Alexander A. Roytvarf

Thinking in Problems

How Mathematicians Find Creative Solutions

Birkhäuser

Alexander A. Roytvarf

Thinking in Problems

How Mathematicians Find Creative Solutions

Birkhäuser

Alexander A. Roytvarf
Rishon LeZion, Israel

ISBN 978-0-8176-8405-1 ISBN 978-0-8176-8406-8 (eBook)
DOI 10.1007/978-0-8176-8406-8
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012950315

Mathematics Subject Classification (2010): 97H10, 97I10, 97K20, 97K50, 97K60, 97K80, 97M50

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.birkhauser-science.com)

Preface

This book is written for people who think that mathematics is beautiful, for people who want to expand their mathematical horizons and sharpen their skills. The best way to sharpen mathematical skills is to use them in solving problems.

Unlike other problem books, this book contains few routine exercises, nor is it a collection of olympiad problems. As we claim in the title, *we aimed to create an atmosphere of real mathematical work for readers*. Therefore, we sought to collect and explore two kinds of problems united by a common methodology. Problems of the first kind include nice theoretical material from the field of mathematics and are designed to teach readers to understand the math and to help them master the mathematical techniques by working on the problems. We kept to a common approach in teaching mathematics: "For effective learning, the learner should discover by himself as large a fraction of the material to be learned as feasible under the given circumstances. This is the principle of active learning (Arbeitsprinzip). It is a very old principle: it underlies the idea of the "Socratic method" [Polya 1962–1965]. Problems of the second kind arise in real-life mathematical research. By necessity, the scope of this book is too narrow for a methodical exposition of applications of mathematical theory to processes and methods in life and in the work place. Such an exposition necessitates including a lengthy introduction to the applied aspects of the real-life problems, so the emphasis of our discussion is on the mathematical aspects of these problems. Having described and explained the theoretical background and methodology in mathematical terms, we invite the reader to go on to obtain mathematical results relating to real-life outcomes. (However, when a lengthy introduction to the applied aspect is not required, such as in the problem group "A Combinatorial Algorithm in Multiexponential Analysis," we depart from this rule and give the reader a nice opportunity to test himself on solving a real-life problem in what is essentially its original formulation.) Thus, we seek to show the reader that the same principles underlie work in both pure and applied mathematics. Some problems in this book pose a serious challenge for most readers; those who are prepared to work hard and undertake this challenge will gain the most out of this work.

The prerequisites for working with this book mostly correspond to the graduate level, so the book is addressed primarily to this category of readers. Undergraduate students will be able to solve a substantial number of problems in this text, including all problems that do not require the reader to wield mathematical skills (mainly in linear algebra and analysis) that are outside the scope of what is usually taught at the undergraduate university level. We also hope that this book will be useful for teachers of higher education working with students of mathematics. Professional mathematicians may find in it material which would be interesting to them (e.g., new problems and new approaches to solving some well-known problems).

For the reader's convenience, we have devised a system of stars marking the problem or problem set to indicate the required background: no stars () indicates elementary material, while one (*), two (**), or three (***) stars correspond to the recommended number of semesters of a university-level math curriculum (a detailed specification of the related key notions and theorems is included in the section "Using the Stars on Problems" below).

Thus, we assume that the relevant definitions will be known to the reader and there is no need to remind him of most of them. However, to facilitate working with this book, in some cases problem statements are preceded by definitions: e.g., "Recall that a function or a map defined on a Cartesian product of vector spaces is multilinear. . .," and "Recall that a real-valued function of one or many real variables is referred to as convex. . .," etc. In addition, the problem groups in this book are prefaced with "zero" ("preliminary" or "introductory") sections containing related key terms, some definitions, and a carefully selected bibliography. (Obviously, *these short introductions are no replacement for regular university courses*.)

The reader is always warned if comprehending a problem requires knowledge that goes beyond what is delineated by the stars: e.g., "Therefore, here is an exercise for readers familiar with multivariate integration," or "This and the next two problems are addressed toward readers familiar with normed vector spaces," or "This problem is addressed to readers familiar with ordinary linear differential equations," etc.; see details in the section "Understanding the Advanced Skill Requirements" below. Those problems that require the reader to have a stronger mathematical background have been emphasized in this book by using a smaller font size. Paragraphs explaining the theoretical background on which these problems are based have also been included.

This book also contains a number of problems that could be successfully solved with the help of some tool outside of the stated curriculum level. Whenever this is the case, the reader is warned: "If you need reference to additional definitions or tools related to this problem then look at corresponding Hint." We introduce the tool within the "Hint," "Explanation," or "Completing Solution" sections (see below), while also discussing the tool and the problem, as required. We recommend that if readers do not encounter any unfamiliar concepts or terms while reading the problem's formulation, they should attempt to solve it. Once they encounter difficulty in solving, they should try to overcome it, and only then, if necessary,

we may offer them appropriate means to overcome this difficulty (in this case, our seeds will fall into already cultivated soil). Therefore, the structure of the text follows a well-known educational method that works well for students of any background: "For efficient learning, an exploratory phase should precede the phase of verification and concept formalization and, eventually, the material learned should be merged in, and contribute to, the integral mental attitude of the learner" [Polya 1962–1965].

The complete problem set consists of (1) problems that stand on their own and (2) problems combined in groups with a common subject. In general, each problem group includes material more or less traditionally related to the field of mathematics, which is indicated in its title, but readers will also notice a number of "nontraditional" inclusions, mainly related to applications – either in other fields of mathematics or in real-life problems (in these cases, the area is indicated). Some of the groups are united by these real-life applications.

In accordance with the aim of the book, the problems are not restricted to the traditional educational categories: analysis, algebra, etc. On the contrary, the suggested solutions are obtained by combining ideas from different branches of mathematics (as is done in real mathematical work). Thus, solving all the problems in any of the problem groups will provide abundant practice in fundamental topics, such as continuity, intermediate value, and implicit function theorems, power-series expansions (analysis), polynomials, symmetry and skew-symmetry, determinants, eigenvalues, Jordan canonical form (algebra), all of which are indispensable for many real-life problems.

Groups of related problems are organized in such a way that the problems build on each other, providing students with a ladder to master increasingly difficult problems. It is worth emphasizing that this organization also corresponds to the sequence which often occurs in real mathematical work: readers are invited first to understand the simplest theoretical concepts, and then to examine applications of these concepts, which allows them to observe additional properties and to return to the theoretical analysis of the generalized concepts. (In some cases, we even consider it necessary that the reader first finds a cumbersome solution by relatively simple tools, which is far from the nicest possible one. We are following the quite obvious idea that (1) the ability to deftly handle such means is useful for the researcher and should be developed, and, more importantly, (2) in this way the reader will be able to see the limits of the method's applicability and further appreciate the more advanced subtle tools to do the same thing more efficiently. Based on our teaching experience, we believe that this "inductive" approach is the most productive for gaining mathematical skill. To gain as much as possible from this approach, we recommend that readers try to solve the problems in each group in the same order as they appear in the book.

The presentation within each problem group is divided into subsections, some of which include introductory, summarizing, or historical materials. To stimulate active perception of the material, in many cases the problems are stated as questions (e.g., "Why?", "Which exactly?", etc.) or assignments (such as "Complete the details," etc.) that do not interrupt the presentation; hence, within any subsection,

not one but a number of closely related problems can be proposed for solving. For readers' convenience, in the "Problems" section, the first word of the each question or assignment (including the frequently used word "Prove") is typeset in a different font (*Lucida Calligraphy*), and keywords in the material exposition are typeset in **bold**. In many cases, we give preference to assignments, as "*Give...*", "*Extend...*", "*Develop...*", "*Find...*", "*Evaluate...*", "*Describe...*", but not "*Prove*" because it is more consistent with actual mathematical practice. We would also like to emphasize that for a mathematician, the word "*Why*?" may be the most important question word. Also, we use the abbreviation "**QED**" (Quod Erat Demonstrandum) to denote the end of a proof, and typeset it in Arial Black. Mathematical and other symbols that we use in the book are traditional and widely used; perhaps the only exception (for the English-speaking reader) is that we prefer to denote the identity matrix (and the identity operator) by "E", and the symbol "I" is reserved for a square root of $-E$ (where the dimension of the vector space is equal to 2). Any other nonstandard, or local, designations are defined in the same place where used.

We have enclosed an explanation of solutions for all the problems to give the reader an opportunity to compare their solutions with someone else's. The solutions to most of the problems are discussed in stages: first a hint, then a more in-depth explanation, and, finally, everything else that needs to be done to complete the solution. First of all, the reader should try to deal with the problem on his own. If he feels that he did not succeed, he should look into the "Hint" and then try to complete the solution of the problem. If that is not enough, the "Explanation" should be used, etc. The sections "Hint," "Explanation," and "Completing Solution" are numbered similarly to the "Problems" sections: e.g., subsections H1.1 (Hints), E1.1 (Explanations), and S1.1 (Completing Solutions) correspond to P1.1 (Problems), and so on. We sought to engage the reader in the process of problem solving as an active participant, so that questions such as "Why?" and offers to fill in the necessary details are presented in these sections in the same way as in "Problems" (readers can see the answers to these questions provided in the "Explanation" or "Completing Solution" sections).

Although the tastes of readers and their styles of thinking may be different, we believe that it is important that readers learn to see the subject from very different angles. Many great discoveries have been made by people who have such an ability! Therefore, for each problem that may be solved by various methods, we sought to consider all the approaches that were known to us. In addition, in the "Hint," "Explanation," and "Completing Solution" sections, we discuss related elementary subjects and other materials, such as ties with more advanced theories, examples of applications, references, etc. The additional materials make these sections an extremely important part of this book. Therefore, in order to get the most benefit from this book, the reader is advised to look through these sections, even if he found a solution to the problem by himself or has encountered the problem previously.

As should be clear from the above, we assume that the reader will be actively working on problems from the book "with pencil and paper." In this regard, we would like to make clear that the "Completing Solution" section of the book

does not contain the presentations of solutions "from 'A' to 'Z'" (which are not needed if a reader thought through the problem, has already read "Hint" – "Explanation," and therefore wants only the remaining instructions, so that his mind has developed a complete solution). Mainly, this section contains details such as proofs of the lemmas that were formulated and used without proof in "Hint" – "Explanation" sections, and further discussion (sometimes including related problem formulations). The solution may be qualified as complete, only if (1) all gaps in the proof are filled and (2) the reader has a clear view of the place of the solved problem within wider mathematical context and would be ready to work on similar or related problems in the future (because the problem is rarely a goal in and of itself).

