negligible mass, and there are n beads of mass m each placed along the string at regular intervals d:

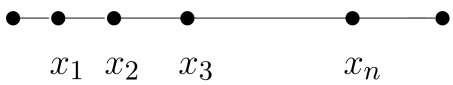

Figure 1

The string is pulled slightly in the y-direction and the beads are displaced to positions $y_1, y_2, ..., y_n$.

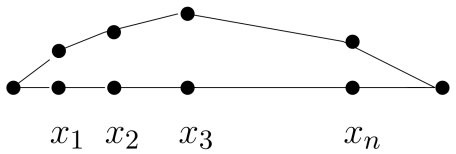

Figure 2

The tension T in the string is a force that pulls the beads towards the initial position of rest. Let α be the angle that the string between the $(j-1)$th and the jth bead makes with the x-axis:

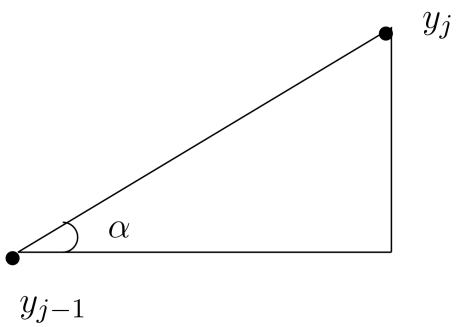

Figure 3

Then the component of T in the downward direction is $T \sin\alpha$. If α is small, then $\cos a$ is close to 1, and $\sin\alpha$ is close to $\tan\alpha$. Thus the downward component of T is approximately

$$T \tan\alpha = T\frac{y_j - y_{j-1}}{d}.$$

Similarly the pull exerted on the jth bead from the other side of the string is

$$T\frac{y_j - y_{j+1}}{d}.$$

Thus the total force exerted on the jth bead is

$$\frac{T}{d}(2y_j - y_{j-1} - y_{j+1}).$$

By Newton's second law of motion

$$\text{Force} = \text{mass} \times \text{acceleration},$$

this force is equal to $m\ddot{y}_j$, where the two dots denote the second derivative with respect to time. So we have

$$m\ddot{y}_j = \frac{-T}{d}(2y_j - y_{j-1} - y_{j+1}). \tag{1}$$

The minus sign outside the brackets indicates that the force is in the 'downward' direction. We have n equations, one for each $1 \le j \le n$. It is convenient to write them as a single vector equation

$$\begin{bmatrix} \ddot{y}_1 \\ \ddot{y}_2 \\ \vdots \\ \ddot{y}_n \end{bmatrix} = \frac{-T}{md} \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & -1 \\ & & & & -1 & 2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{2}$$

or as

$$\ddot{\boldsymbol{y}} = \frac{-T}{md} L\boldsymbol{y}, \tag{3}$$

where $\boldsymbol{y}$ is the vector with n components $y_1, y_2, \ldots, y_n$ and L is the $n \times n$ matrix with entries $l_{ii} = 2$ for all i, $l_{ij} = -1$ if $|i - j| = 1$, and $l_{ij} = 0$ if $|i - j| > 1$. (A matrix of this special form is called a *tridiagonal* matrix.)

Let us drop the factor $-T/md$ (which we can reinstate later) and study the equation

$$\ddot{\boldsymbol{y}} = L\boldsymbol{y}. \tag{4}$$

We want to find solutions of this equation; i.e., we want to find $\boldsymbol{y}(t)$ that satisfy (4). In this we are guided by two considerations. Our experience tells us that the motion of the string is oscillatory; the simplest oscillatory function we know of is $\sin t$, and its second derivative is equal to itself with a negative sign. Thus it would be reasonable to think of a solution

$$\boldsymbol{y}(t) = (\sin \omega t)\boldsymbol{u}. \tag{5}$$

If we plug this into (4), we get

$$-\omega^2(\sin \omega t)\boldsymbol{u} = (\sin \omega t)L\boldsymbol{u}.$$

So, we must have

$$L\boldsymbol{u} = -\omega^2\boldsymbol{u}.$$

In other words $\boldsymbol{u}$ is an *eigenvector* of L corresponding to *eigenvalue* $-\omega^2$.