There are many differences between the present book and several widely known collections of nonstandard problems. For example, our work differs from a remarkable collection [Steinhaus 1958] in that it explores more advanced topics at the college undergraduate and graduate level.

The main difference from another brilliant collection of problems for college exit exams [Arnold 1991, 1989 Russian] Is that this book includes a smaller number of topics and larger number of problems per topic, allowing detailed and gradual topic development. A few problems from Arnold collection were included in the corresponding problem groups in our book where they fit logically into the problem sequence.

Finally, the present book differs from the famous Polya et al. [1964] by the relatively small number of topics that are explored deeply and in its orientation toward readers with a relatively limited experience in mathematics – namely, undergraduate and graduate students. Unlike Polya, we focus on stimulating the reader to combine ideas from different branches of mathematics to solve problems.

We have achieved our goal if the reader becomes more adept in solving real mathematical problems and we will be quite satisfied if the reader develops a taste for this kind of work.

Some of the problems (especially those related to applied topics) and solutions in this book we have suggested ourselves. Where this is not the case, we have tried to give credit to the authors of the problems and solutions. Absence of a reference means that we do not know the primary source and also that the fact (or method) in question has become a part of "mathematical folklore."

Rishon LeZion, Israel Alexander A. Roytvarf

Using the Stars on Problems

"No stars" refers to the following material:

- *Basic arithmetic operations with integer numbers*, *greatest common divisor of natural numbers as smallest-value linear combination of these numbers with integral coefficients*, and a notion of *unique prime factorization of integer numbers*
- *Basic arithmetic operations with rational numbers*, *root*s and a notion of *irrational and real numbers, decimal expansion of rational and irrational numbers*
- Usual fundamental course in *elementary plane and solid geometry*
- *Identities for exponential and trigonometric functions* and related formulas
- *Calculation with elementary* (polynomial, rational, exponential, logarithmic, and trigonometric) *functions* and *their graphs*
- *Linear equations and inequalities and systems of two or three equations*, *quadratic equations and inequalities*, and *trigonometric equations*
- *Binomial coefficients and related elementary combinatorial formulas*
- "*Dirichlet boxes principle*" (also known as the *pigeonhole principle*)
- *Set-theoretic language for geometrical sets* (using notions such as membership, inclusion, union, intersection, set-difference, empty set), and the *combinatorial principle of inclusion–exclusion*
- *Basic elements of mathematical theory* (axioms, theorems, and inverse theorems) and *basic types of mathematical argumentation* including deduction, the method of mathematical induction and proof by contradiction.

In addition to that, "one star" refers to the following material:

- *Algebra of real and complex numbers*, algebraic operations, including *division with remainder* and *Euclid's algorithm*, with the segments of real axis, notions of a *group*, *ring*, and *field*, *Euler's totient function, congruences modulo natural numbers*, and *Unique prime factorization theorem* for integer numbers
- *Basic set-theoretic notions and theorems* (sets and maps, natural numbers and countable sets, potential and actual infiniteness, Cantor–Bernstein–Schroeder theorem and linear segment and square as continual sets, algebraic operations on

sets including Cartesian (or, direct) product and respective operations on cardinal numbers, and Cantor's theorem $2^{\#X} > \#X$ and uncountability of continuum)

- *Basic elements of topology of Euclidean straight line* (equivalence of Dedekind's, Cantor's, and Weierstrass's approaches, and Hilbert's description of real field as Archimedean-ordered complete field, compactness of segments (Borel's lemma about open cover of the segment)), *supremum and infimum*, *infinite numerical sequence and its upper and lower limits, convergence and limit*, Cauchy criterion, Bolzano-Weierstrass lemma, Stolz's theorem and its applications, *commutativity of limits with arithmetic operations and the other elementary functions* (see the "no stars" section above), *limits of functions*, *"remarkable" limits* $\lim\limits_{x\to 0}\frac{\sin x}{x}=1$, $\lim\limits_{n\to\infty}\left(1+\frac{1}{n}\right)^n=\lim\limits_{n\to\infty}\left(1+\frac{1}{1!}+\ldots+\frac{1}{n!}\right)=e$ *and their applications*, *comparison of different infinitesimal and infinitely large quantities to each other, and symbols "O" and "o"*
- Basic theorems about *continuity and uniform continuity of functions of one variable* (classification of discontinuities, discontinuities of monotonic functions, continuity of inversion of monotonic continuous function, continuity of superposition, continuity of elementary functions, theorems by Bolzano, Cauchy, Weierstrass, and Cantor)
- *Differential calculus of functions of one variable*: the first- and higher-order derivatives and differentials; continuity of differentiable function; Taylor's expansion theorem, expansions of elementary functions; differentiable change of variable and differentiation of superposition; Leibnitz derivation rule; differentiation of elementary functions; usage of derivatives to explore monotonicity and extrema; convexity (concavity) and inflections and asymptotic behavior; Darboux, Rolle, Lagrange, and Cauchy Intermediate value theorems and their applications; limits of derivatives; "L'Hôpital's rule"; and Implicit function theorem for one variable
- *Solution of systems of linear equations*: linear dependence, vector spaces, and vector and affine subspaces; bases; rank of matrix; Kronecker-Capelli theorem; linear transformations and their matrix representation; similar matrices; and linear changes of coordinates
- *Multi-linearity of functions* and *skew-symmetry* (introduction): permutations, determinants, and oriented volumes

In addition to that, "two stars" refers to the following material:

- *Elements of algebra of polynomials* (division with remainder, Euclid's algorithm, irreducible decomposition, Bezout's theorem, the fundamental theorem of algebra (without a complete proof), roots of polynomials over the real and rational field, multiple roots and derivative, resultant and discriminant, field of fractions (rational functions) and partial fraction expansion over complex and real fields, the fundamental theorem on symmetric polynomials), and *polynomial change of variable* (putting polynomial in polynomial)

- *Definite (Riemann) integrals and primitives (indefinite integrals) of one-variable functions*: Darboux integrals, Newton-Leibnitz formula, methods of integration, changes of variables in integrals, integration of elementary functions, classes of integrable functions, improper integrals, mean value theorems, integrals dependent on parameters, applications of integral calculus to geometry and physics
- *Jordan canonical form of linear operators*: eigenvalues and eigenspaces, root subspaces, algebraic and geometric multiplicity, characteristic and minimal polynomials, Cayley-Hamilton theorem, Jordan canonical forms over complex and over real numbers
- *Euclidean and Hermitian (or unitary) finite-dimensional spaces* (scalar product, orthonormal bases), Cauchy-Schwarz-Bunyakovskii (CSB) inequality, Gram-Schmidt orthogonalization, adjoint operators (in Euclidean and Hermitian cases), spectral theorems for orthogonal and unitary operators, and symmetric and self-adjoint operators, and theorems of *Elementary geometry* in *Cartesian coordinate* language
- *Bilinear and quadratic forms*: matrix representation of bilinear form and linear changes of coordinates, bilinear forms and linear functionals, duality, Lagrange diagonalization of quadratic forms, Sylvester's theorem, rank and signature, positive (negative) definite and semidefinite quadratic forms, diagonalization by an orthogonal linear transformation, simultaneous diagonalization of two quadratic forms and relative eigenvalues, diagonalization of skew-symmetric bilinear forms, and classification of real and complex quadrics
- *Basic elements of topology of Euclidean space*, *multiple and iterated limits*, and basic theorems (similar as listed in the "one-star" section above) about *continuity and uniform continuity of multi-variable functions*
- *Differential calculus of multi-variable functions on Euclidean domains and smooth varieties* (curves and surfaces in Euclidean spaces): differential, continuity of differentiable function, directional derivatives, partial derivatives; gradient; the first- and higher-order differentials; Clairaut's (or, Schwarz's) theorem about mixed derivatives; extrema and other critical points and Hessians; convexity (concavity); Taylor's expansion theorem; Euler theorem about homogeneous functions; intermediate value theorems; diffeomorphisms and implicit function theorem; change of variable and Jacobian, polar, cylindrical, and spherical coordinate systems; and smooth surfaces in Euclidean spaces and critical points of differentiable functions on such surfaces (conditional extrema and Lagrange multipliers)
- *Numerical series*, and *series and sequences of functions*: convergence and uniform convergence, Leibnitz, Cauchy, d'Alembert, Abel's, Dirichlet's and integral tests for convergence, Dini's and Weierstrass theorems about uniform convergence, and theorems about integrability and differentiability of series' sum
- *Power series of one variable* (formal series, absolute convergence; Abel's lemma; Cauchy-Hadamard formula for convergence radius; and Taylor series of elementary functions, their convergence radii, and singular points on boundaries of discs of convergence in complex domain), *multi-variable power*

series (formal series, normal convergence, majorization), and *formal change of variable* (putting formal power series in formal power series) and its convergence.

In addition to that, "three stars" refers to the following material:

- *Fourier series of one variable*: Fourier coefficients, mean summability, Dirichlet's integral, and Dini's summability test
- *Weierstrass approximation theorem* for continuous functions on a closed segment of a straight line
- *Ordered sets, set-theoretic axiom of choice and equivalent statements* (nonemptiness of Cartesian product $\prod_{i \in I} X_i$ for any nonempty I and X_i, dependent choice principle, Hausdorff and Kuratowski-Zorn lemmas, existence of well-ordering of any set and Cantor's theorem $\forall X,Y : \#X \geq \#Y \text{ or } \#X < \#Y$, and *Transfinite induction principle*
- *Elements of abstract algebra*: direct products of groups and rings, residue classes with respect to subgroup, invariant (or, normal) subgroups and ideals, modules, "fundamental" and Noether homomorphism theorems, finite groups and Lagrange theorem, action of group and stationary subgroups, group of transformations, symmetric groups, cyclic groups, Sylow theorems, solvable groups, structure theorem for finitely generated Abelian groups, rings with Euclid's algorithm (or, Euclidean rings), rings of principal ideals and unique prime factorization theorem, one-sided ideals of noncommutative ring, fields and bodies, Noetherian rings, rings of formal series, unique prime factorization property of polynomials over factorial ring of coefficients (commutative ring of unique prime factorization property), Gauss lemma, ring of quotients, rational fractions and partial fraction expansion, elements of theory of fields (finite and algebraic extensions, algebraic closure, Galois finite-element fields, roots of unity, transcendental extensions), and using commutative diagrams
- *Multi-linearity of functions* and *skew-symmetry* (continuation): exterior forms, vector product, and mixed product.

The reader who wishes to brush up on elementary material corresponding to "No stars," and look at it from a more advanced point of view, can do so by referring to the excellent books [Arnol'd 2004 Russian], [Vilenkin 1969–1,2], [Sierpiński 1969], [Sivashinsky 1967 Russian], [Stoll 1975], [Steinhaus 1938], [Klein 1924, 1925], [Hilbert 1830], [Niven 1961], [Landau 1927, 1930], [Courant et al. 1941], [Polya 1954, 1962–1965, 1971], [Rényi 1967], [Lakatos 1976, 1980], [Hadamard 1954/1996, 2008, 1898–1901 French], [Hilbert et al. 1932], [Zetel 1962 Russian], [Dieudonné 1968 French], [Choquet 1964 French], and [Kline 1981, 1982, 1986]. With regard to the material corresponding to the "one star" to "three stars," they can be found in the textbooks and monographs, which are usually recommended for the study of mathematics at the university; the reader can use the ones that are listed below, or others – by his/her own choice:

On calculus and analysis and their applications: [Lang 1999], [Hardy 1908], [Landau 1934], [Fichtengoltz 1970 Russian], [Whittaker et al. 1927], [Titchmarsh 1939], [Apostol 1967, 1969], [Cartan 1967], [Courant 1992], [Dieudonné 1960], [Hardy et al. 1956], [Polya et al. 1964], [Folland 1999], [Rudin 1976, 1986], [Schwartz 1967], [Bourbaki 1976], [Hörmander 1983], [Spivak 1994], [Zel'dovich 1963/Zel'dovich et al. 1982], [Zel'dovich et al. 1972, 1973 Russian].

On linear and multilinear algebra and geometry, polynomials, and abstract algebra: [Artin 1957], [Dummit et al. 2003], [Gelfand 1971], [Glazman et al. 1969], [Halmos 1974], [Berger 1977 French], [Lelong-Ferrand 1985 French], [Kostrikin et al. 1980], [Van der Waerden 1971, 1967], [Lang 1965, 1999], [Artin 1991], [Zariski et al. 1958–1960], [Atiyah et al. 1994], [Bourbaki 1970–1981, 1972], [Kargapolov et al. 1977], [Hall 1975], [Herstein 2005], [Strang 2009].

On set-theoretic notions and theorems, axiom of choice (AC) and its equivalents: [Hausdorff 1914], [Arkhangelsky et al. 1974], [Engelking 1985], [Kuratowski et al. 1976], [Jech 2006], [Bourbaki 1970], [Zermelo 2010], [Fraenkel et al. 1973], [Gödel 1940], [Cohen 2008], [Shoenfield 1967], [Barwise ed. 1977], [Moore 1982].

Understanding the Advanced Skill Requirements

If readers are warned of an advanced requirement for a problem that is not stipulated by stars, it refers to the following material:

- *Fundamentals of differential geometry of planar curves*: rectifiability and length, natural parametrization, osculating circle and curvature, Frenet equations, Gauss map and Gauss-Bonnet formula (version for curves), focal points, evolutes and evolvents, envelopes, and equidistants
- *Differential geometry of hypersurfaces in Euclidean spaces*: local parametrization (coordinate system) and its change, tangent bundle and field of unit normals, orientability, first quadratic form (Riemannian metric), ($n-1$-dimensional) area, tangency order, osculating hyperspheres and second quadratic form (principal curvatures, mean curvature, Gaussian curvature, Gauss map), Gauss-Bonnet formula, focal (caustic) points, envelopes, and equidistants
- "*Multivariate integration*": integral calculus of functions on Euclidean domains and on smooth varieties (curves and surfaces in Euclidean spaces): tangent vector fields, differential forms and their integration, Fubini theorem, Change-of-variable formula, exterior differentiation of differential form, closed and exact forms, divergence, curl, Laplace-Beltrami operator (Laplacian), the general "Stokes" (Newton-Leibnitz-Gauss-Ostrogradskii-Green-Stokes-Poincaré) formula and its classical special cases
- *Language of differentiable manifolds*: local coordinate systems and their changes; submanifolds; tangent bundle and vector fields; cotangent bundle and differential forms; tensors; smooth maps and tangent and cotangent maps; Lie commutator and exterior differentiation and their commutativity with tangent and cotangent maps, respectively; Lie differentiation; Cartan's infinitesimal (chain) homotopy formula, theorems of differential and integral calculus from Sections "Using the Stars on Problems" and "Multivariate Integration" in terms of manifolds
- *Elements of general topology*: topological spaces; convergence and continuity; bases of neighborhoods, interior, boundary and closure; Boolean operations with open and closed sets; connectedness; compactness

- *Metric spaces*: distance function (metric), countable bases of neighborhoods of point, Cauchy criterion, completeness, boundedness, compactnesss
- *Normed vector spaces*: norm, metric and topology, norms and centrally symmetric open convex sets, Riesz theorem (local compactness ⇔ finite dimensionality), topological equivalence of norms on finite-dimensional space, compact sets in that space, norm of linear operator, Hilbert space, and Banach space
- *Elements of complex analysis*: Cauchy-Riemann equations, Taylor's and Laurent expansions and Cauchy integral formula for the coefficients, openness of image, residues, Argument principle, Rouché's theorem, Maximum principle and Mean value theorem, Liouville theorem and "the fundamental theorem of algebra," Joukowsky map (transform), rational functions on Gaussian (Riemannian) sphere, fractional-linear automorphisms, analytic continuation, singularities and many-valued functions (algebraic and logarithmic ramifications and Riemannian surfaces), Casorati-Sokhotskii-Weierstrass and Picard theorems, elements of multi-variable complex analysis (meromorphic forms as closed differential forms, Taylor's expansion, biholomorphic maps and implicit function theorem, analytic continuation principle, domain of holomorphy, Hartogs theorem)
- *Elements of probability theory*: probability measure, univariate/multivariate probability distribution (density function) and cumulative distribution function, and mathematical expectation (mean value) and variance
- *Elements of mathematical statistics*: covariance matrix (variances and correlation coefficients), univariate/multivariate normal (Gaussian) distribution, parameter estimation, Maximal likelihood principle, and unbiased estimate
- *Elements of theory of linear ordinary differential equations (ODE)*: unique existence and maximal possible growth of solution, Liouville theorem about Wronskian, Sturm-Liouville boundary eigenvalue problem and Sturm theorem about zeros, quasipolynomials and solution of autonomous system

This material can be learned from the following textbooks and monographs (as well as from other sources of the reader's choice):

On differential geometry of curves and surfaces: [Blaschke 1950], [Finnikov 1952 Russian], [Favard 1957], [Cartan 1967], [do Carmo 1976], [Thorpe 1979], [Dubrovin et al. 1986], [Kreyszig 1991], [Pressley 2010].

On multivariate integration: [Whitney 1957], [Spivak 1965, 1994], [Arnol'd 1989], [Rudin 1976], [Cartan 1967], [Warner 1983], [Dubrovin et al. 1986], [Sternberg 1964], [Schwartz 1967], [Bott et al. 1982].

On differential manifolds: [de Rham 1955 French], [Bishop et al. 1964], [Milnor 1997], [Golubitsky et al. 1974], [Bruce et al. 1993], [Arnol'd 1978, 1989], [Arnol'd et al. 1982], [Guillemin et al. 1974], [Hirsch 1976], [Warner 1983], [Bourbaki 1967–1971], [Lang 2002], [Schwartz 1968], [Dubrovin et al. 1986], [Wallace 2006], [Kosinski 2007].

On General topology: [Hausdorff 1914], [Bourbaki 1960], [Kuratowski 1966], [Chinn et al. 1966], [Kelley 1975], [Munkres 1974], [Arkhangelsky et al. 1974], [Engelking 1985].

On metric and normed spaces: [Kuratowski 1966], [Kelley 1975], [Dieudonné 1960], [Rudin 1973, 1986], [Kirillov et al. 1988], [Kantorovich 1972], [Banach 1932], [Akhiezer et al. 1950], [Dunford et al. 1957], [Hille et al. 1957], [Yosida 1965], [Edwards 1965], [Halmos 1967], [Riesz et al. 1972], [Reed et al. 1972], [Kolmogorov et al. 1976], [von Neumann 1932], [Bourbaki 1960, 1981], [Krein ed. & coaut. 1972].

On complex analysis: [Ahlfors 1979], [Bochner et al. 1948], [Cartan 1961], [Dieudonné 1960], [Whittaker et al. 1927], [Titchmarsh 1939], [Hurwitz/Courant 1964], [Polya et al. 1964], [Schwartz 1967], [Rudin 1986], [Narasimhan 2000], [Hörmander 1966], [Privalov 1984 Russian], [Lavrentiev et al. 1973 Russian], [Evgrafov 1991 Russian].

On probability theory and math statistics: [Poincaré 1912], [Borel 1963], [Kac 1956, 1957, 1959], [Kac et al. 2004], [Kolmogorov 1974], [Feller 1966], [Loeve 1977–1978], [Shiryaev 1989], [Guillemin et al. 1996], [Cramér 1946], [Van der Waerden 1957].

On ODE: [Arnol'd 1975, 1978], [Sansone 1948–1949 Italian], [Coddington et al. 1955], [Lefschetz 1957], [Kamke 1977], [Hartman 1964], [Hille 1969], [Levitan et al. 1970], [Arrowsmith et al. 1982], [Petrovski 1984], [Edwards et al. 1999], [Nagle et al. 2003].

Acknowledgments

I wish to thank my colleagues Professors Victor Katsnelson and Yosef Yomdin from Weizmann Institute of Science, Ann Melnikov from University of Haifa, Henryk Zoladek from University of Warsaw, Pierre Milman from University of Toronto, and Alexander Tovbis from University of Central Florida for their interest and support at various stages of the manuscript preparation.

I am grateful to my friend Alexander Zbarsky for his support and great help in preparing and editing the manuscript.

And, I will always be grateful to my teacher Professor Vladimir Igorevich Arnol'd (1937–2010).