So our problem has been reduced to a problem on matrices: find the eigenvalues and eigenvectors of the tridiagonal matrix L. In general, it is not easy to find eigenvalues of a (tridiagonal) matrix. But our L is rather special. The calculation that follows now is very ingenious, and remarkable in its simplicity.

The characteristic equation $L\boldsymbol{u} = \lambda\boldsymbol{u}$ can be written out as

$$-u_{j-1} + 2u_j - u_{j+1} = \lambda u_j, 1 \le j \le n, \tag{6}$$

together with the *boundary conditions*

$$u_0 = u_{n+1} = 0. \tag{7}$$

The two conditions in (7) stem from the fact that the first and the last row of the matrix L are different from the rest of the rows. This is because the two endpoints of the string remain fixed – their displacement in the y-direction is zero. The trigonometric identity

$$\begin{aligned}\sin(j+1)\alpha + \sin(j-1)\alpha &= 2\sin j\alpha\cos\alpha \\ &= 2\sin j\alpha\left(1 - 2\sin^2\frac{\alpha}{2}\right),\end{aligned}$$

after a rearrangement, can be written as

$$-\sin(j-1)\alpha + 2\sin j\alpha - \sin(j+1)\alpha = \left(4\sin^2\frac{\alpha}{2}\right)\sin j\alpha. \tag{8}$$

So, the equations (6) are satisfied if we choose

$$\lambda = 4\sin^2\frac{\alpha}{2}, \quad u_j = \sin j\alpha. \tag{9}$$

There are some restrictions on α. The vector $\boldsymbol{u}$ is not zero and hence α cannot be an integral multiple of π. The first condition in (7) is automatically satisfied, and the second dictates that $\sin(n+1)\alpha = 0$.

This, in turn means that $\alpha = k\pi/(n+1)$. Thus the n eigenvalues of L are

$$\lambda = 4\sin^2\frac{k\pi}{2(n+1)}, k = 1, 2, \ldots, n. \tag{10}$$

You can write out for yourself the corresponding eigenvectors.

What does this tell us about our original problem? You are invited to go back to ω and to the equation (3) and think. A bit of 'dimension analysis' is helpful here. The quantity T in (3) represents a force. So its units are $\frac{\text{mass}\times\text{length}}{(\text{time})^2}$. The units of $\frac{T}{md}$ are, therefore $(\text{time})^{-2}$. So, after the factor $\frac{-T}{md}$ is reinstated, the quantity ω represents a frequency. This is the frequency of oscillation of the string. It is proportional to $\sqrt{T/md}$. So, it increases with the tension and decreases with the mass m of the beads and the distance d between them. Does this correspond to your physical experience?

We can go in several directions from here. Letting d go to zero we approach the usual string with uniformly distributed mass. The matrix L then becomes a differential operator. The equation corresponding to (3) then becomes Euler's equation for the vibrating string. We can study the problem of beads on a *heavy* string. Somewhat surprising may be the fact that the same equations describe the flow of electricity in telephone networks.

The study of the vibrating string led to the discovery of Fourier Series, a subject that eventually became 'harmonic analysis', and is behind much of modern technology from CT scans to fast computers.

I end this talk by mentioning a few more things about Lagrange. Many ideas in mechanics go back to him. It has been common to talk of 'Lagrangian Mechanics' and 'Hamiltonian

Mechanics' as the two viewpoints of this subject. Along with L Euler he was the founder of the *calculus of variations*. The problem that led Lagrange to this subject was his study of the *tautochrone*, the curve moving on which a weighted particle arrives at a fixed point in the same time independent of its initial position. The Lagrange method of undetermined multipliers is one of the most used tools for finding maxima and minima of functions of several variables. Every student of group theory learns Lagrange's theorem that the order of a subgroup H of a finite group G divides the order of G. In number theory he proved several theorems, one of which called 'Wilson's theorem' says that n is a prime if and only if $(n-1)! + 1$ is divisible by n. In addition to all this work Lagrange was a member of the committee appointed by the French Academy of Sciences to standardise weights and measures. The metric system with a decimal base was introduced by this committee.