Contents

Jacobi Identities and Related Combinatorial Formulas

Problems

P1.0

Preliminaries. The derivatives of a quadratic trinomial $F = ax^2 + bx + c$ at its roots x_1, x_2 add up to zero (Vieta's theorem). Alternatively, we can write this as $1/F'(x_1) + 1/F'(x_2) = 0$ or, equivalently, $1/(x_1 - x_2) + 1/(x_2 - x_1) = 0$. Also, there is an obvious identity: $x_1/(x_1 - x_2) + x_2/(x_2 - x_1) = 1$. C.G.J. Jacobi's concern with questions of dynamics and geometry led him to derive far-reaching generalizations of these identities for polynomials of arbitrary degree in his Ph.D. thesis (Jacobi 1825). It enabled him to solve several complex problems. For example, he squared an ellipsoid's surface, determined the geodesics on this surface, and described the dynamics with two motionless centers of gravity (Jacobi 1884).

In this problem group, readers will become acquainted with the Jacobi identities and related formulas (P1.1**, P1.2**, P1.3**, P1.4**, P1.5**, P1.6**, and P1.7**) and their close ties with complex analysis (sections P1.8***, P1.9***, P1.10***, and P1.11***); sections P1.12** and P1.13*** are dedicated to less traditional applications of the Jacobi identities to linear differential equations. Some final notes about further generalizations and applications and a short guide to the literature are provided in section P1.14***.

P1.1**

Prove the following identities (Jacobi 1825). For distinct complex numbers (or elements of any other field) $x_1, \ldots, x_n$,

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_1,

$$\sum_{i=1}^{n} \frac{x_i^m}{\prod_{j\neq i}(x_i - x_j)} = \begin{cases} 0, & \text{for } m = 0, \ldots, n-2, \\ 1, & \text{for } m = n-1. \end{cases}$$

P1.2**

With $x_1,\ldots,x_n$ as above, *establish* that

$$\sum_{i=1}^{n} \prod_{j\neq i} \frac{x_j}{x_j - x_i} = 1\,.$$

P1.3**

(Jacobi 1825): With $x_1,\ldots,x_n$, as above and for any $y_1,\ldots,y_{n-1}$, *establish* that

$$\sum_{i=1}^{n} \frac{\prod_{k=1}^{n-1}(x_i - y_k)}{\prod_{j\neq i}(x_i - x_j)} = 1.$$

P1.4**

(Jacobi 1825): *Check* that $u_i = \prod_{1\leq j\leq n}(x_i - y_j)/\prod_{j\neq i}(x_i - x_j)$, $i = 1,\ldots,n$ is a solution of the system of linear equations

$$\sum_{i=1}^{n} \frac{u_i}{x_i - y_j} = 1, \qquad j = 1, \ldots, n$$

($x_1,\ldots, x_n$ are as above, and $x_i \neq y_j$, $\forall i, j$). *When* is it a unique solution?

Give a geometric interpretation for the preceding solution when $n = 2$ and either $x_2 > y_2 > x_1 > y_1$ or $x_1 > y_1 > x_2 > y_2$.

P1.5[**]

(Jacobi 1825): *Prove* the extension of the identities from section P1.1[**] to $m \geq n$: for $x_1,\ldots,x_n$ as above,

$$\sum_{i=1}^{n} \frac{x_i^m}{\prod_{j\neq i}(x_i - x_j)} = \sum_{\sum v_i = m-n+1} \prod_{i=1}^{n} x_i^{v_i},$$

where the sum on the right-hand side is taken over sets of natural numbers v_i such that $\sum v_i = m - n + 1$.

P1.6[**]

Establish the extension of the Jacobi identities for $m = -1$. For distinct nonzero $x_1,\ldots,x_n$ show that

$$\sum_{i=1}^{n} \frac{x_i^{-1}}{\prod_{j\neq i}(x_i - x_j)} = \frac{(-1)^{n-1}}{\prod_{i=1}^{n} x_i}.$$

P1.7[**]

Extend the Jacobi identities for every $m = -1, -2, -3,\ldots$.

P1.8[***]

This section and the next two sections are for readers familiar with elements of complex analysis. For $x_1,\ldots,x_n$ as above, *prove* that

$$(2\pi i)^{-1} \oint_{\partial D} \frac{f(z)dz}{\prod_{k=1}^{n}(z - x_k)} = \sum_k \frac{f(x_k)}{\prod_{j\neq k}(x_k - x_j)};$$

in this formula, the function f is holomorphic in the bounded open domain with **rectifiable** boundary $D \subset \mathbb{C}$, free of self-intersections and continuous on its closure $D \cup \partial D$. D contains $x_1,\ldots,x_n$. The right-hand side of this formula, as a function of

$x_1,\ldots,x_n$, is defined in D^n minus the union of hyperplanes $\{x_i = x_j\}$. The formula shows that it can be holomorphically extended to all of D^n by defining it via the left-hand side. (*How* can one show the left-hand side is holomorphic?) In particular, the Jacobi identities $[f(z) = z^m]$ are holomorphically extended to $\mathbb{C}^n$. A short computation gives the extension to the diagonal of D^n. For $n = 2$,

$$\lim_{x_1 \to x_2} \left[\frac{f(x_1)}{x_1 - x_2} + \frac{f(x_2)}{x_2 - x_1} \right] = f'(x_2).$$

Check that, in general,

$$\lim_{x_1,\ldots,x_n \to a} \sum_i \frac{f(x_i)}{\prod_{j \neq i} (x_i - x_j)} = \frac{f^{[n-1]}(a)}{(n-1)!}.$$

P1.9[***]

Let f be holomorphic in an open disc $D \subset \mathbf{C}$ and continuous on its closure $D \cup \partial D$. Let the center of D be the origin, and let $f = \sum_{m \geq 0} a_m z^m$ be the Taylor-series expansion in D. *Show* that for any $x_1,\ldots,x_n \in D$,

$$(2\pi i)^{-1} \oint_{\partial D} \frac{f(z)dz}{\prod_{k=1}^{n} (z - x_k)} = \sum_{m \geq n-1} a_m \sum_{\sum \nu_k = m-n+1} \prod_{k=1}^{n} {x_k}^{\nu_k}.$$

(A special case of this formula, for $x_1 = \ldots = x_n$,

$$(2\pi i)^{-1} \oint_{\partial D} \frac{f(z)\, dz}{(z-x)^n} = \sum_{m \geq n-1} a_m \binom{m}{n-1} x^{m-n+1},$$

also results from differentiating the Taylor series, with regard to the Cauchy integral formula for the Taylor coefficients.)

P1.10[***]

A point x is called a **critical point** of a holomorphic function F if $F'(x) = 0$. A value of F is **critical** when some of the preimages are critical. Noncritical values are also referred to as **regular**. *Establish* the following generalization of the formula in section P1.8[***]: Let functions f and F be holomorphic in a bounded open domain

$D \subset \mathbf{C}$ with rectifiable boundary ∂D, free of self-intersections and continuous on its closure $D \cup \partial D$; then for any regular value y of F, with no preimages on ∂D,

$$(2\pi i)^{-1} \oint_{\partial D} \frac{f(z)\,dz}{F(z)-y} = \sum_j \frac{f\left(F_j^{-1}(y)\right)}{F'\left(F_j^{-1}(y)\right)};$$

in this formula, the summation is performed over all preimages of y in D. (*Prove* that y has a finite number of preimages in D.) A further development of the subject, including a multivariate version of this formula, a many-dimensional generalization of the Jacobi identities, and so on, is contained in Arnol'd et al. (1982) and references therein. For versions of this formula for ramifying f, see Pakovich et al. (2004).

P1.11[***]

It is a well-known fact that a holomorphic function is divisible by a polynomial with a remainder, the remainder being a polynomial of degree that is smaller by one than the degree of the divisor polynomial:

$$f(z) = q(z)F(z) + r(z), \quad r = \sum r_l z^l, \quad \deg\ r = \deg\ F - 1\ .$$

In this section, readers unfamiliar with holomorphic functions may consider polynomials instead of holomorphic functions f. Let F be a polynomial of degree n with simple roots $x_1, \ldots, x_n$. *Prove* that the leading coefficient of the remainder is given by

$$r_{n-1} = \sum_{i=1}^{n} \frac{f(x_i)}{\prod_{j\neq i}(x_i - x_j)}\ .$$

Generalize this formula to

$$r_{n-k} = (-1)^{k-1} \sum_{i=1}^{n} \frac{f(x_i)\ \sigma_{k-1}(x_1, \ldots \hat{x}_i \ldots, x_n)}{\prod_{j\neq i}(x_i - x_j)}, \qquad k = 1, \ldots, n,$$

where σ_k are the usual elementary symmetric polynomials in $n-1$ variables (circumflexes indicate the missing corresponding terms).

We suggest further practicing with the Jacobi identity techniques by solving the following two problems in the theory of ordinary linear differential equations.

P1.12[**]

(*This problem assumes no familiarity with differential equations.*) Let an algebraic equation $\lambda^n + a_1\lambda^{n-1} + \ldots + a_n = 0$ have simple roots $\lambda_1,\ldots,\lambda_n$. *Prove* that for any numbers $p_0,\ldots, p_{n-1}$, a function $x(t) = \sum_{i=1}^{n} \frac{c_1\lambda_i^{n-1}+\ldots+c_n}{\prod_{j\neq i}(\lambda_i-\lambda_j)} e^{\lambda_i t}$, where $c_1 = p_0$, $c_2 = p_1 + a_1p_0, \ldots, c_n = p_{n-1} + a_1p_{n-2} + \ldots + a_{n-1}p_0$, satisfies the differential equation $x^{[n]}(t) + a_1x^{[n-1]}(t) + \ldots + a_nx(t) = 0$ and initial conditions $x(0) = p_0,\ldots, x^{[n-1]}(0) = p_{n-1}$. Specifically, $x(t) = \sum_{i=1}^{n} \frac{e^{\lambda_i t}}{\prod_{j\neq i}(\lambda_i-\lambda_j)}$ corresponds to the initial conditions of $p_0 = \ldots = p_{n-2} = 0$, $p_{n-1} = 1$. (Passing to the limits in these formulas allows one to deriver similar formulas for multiple roots using linear combinations of functions of the form $e^{\lambda_i t}t^k$.)

P1.13[***]

(N. Roytvarf, personal communication (2003)). *This problem, of practical importance for radio engineering and electronics, is addressed to those readers who are familiar with the elements of the theory of ordinary linear differential equations including the direct and inverse* ***Laplace transform***. Consider a function of time that has zero value for $t < 0$ and is piecewise quasipolynomial on $t \geq 0$: $f(t) = \sum_k \theta(t - t_k) \sum_j \big(P_{kj}(t)\cos\beta_{kj}t + Q_{kj}(t)\sin\beta_{kj}t\big)e^{\alpha_{kj}t}$, where $0 = t_0 < t_1 < \ldots$, P_{kj}, Q_{kj} are polynomials and θ is the usual **Heaviside step-function**, $\theta(t) = \begin{cases} 0, & \text{for } t<0, \\ 1, & \text{for } t\geq 0. \end{cases}$

Let us consider the following initial value (or Cauchy) problem: determine a function $x(t)$ on $\{t \geq 0\}$ that is $n - 1$ times continuously differentiable, including the points 0, t_1, t_2, …, and satisfies the differential equation $x^{[n]} + a_1x^{[n-1]} + \ldots + a_nx = f$ on the segments between the points 0, t_1, t_2, …, ∞ and the initial conditions $x(0) = p_0,\ldots, x^{[n-1]}(0) = p_{n-1}$. (Replacing 0 with any other point t_0 does not make any essential changes.)

Certainly, one might deal with this problem by solving the successive Cauchy problems on the segments between the points $0, t_1, t_2, \ldots, \infty$, setting the final value of $x,\ldots, x^{[n-1]}$ on the former segment as the initial value for the latter one (which also shows the unique solvability of the problem as a whole). However, the usual algorithm by the Laplace transform automatically leads to the same solution. *Prove* it, and then you may use this remarkable algorithm in your practical work!

P1.14[***]

Final notes. The Jacobi identities are applied to various problems in geometry and mathematical physics. Readers will encounter multiple examples in Charlier (1927), Landau and Lifshitz (1973), Landau et al. (1982), Arnol'd (1989), Arnol'd et al. (1985), and references therein. Some of these examples will be discussed in the "Jacobi Elliptic Coordinates" and "Gravitation of Spheres and Ellipsoids" problem groups, in volume 2 of this book.

A good illustration of the close ties between the Jacobi identities and complex analysis is the fact that identities from section P1.1[**] are obtained via a straightforward calculation of the sums of **residues** of differential forms $\omega_m = \frac{z^m dz}{\prod (z-x_i)}$ on the Gaussian (Riemannian) sphere, taking into account that ω_n is holomorphic at infinity for $m = 0,\ldots, n-2$ and has there a pole of residue of -1 for $m = n-1$ (section S1.5). Further generalizations of the Jacobi identities can be obtained using the language and tools of (multivariate) complex analysis; these generalizations are applied within a related mathematical branch – algebraic geometry; experienced readers may consult Arnol'd et al. (1982) and references therein (pay special attention to the following works: (1) Petrovsky and Oleynik, On the topology of real algebraic manifolds (1949), (2) Khovanskii, Newton polyhedra, and the Euler–Jacobi formula (1978), and (3) Griffiths and Harris, Residues, and zero cycles on algebraic varieties (1978)). Griffiths and Harris, residues, and zero cycles on algebraic varieties (1978).

Hint

H1.1

We suggest proceeding by any of the following three methods:

(1) Algebraic method: For each m, the left-hand side of the corresponding Jacobi identities, once denominators have been cleared, becomes a polynomial of degree $max(m, n-2)$ that has $n-1$ different roots. (The same method applies to any field.)

(2) Analytic method, using implicit function theorem: The left-hand sides are proportional to the first derivatives of the functions $\sum_i x_i^{m+1}$, with respect to the constant term of the polynomial $F(z) := \prod(z - x_i)$, while the other coefficients are kept fixed (N. Roytvarf, pers. Commun.).

(3) Method using complex analysis: The Jacobi identities can be established based on the formula from section P1.8[***].

H1.2

This can be done using methods similar to Method 1 or 2 above. One can even take the same first steps as in solving problem P1.1**.

H1.3

Rearrange the left-hand side, presenting it as a polynomial in $y = (y_1,\ldots,y_{n-1})$, and then use the identities from section P1.1**.

H1.4

Verify directly that u satisfies the equations. This solution is unique if and only if the matrix of the system is nonsingular; therefore, it can be unique only if $y_1,\ldots, y_n$ are distinct. With distinct $x_1,\ldots, x_n$ and $x_i \neq y_j$, $\forall i,j$, is the distinctness of $y_1,\ldots, y_n$ also sufficient for that uniqueness?

H1.5

The first step may be the same as in the algebraic method of proving identities in section P1.1**: multiplication of the left-hand side by the Vandermonde determinant gives the polynomial

$$P_{n,m}(x) := V_n(x_1,\ldots,x_n)\cdot\sum_{i=1}^{n}\frac{x_i^m}{\prod_{j\neq i}(x_i-x_j)} = \sum_{i=1}^{n}(-1)^{n-i}{x_i}^m\cdot V_{n-1}(x_1,\ldots\hat{x}_i\ldots,x_n),$$

and we must show that

$$P_{n,m}(x) = V_n(x)\quad\cdot\sum_{\sum v_i=m-n+1}\prod_{i=1}^{n}{x_i}^{v_i}.$$

$P_{n,m}(x)$ goes to zero if $x_i = x_j$ for some $i \neq j$. Therefore, by virtue of the relative primality of the elements $x_i - x_j$ in the ring of polynomials with rational coefficients $\Theta[x]$, $P_{n,m}(x)$ is divisible by the product of those elements, that is, by $V_n(x)$. The quotient is a homogeneous and symmetric polynomial in x (as $P_{n,m}$ and V_n both are homogeneous and skew-symmetric) of degree $m - n + 1$. Moreover,

by the famous **Gauss lemma** in algebra of polynomials (Van der Waerden 1971, 1967; Lang 1965), the quotient has integral coefficients

$$P_{n,m}(x) = V_n(x) \cdot Q_{n,m}(x), \qquad Q_{n,m}(x) \in \mathbf{Z}[x].$$

Show that the polynomials $Q_{n,m}$ satisfy the following recursion equations (*discrete Cauchy problem*):

$$Q_{2,m}(x_1, x_2) = \sum_{k=0}^{m-1} {x_1}^k {x_2}^{m-1-k} \qquad (m \geq 1),$$

$$Q_{n,m}(x_1, \ldots, x_n) = \sum_{k=0}^{m-n+1} \binom{m}{k} {x_n}^k \cdot Q_{n-1,m-k-1}(x_1 - x_n, \ldots, x_{n-1} - x_n),$$

$$n>2, \ \ m \geq n-1 \ .$$

Lastly, by induction on n show that the unique solution of this discrete problem is

$$Q_{n,m}(x) = \sum_{\sum v_i = m-n+1} \prod_{i=1}^{n} {x_i}^{v_i} .$$

Readers will arrive at a shorter proof if they pay attention to the fact that the sequence of the left-hand sides of the equalities in section P1.5**, $L_{m:\ m=0,1,\ldots}$, and the similar sequence of the right-hand sides, $R_{m:\ m=0,1,\ldots}$ (complemented by $R_0 = \ldots = R_{n-2} = 0$), obey the same recursion relation for $m \geq n$: $\sum_{k=0}^{n} (-1)^k \sigma_k L_{m-k} = 0$, $\sum_{k=0}^{n} (-1)^k \sigma_k R_{m-k} = 0$ ($\sigma_0 = 1$ and, for $k > 0$, σ_k are the usual elementary symmetric polynomials in $x_1, \ldots, x_n$); we leave it to the interested reader to fill in the details. See also a proof in Givental (1984).

H1.6

This is the same identity as in section P1.2** (indeed, multiply both sides by $\prod x_i$), which makes the same approach relevant. Alternatively, identities from section P1.6** (and, hence, from section P1.2**) can be derived directly from the Jacobi identities in section P1.1**. Specifically, substituting $x_1, \ldots, x_n$ as in section P1.6**, $x_{n+1} = 0$ and $m = 1$ in these identities gives

$$\sum_{i=1}^{n} \frac{1}{x_i \cdot \prod_{j \neq i} (x_i - x_j)} + \frac{(-1)^n}{\prod_{i=1}^{n} x_i} = 0,$$

which is equivalent to that of section P1.6**.

H1.7

The answer is

$$\sum_{i=1}^{n} \frac{x_i^{-m}}{\prod_{j\neq i}(x_i - x_j)} = \frac{(-1)^{n-1} \sum \prod_{i=1}^{n} x_i^{m-1-v_i}}{\prod x_i^m}, \qquad m = 1, 2, 3, \ldots .$$

In looking for an answer, use section P1.5**, applying x_i^{-1} for x_i. (Fill in the details.)

H1.8

The Cauchy-type integral on the left-hand side of the first formula can be calculated using the **residue** technique. Alternatively, this integral can be calculated directly, replacing the contour ∂D by the union of small circles centered at x_i. Lastly, the formula being proved is a special case of a formula in section P1.10***; therefore, it can also be proved using the method suggested for solving section P1.10***. That the left-hand side is holomorphic can be deduced from its differentiability in each x_i and continuity on $(x_1,\ldots,x_n)$ Dieudonné (1960). Actually, the continuity does not require a special proof - by the fundamental Hartogs theorem. For more on this see Bochner and Martin (1948), Cartan (1961), and Hörmander (1966). The second limit formula is immediately obtained if we go over to the limit on the left-hand side and apply the Cauchy integral formula to get the Taylor coefficients.

H1.9

The required formula is immediately obtained after putting the Taylor-series expansion of f on the left-hand side of the first formula in section P1.8*** and using the identities in sections P1.1** and P1.5**.

H1.10

Replace the contour ∂D by the union of small circles centered at $F_j^{-1}(y)$, and change the variable $z \mapsto u := F(z)$ in the obtained integrals.

H1.11

Those familiar with the elements of complex analysis may deal with the formula for r_{n-1} using the first formula in section P1.8***. Substitute f on its left-hand side by $q \cdot \prod_{1 \le i \le n}(z - x_i) + r$, and apply the Cauchy integral theorem and identities from section P1.1**, taking into account that $\deg r = n - 1$. On the other hand, all r_l can be obtained from the fact that $r(z)$ is the Lagrange interpolating polynomial of degree $n - 1$, which takes values $f(x_i)$ at the points x_i:

$$r(z) = \sum_{i=1}^{n} f(x_i) \cdot \prod_{j \neq i} \frac{z - x_j}{x_i - x_j} \ .$$

H1.12

Experienced readers familiar with applications of the Laplace transform for linear differential equations may solve these initial value problems for both simple and multiple characteristic roots using this method. (The same method will enable a different approach to establish the identities in section P9.5*!) Less experienced readers may prove the formula in section P1.12** by following these steps:

1) Verify that any of the functions $e^{\lambda_i t}$ satisfy the differential equations, so any of their linear combinations satisfy them as well.
2) Apply the identities in section P1.1** for $t = 0$ to obtain that $x(0) = p_0$;.
3) Differentiate $x(t)$ at $t = 0$ and apply the identities in section P1.1**, the first identity in section P1.5** (for $m = n$), a relation $x(0) = p_0$ from the previous step, and, lastly, Vieta's formula $-a_1 = \lambda_1 + \ldots + \lambda_n$ to obtain that $x'(0) = p_1$.
4) Reason by induction: differentiating $x^{[k-1]}(t)$ at $t = 0$ and applying the identities in sections P1.1** and P1.5**, the relations $x(0) = p_0, \ldots, x^{[k-1]}(0) = p_{k-1}$ from the previous induction steps, and, lastly, Vieta's formulas for $a_1, \ldots, a_k$ will bring, after some combinatorial computations, that $x^{[k]}(0) = p_k$ $(k \le n - 1)$.

   (Fill in the details.)

H1.13

The linearity of the Laplace transform with respect to the initial conditions and the right-hand sides allows us to restrict ourselves to zero initial conditions and $f(t) = \theta(t - a)\ [p(t) \cos \beta t + q(t) \sin \beta t]\ e^{\alpha t}$. (Why?) Let, for simplicity, $f(t) = \theta(t - a) e^{\alpha t}$ $(a \ge 0)$. Assume also that the characteristic roots $\lambda_1, \ldots, \lambda_n$ are simple

and distinct from α, and include α in the set of λ_i as λ_0. Verify that applying the Laplace transform yields the solution $x(t) = \theta(t-a)e^{\alpha a} \cdot \sum_{i=0}^{n} \frac{e^{\lambda_i(t-a)}}{\prod_{j\neq i}(\lambda_i - \lambda_j)}$. Now, a direct computation using the identities from section P1.1** will show that $x(a+0) = 0, \ldots, x^{[n-1]}(a+0) = 0$. (Verify this!) **QED**. Similar arguments yield the desired result for any β, p, q, and $\lambda_1, \ldots, \lambda_n$. (We leave it to the reader to complete the details.)

Readers familiar with **generalized functions** (**distributions**) and **fundamental solutions** may prefer a different approach, as follows. The fundamental solution of the differential operator has $n-2$ continuous derivatives [which may be found by applying either identities from section P1.1** or, as in Schwartz and Huet (1961), **generalized differentiation**]. The solution corresponding to a right-hand-side f may be found by convolving the fundamental solution with f, which increases the number of continuous derivatives by one. (Fill in the details.)

Explanation

E1.1

Algebraic method in section H1.1. Multiplication by the Vandermonde determinant

$$V_n(x_1, \ldots, x_n) := \det \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \\ . & \ldots & . \\ {x_1}^{n-1} & \ldots & {x_n}^{n-1} \end{pmatrix} = \prod_{1 \le i < j \le n} (x_j - x_i)$$

yields the left-hand sides of the Jacobi identities in the form

$$P_m(x) := V_n(x_1, \ldots, x_n) \cdot \sum_{i=1}^{n} \frac{x_i^m}{\prod_{j\neq i}(x_i - x_j)} = \sum_{i=1}^{n} (-1)^{n-i} {x_i}^m \cdot V_{n-1}(x_1, \ldots \hat{x}_i \ldots, x_n)$$

(circumflexes, or "hats," mean that the corresponding term is missing). $P_m(x)$ is a polynomial with integer coefficients, of degree not exceeding max $(m, n-2)$ in any x_i. As a polynomial in one variable, say x_n, with parameters $x_1, \ldots, x_{n-1}$, it has roots $x_1, \ldots, x_{n-1}$. Hence, $P_m = 0$ for $m < n-1$.

Analytic method, using the implicit function theorem, in section H1.1. Derivatives of the functions $\sum_i {x_i}^{m+1}$ with respect to the constant term of the polynomial $F(z) = \prod(z - x_i)$, while the remaining coefficients are kept fixed, can be expressed in terms of x_i using the implicit function theorem. For $m < n-1$, those functions do not depend on the constant term of F, and therefore the derivatives are equal to zero.

Method using complex analysis in section H1.1. Denote both sides of the formula in section P1.8*** with $f(z) = z^m$ by $\varphi(x)$ [$x = (x_1,\ldots,x_n)$]. The right-hand side is homogeneous of degree $m - n + 1$ in x, $\varphi(\varepsilon x) = \varepsilon^{m-n+1}\varphi(x)$. Now calculate the limit of the left-hand side, with εx, instead of x, as $\varepsilon \to 0$, and compare the result with the same limit of the right-hand side for $m = 0,\ldots, n - 1$.

E1.3

The required representation looks like this:

$$\sum_{i=1}^{n} \frac{\prod_{k=1}^{n-1}(x_i - y_k)}{\prod_{j\neq i}(x_i - x_j)} = \sum_{i=1}^{n} \frac{x_i^{n-1}}{\prod_{j\neq i}(x_i - x_j)} + \sum_{m=1}^{n-1} (-1)^m \sigma_m(y) \cdot \sum_{i=1}^{n} \frac{x_i^{n-1-m}}{\prod_{j\neq i}(x_i - x_j)},$$

where σ_m are the usual elementary symmetric polynomials.

E1.4

Substituting u, defined in this way, into all of the indicated equations turns each equation into an identity of the type considered in section P1.3**. With distinct $x_1,\ldots, x_n$ and $x_i \neq y_j$, $\forall i,j$, the solution is unique if and only if $y_1,\ldots, y_n$ are distinct, which follows from an elementary algebraic identity

$$\det\left((x_i - y_j)^{-1}\right)_{i,j=1,\ldots,n} = \prod_{1\le i<j\le n}(x_j - x_i)(y_i - y_j) \Big/ \prod_{i,j=1}^{n}(x_i - y_j)$$

or, using $-y$ for y to restore the symmetry between arrays x and y,

$$\det\left((x_i + y_j)^{-1}\right)_{i,j=1,\ldots,n} = \prod_{1\le i<j\le n}(x_j - x_i)(y_j - y_i) \Big/ \prod_{i,j=1}^{n}(x_i + y_j)$$

(if $x_i + y_j \neq 0$, $\forall i,j$). Readers can establish this identity by multiplying columns of a matrix $((x_i + y_j)^{-1})$ by $(x_j + y_1)\ldots(x_j + y_n)$, respectively, factorizing the obtained matrix and calculating determinants of the factors and using the identity for a Vandermonde determinant from section E1.1.

Thus, the linear system from section P1.4** is uniquely solvable, and $u_1,\ldots, u_n > 0$ when either $x_n > y_n > x_{n-1} > \ldots > y_1$ or $x_1 > y_1 > x_2 > \ldots > y_n$. (Why?) For $n = 2$, geometrically this means that **confocal** (having common foci) ellipse and hyperbola intersect each other at $2^n = 4$

points $\left(\pm u_1^{1/2}, \pm u_2^{1/2}\right)$, symmetric with respect to the axes of a Cartesian coordinate system, such that one of the axes contains the foci and the origin is in the middle between them. Under fixed x_1, x_2, the smallest and the biggest of y_1, y_2 correspond, respectively, to the ellipse and the hyperbola. (Find semiaxes of these quadrics and the distance between the foci.) One can prove the conversion: given three points F_1, F_2, and A on the Euclidean plane, there exist unique pairs of ellipse and hyperbola both having F_1, F_2 as their foci and both passing via point A; these confocal quadrics intersect at right angles at each of the four points noted above. (Can the same be stated without exceptions if some or all of the points F_1, F_2, and A coincide with each other? Describe the ellipses and the hyperbolas in these cases.) The parameters y_1 and y_2 are called **Jacobi elliptic coordinates**. These coordinates are the same for the four intersection points (in case of degeneration, some of those four points may coincide). A multidimensional generalization, as described by the Jacobi theorem, and related subjects will be discussed in the "Jacobi Elliptic Coordinates" problem group (volume 2 of this book).

Coordinate systems are called elliptic if any curves of a constant coordinate value (called the constant coordinate curves) are quadrics or segments of quadrics of a fixed confocal family. In these systems, either the Jacobi coordinates or other elliptic coordinates are used, such as (r,θ), where r is the sum of semiaxes of a coordinate ellipse and θ is the slope angle of an asymptote for a corresponding half-branch of a coordinate hyperbola. Another alternative if the ellipse's foci do not coincide is to use (t,θ), where $t = \ln(r/c) = \operatorname{arch}(e^{-1})$, c is the half-distance between the two foci, and e is the **eccentricity** of the ellipse. (The eccentricity of an ellipse is the ratio of c to the length of the major semiaxis or, equivalently, the ratio of the distance between a point on the ellipse and its focus to the distance between that point and a **directrix** of the ellipse closest to this focus.) Crossing the rectilinear segment between the foci at its interior point we will observe a discontinuity of the first kind (a jump) in θ (the maximal jump, from $\mp \pi/2$ to $\pm \pi/2$, is in the middle of the segment; there are no jumps on the foci since θ is considered modulo 2π). The preceding Cartesian coordinates can be expressed as $\frac{1}{2}(r + c^2/r) \cdot \cos\theta = c \cdot \operatorname{ch} t \cdot \cos\theta$, $\frac{1}{2}(r - c^2/r) \cdot \sin\theta = c \cdot \operatorname{sh} t \cdot \sin\theta$, and the semiaxes of the coordinate ellipse and hyperbola are equal to $c \cdot \operatorname{ch} t$, $c \cdot \operatorname{sh} t$, and $c \cdot |\cos\theta|$, $c \cdot |\sin\theta|$, respectively. We leave it to interested readers to prove these formulas. (This may require some advanced technical elements, such as the **Joukowsky map** that is usually taught in courses on complex analysis, which will be discussed in the "Jacobi Elliptic Coordinates" problem group.) Readers should not confuse elliptic coordinates (r,θ) with generalized polar coordinates for which the constant coordinate curves are homothetic concentric ellipses and rectilinear rays having their vertex at the ellipses' common center; here, r is a parameter that has the dimensionality of length and is proportional to the dilatation coefficient of the ellipse, and θ is the polar angle of the ray. These are not orthogonal coordinates unless the foci coincide. (Why?) The Cartesian coordinates corresponding to such generalized polar coordinates can be expressed as $ar \cdot \cos\theta$, $br \cdot \sin\theta$, where a, b are dimensionless positive constants proportional to the ellipses' axes. A common limit case of the elliptic and the generalized polar coordinates (r,θ) in the case of coinciding foci are the usual polar coordinates.

E1.5

The first of the recursion equations (for $Q_{2,m}$) is derived directly. To derive the second one algebraically, put $P_{n,m}(x) = V_n(x) \cdot Q_{n,m}(x)$ into the formula that defines $P_{n,m}$ and replace arguments $x_1, \ldots, x_n$ by $x_1 - x_n, \ldots, 0$, respectively. We will obtain the required equation if we take into account that the Vandermonde determinant is not affected by this replacement (why?) and then apply the identities from section P1.1** (for $n \geq 3$).

Those familiar with elements of complex analysis can arrive at those recursion equations by a shorter route, expressing the left-hand side of the identities in section P1.5** via the Cauchy integral, in accordance with the first formula in section P1.8*** ($f(z) = z^m$), and then substituting z by $z + x_n$ and applying the identities from section P1.1**:

$$\begin{aligned} Q_{n,m}(x) &= \sum_{k=1}^{n} \frac{{x_k}^m}{\prod_{j\neq k}(x_k - x_j)} = \frac{1}{2\pi i} \oint_{\partial D} \frac{z^m dz}{\prod_{k=1}^{n}(z - x_k)} = \frac{1}{2\pi i} \oint_{\partial D - x_n} \frac{(z + x_n)^m dz}{\prod_{k=1}^{n}(z - (x_k - x_n))} \\ &= \frac{1}{2\pi i} \sum_{l=0}^{m-n+1} \binom{m}{l} x_n^l \oint_{\partial D - x_n} \frac{z^{m-l-1} dz}{\prod_{k=1}^{n}(z - (x_k - x_n))} \\ &= \sum_{l=0}^{m-n+1} \binom{m}{l} x_n^l \, Q_{n-1,m-l-1}(x_1 - x_n, \ldots, x_{n-1} - x_n) \, . \end{aligned}$$

We suggest solving these recursion equations using multi-index notations. A multi-index is an ordered set of natural numbers (including zero) of a fixed (finite) dimension $\alpha = (\alpha_1, \ldots, \alpha_l)$; its norm is $|\alpha| := \sum \alpha_i$; for α, β of the same dimension, $\alpha \geq \beta$ if this inequality holds for each component; furthermore, for a complex vector $y = (y_1, \ldots, y_l)$, $y^\alpha := \prod y_i^{\alpha_i}$; we will also use the notation $\binom{\alpha}{\beta} := \prod \binom{\alpha_i}{\beta_i}$. Denote by $\sigma_{\alpha,k}(y)$ a polynomial in l variables $y_1, \ldots, y_l$, obtained from an elementary symmetric polynomial in $|\alpha|$ variables $\sigma_k(t_1, \ldots, t_{|\alpha|})$ by the substitution

$$t_1 = \ldots = t_{\alpha_1} = y_1, \quad \ldots \quad , \quad t_{\alpha_{l-1}+1} = \ldots = t_{\alpha_l} = y_l \, .$$

By the inductive assumption,

$$\begin{aligned} Q_{n-1,n+k-2}(x_1 - x_n, \ldots, x_{n-1} - x_n) &= \sum_{|\alpha|=k} \prod_{j=1}^{n-1} (x_j - x_n)^{\alpha_j} = (-1)^k \sum_{|\alpha|=k} \prod_{j=1}^{n-1} (x_n - x_j)^{\alpha_j} \\ &= \sum_{|\alpha|=k} \sum_{s=0}^{k} (-x_n)^{k-s} \sigma_{\alpha,s}(x_1, \ldots, x_{n-1}) \\ &= \sum_{s=0}^{k} (-x_n)^{k-s} \sum_{|\alpha|=k} \sigma_{\alpha,s}(x_1, \ldots, x_{n-1}) \, . \end{aligned}$$

Putting this into the second of the recursion equations and substituting k by $m - n - k + 1$ yields

$$Q_{n,m}(x) = \sum_{s=0}^{m-n+1} (-1)^s x_n^{m-n-s+1} \sum_{s \le |\alpha| \le m-n+1} (-1)^{|\alpha|} \binom{m}{|\alpha|+n-1} \cdot \sigma_{\alpha,s}(x_1, \ldots, x_{n-1}).$$

Therefore, the desired formula $Q_{n,m}(x) = \sum_{|\alpha|=m-n+1} x^\alpha$ is equivalent to the identities

$$\sum_{|\beta| \le m} (-1)^{|\beta|} \binom{m+n}{|\beta|+n} \cdot \sigma_{\beta,k}(x) = (-1)^k \sum_{|\alpha|=k} x^\alpha, \qquad \forall n \ge 1,$$
$$\forall x = (x_1, \ldots, x_n), \ \forall m \ge k,$$

where the sum on the left-hand side is taken over the multi-indices of dimension n with $|\beta| \le m$, and on the right-hand side over the multi-indices of the same dimension and norm k. (Why?) To proceed, consider any multi-index α of dimension n and norm k. The addend x^α from the right-hand side is encountered $\binom{\beta}{\alpha}$ times in $\sigma_{\beta,k}(x)$ (as it is equal to the number of ways that you can simultaneously select α_1 elements from β_1, …, α_n elements from β_n). Thus, these identities are equivalent to the identities

$$\sum_{\beta \ge \alpha,\ |\beta| \le |\alpha|+k} (-1)^{|\beta|-|\alpha|} \binom{|\alpha|+k+n}{|\beta|+n} \binom{\beta}{\alpha} = 1, \quad \forall k \ge 0,\ \forall n > 0,\ \forall \alpha \text{ of dim } n,$$

that are discussed in the problem group "Some Problems in Combinatorics and Analysis That Can Be Explored Using Generating Functions" below. Applying them finalizes the proof.

E1.8

The partial fraction decomposition of the integrand yields

$$\frac{f(z)dz}{\prod_{i=1}^{n}(z-x_i)} = \sum_{i=1}^{n} \frac{c_i}{z-x_i} + O(1), \qquad \text{with} \quad c_i = \frac{f(x_i)}{\prod_{j \ne i}(x_i - x_j)}.$$

Having done this, apply the **residue** formula. Alternatively, one may directly apply the Cauchy integral formula to the function $f_i(z) := f(z)/\prod_{j \ne i}(z - x_j)$, which is holomorphic inside a small disc centered at x_i.

E1.10

By the Cauchy theorem, $\oint\limits_{\partial D} \frac{f(z)\,dz}{F(z)-y} = \sum \oint\limits_{\partial D_j} \frac{f(z)\,dz}{F(z)-y}$, where D_j are nonintersecting disks in D with the centers ${F_j}^{-1}(y)$. The restrictions of F to D_j will be diffeomorphisms if the radii are sufficiently small (by the inverse function theorem). Change the variable $z \mapsto u := F(z)$ in the integrals, and apply the Cauchy integral formula.

Completing the Solution

S1.1

Algebraic method in section H1.1. As we have already seen, for $m < n - 1$, $P_m(x)$ is the zero polynomial and, therefore, the zero function, which proves the Jacobi identities in this case. $P_{n-1}(x)$ is equal to $\prod_{i<n}(x_n - x_i)$ multiplied by the leading coefficient of $P_{n-1}(x)$,

$$P_{n-1}(x) = V_{n-1}(x_1, \ldots, x_{n-1}) \cdot \prod_{i<n} (x_n - x_i) = V_n(x_1, \ldots, x_n)\ .$$

[The equality $P_{n-1}(x) = V_n(x_1,\ldots,\ x_n)$ is also deduced by decomposing the determinant V_n with respect to its last row.] This completes the proof.

For readers familiar with the matrix calculus, we can introduce another version of the same proof. The Vandermonde matrix (which will also be denoted by V) is invertible for mutually different $x_1,\ldots,x_n$, $V \circ V^{-1} = E$:

$$\begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \\ \vdots & & \vdots \\ x_1^{n-1} & \cdots & x_n^{n-1} \end{pmatrix} \cdot \begin{pmatrix} * & \prod\limits_{j\neq 1} (x_1 - x_j)^{-1} \\ * & \cdots \\ * & \prod\limits_{j\neq n} (x_n - x_j)^{-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix};$$

in this identity the elements of the last column in V^{-1} are calculated directly by formulas for the inverse matrix entries, and the remaining elements, which do not affect the proof, are marked by asterisks. This matrix identity is equivalent to the n^2 scalar identities

$$\sum_k V_{ik} V_{kj}^{-1} = \delta_{ij}, \qquad i,j = 1,\ldots,n,$$

and among them, n identities corresponding to $i = 1,\ldots,n$, $j = n$ are the Jacobi identities.

Analytic method, using the implicit function theorem, in section H1.1. Let z_0 be a simple root of a polynomial equation

$$0 = F(z) = z^n + a_1 z^{n-1} + \ldots + a_n,$$

$F'(z_0) \neq 0$. By the implicit function theorem, the equation

$$0 = z^n + a_1 z^{n-1} + \ldots + a_n + u$$

with fixed $a_1,\ldots,a_n$ defines z as a unique smooth function of u such that $z(0) = z_0$ in a small neighborhood of $u = 0$. Its first derivative at $u = 0$ is

$$\frac{\partial z}{\partial a_n}\Big|_{a_1,\ldots,a_{n-1}=const} = \frac{dz}{du}\Big|_{u=0} = -\frac{1}{F'(z_0)}\ .$$

Similarly, if a_i are coefficients of the polynomial $F(z) = \prod(z - x_i)$, then the derivatives of $\sum_i x_i^{m+1}$ with respect to a_n will be

$$\frac{\partial \sum_i x_i^{m+1}}{\partial a_n}\Big|_{a_1,\ldots,a_{n-1}=const} = -(m+1)\sum_i \frac{x_i^m}{F'(x_i)} = -(m+1)\sum_i \frac{x_i^m}{\prod_{j\neq i}(x_i - x_j)},$$

and thus they will be proportional to the left-hand sides of the identities that are being proved. The next step consists in the calculation of those derivatives. For $m < n - 1$, $\sum_i x_i^{m+1}$ are functions of $a_1,\ldots,a_{n-1}$ (by the symmetric polynomial theorem); therefore, they have zero partial derivatives with respect to a_n. Furthermore, by the same symmetric polynomial theorem, there exists a polynomial p in n variables such that

$$\sum_i x_i^n = p(a_1, \ldots, a_n),$$

and evidently p depends on the last variable linearly and does not have a constant term,

$$p = p_1(a_1, \ldots, a_{n-1}) + ca_n, \qquad p_1(0, \ldots, 0) = 0.$$

It remains to prove that $c = -n$. For this, set $a_1 = \ldots = a_{n-1} = 0$ and $a_n = 1$. In this case, the polynomial $F(z)$ will be $z^n + 1$, and $x_1,\ldots,x_n$ will be the roots of -1 of degree n, so that

$$-n = \sum_i x_i^n = p(0, \ldots, 0, 1) = ca_n = c,$$

which completes the proof.

Method using complex analysis in section H1.1 (Arnol'd et al. 1982). The holomorphy, and thus continuity, and thus boundedness, of $\varphi(\varepsilon x)$ near the origin implies that for $m < n - 1$, $\varphi(x) = 0$ [otherwise, $\varphi(\varepsilon x)$ would grow as ε^{m-n+1}]. For $m = n - 1$, $\varphi(x)$ is homogeneous of degree 0, that is, $\varphi(\varepsilon x)$ does not depend on ε. In particular, it is equal to the limit of the left-hand side of the formula in section P1.8*** when $\varepsilon \to 0$. This limit is a classic Cauchy integral, expressing the residue of $1/z$ at the origin and equal to the increase in the value of the logarithm as its argument goes once around the origin:

$$\lim_{\varepsilon\to 0} \oint_{\partial D} \frac{z^{n-1}\,dz}{\prod_{k=1}^{n}(z-\varepsilon x_k)} = \oint_{\partial D} \frac{dz}{z} = \oint_{\partial D} d\ln z = 2\pi i,$$

which completes the proof.

Other arguments using complex analysis are discussed in section S1.5 below.

S1.2

Using the algebraic method as in section H1.1, take a polynomial

$$P(x) := V_n(x_1,\ldots,x_n)\cdot\sum_{i=1}^{n}\prod_{j\neq i}\frac{x_j}{x_j-x_i} = \sum_{i=1}^{n}(-1)^{i-1}\,V_{n-1}(x_1,\ldots\,\hat{x}_i\ldots,x_n)\cdot\prod_{j\neq i}x_j$$

(which is considered a polynomial in one variable x_n, with parameters $x_1,\ldots,x_{n-1}$). Similarly to section S1.1, multiplication of the polynomial $\prod_{i<n}(x_n - x_i)$ by the ratio of the constant terms of $P(x)$ and itself will be equal to $P(x)$:

$$P(x) = V_{n-1}(x_1,\ldots,x_{n-1})\cdot\prod_{i}(x_n - x_i) = V_n(x_1,\ldots,x_n),$$

which completes the proof. Using the analytic method as in section H1.1, one will have

$$(-1)^n = \frac{\partial\prod_{i=1}^{n}x_i}{\partial a_n}\bigg|_{a_1,\ldots,a_{n-1}=const} = -\sum_{i}\frac{\prod_{j\neq i}x_j}{F'(x_i)} = -\sum_{i}\prod_{j\neq i}\frac{x_j}{x_i-x_j},$$

which completes the proof.

S1.3

One immediately obtains from this representation, taking into account the identities in section P1.1**, which correspond to $m = 0,\ldots,n-2$, an identity

$$\sum_{i=1}^{n} \frac{\prod_{k=1}^{n-1}(x_i - y_k)}{\prod_{j\neq i}(x_i - x_j)} = \sum_{i=1}^{n} \frac{x_i^{n-1}}{\prod_{j\neq i}(x_i - x_j)}, \qquad (*)$$

and application of the last identity in section P1.1** completes the proof.

Note on identities in section P1.1:** the last identity can be derived from the first $n-1$ ones. Indeed, the identity (*) was derived from the first $n-1$ identities from section P1.1**, and it holds for all $y_1,\ldots,y_{n-1}$. Setting those variables to, respectively, $x_1,\ldots,x_{n-1}$ turns the first $n-1$ terms in the sum on the left-hand side of (*) into zero, and the last term into 1. Thus, the right-hand side of (*) must be equal to 1.

S1.4

Establishing the identity from section E1.4. Multiplication of the columns of the matrix $((x_i + y_j)^{-1})$ by $(x_j + y_1)\ldots(x_j + y_n)$, respectively, brings a matrix with elements $f_i(y_j)$, $f_i(y) = \prod_{k\neq i}(y + x_k) = \sum_{m=0}^{n-1} \sigma_m(x_1,\ldots,\widehat{x_i},\ldots,x_n)\, y^{n-1-m}$ [σ_m are the usual elementary symmetric polynomials ($\sigma_0 \equiv 1$), and the hat means that the corresponding term is missing]. The matrix $(f_i(y_j))$ is a product of a Vandermonde matrix (y_j^{i-1}) and a matrix with the entries $\sigma_{n-j}(x_1,\ldots,\widehat{x_i},\ldots,x_n)$ $(i,j = 1,\ldots,n)$. The determinant of the Vandermonde matrix is equal to $\prod_{1\le i<j\le n}(y_j - y_i)$. The determinant of the second matrix depends only on $x_1,\ldots, x_n$, and not on $y_1,\ldots, y_n$, and so it is proportional (with a numerical coefficient) to the product $\prod_{1\le i<j\le n}(x_j - x_i)$ since $\det(f_i(y_j))$ is symmetric with respect to arrays x and y and since

$$\varphi(x)\cdot\psi(y) \equiv \varphi(y)\cdot\psi(x) \quad \Leftrightarrow \quad \varphi(x)/\psi(x) \equiv const \equiv \varphi(y)/\psi(y).$$

The coefficient of proportionality equals one because both of the proportional polynomials under consideration contain the monomial term $\prod x_i^{i-1}$ with coefficient equal to 1.

Alternatively, readers could prove the relation $\det\left(\sigma_{n-j}(x_1,\ldots,\widehat{x_i},\ldots,x_n)\right) \propto \prod_{1\le i<j\le n}(x_j - x_i)$ by the following reasoning. The determinant on the left-hand side is a homogeneous polynomial of degree $0 + \ldots + (n-1) = \binom{n}{2}$ in $x_1,\ldots, x_n$, becoming zero if $x_i = x_j$ for some $i \neq j$. Therefore,

arguments identical to those in section S3.2 ("A Combinatorial Algorithm in Multiexponential Analysis" problem group below) show that this determinant is proportional to $\prod_{1 \le i < j \le n} (x_j - x_i)$. (Fill in the details; note that a similar method can be used to calculate the Vandermonde determinant.)

S1.5

The final step of the proof uses the implication $\sum k_\alpha x^\alpha = 0 \Rightarrow k_\alpha = 0, \forall \alpha$, expressing the linear independence of the monomial functions x^α of fixed dim α and $|\alpha|$ (in the vector space of dimension dim α over an infinite field). In other words, a homogeneous polynomial with coefficients from an infinite field defines a zero function (if and) only if all the coefficients are zero. Readers may establish it themselves or look through section E12.39 (proof of lemma 1) from the problem group "Least Squares and Chebyshev Systems" below.[1] [In addition, readers may generalize it by proving the linear independence of the monomial functions x^α with different α, or, in other words, by proving that a polynomial with coefficients from an infinite field defines a zero function (if and) only if all the coefficients are zero. This can be proved using quite elementary means (Lang 1965; Van der Waerden 1971, 1967).]

This paragraph is addressed toward readers with advanced knowledge of complex analysis. According to the formula in section P1.8***, the left-hand sides of the Jacobi identities (and their extensions for other integers m) are the sums of residues of meromorphic differential forms $\omega_m = \frac{z^m dz}{\prod (z - x_i)}$, respectively, at the points $x_1, \ldots, x_n$ on the Gaussian (Riemannian) sphere $\overline{\mathbf{C}}$. (Fill in the details.) Since the sum of all residues in $\overline{\mathbf{C}}$ is zero (because of the compactness of $\overline{\mathbf{C}}$), the sum of residues at $x_1, \ldots, x_n$ must be equal to the minus residue at infinity. A change in the variable $u = z^{-1}$ shows that for $m \le n - 2$ ω_m have no poles at infinity, so we have zero residues there (complete the details), which provides the corresponding Jacobi identities with one more proof. For $m \ge n - 1$, ω_m do have poles at infinity, and finding the residues also provides the corresponding identities with one more proof. Calculating the residue is quite routine for $m = n - 1$ (we recommend that readers fill in all the details), but for $m > n - 1$ combinatorial considerations similar to those discussed previously are required.

S1.8

To find the coefficients c_k of the partial fraction decomposition, one can multiply both parts of this formula, with indeterminate c_l, by $\prod(z - x_l)$ and then insert $z := x_k$, which will complete the proof. Alternatively, finding the integral on the left-hand side by direct computation using the Cauchy integral formula one will have

[1] The proof in section E12.39 is done for the complex field $\mathbb{C}$. However, it is applicable for any infinite field because a finitely generated extension of the rational field is embeddable in $\mathbb{C}$. (Readers should complete all details.)

$$\frac{f(x_k)}{\prod\limits_{j\neq k}(x_i-x_j)} = f_k(x_k) = (2\pi i)^{-1} \oint\limits_{\partial D} \frac{f_k(z)\,dz}{z-x_k} = (2\pi i)^{-1} \oint\limits_{\partial D} \frac{f(z)\,dz}{\prod\limits_{l=1}^{n}(z-x_l)}\,.$$

QED

S1.10

In accordance with the change-of-variables rule and the Cauchy integral formula,

$$\frac{1}{2\pi i}\sum_j \oint\limits_{\partial D_j} \frac{f(z)\,dz}{F(z)-y} = \frac{1}{2\pi i}\sum_j \oint\limits_{F\left(\partial D_j\right)} \frac{f\left(F_j^{-1}(u)\right)\,du}{F'\left(F_j^{-1}(u)\right)\cdot(u-y)} = \sum_j \frac{f\left(F_j^{-1}(y)\right)}{F'\left(F_j^{-1}(y)\right)}.$$

S1.13

The fundamental solution may be characterized as having a form $\theta(t)\cdot x_0(t)$, where $x_0(t)$ is the usual (smooth) solution on $-\infty < t < \infty$ of the homogeneous Cauchy problem $x^{[n]}(t) + a_1x^{[n-1]}(t) + \ldots + a_nx(t) = 0$, $x(0) = 0,\ldots, x^{[n-2]}(0) = 0$, $x^{[n-1]}(0) = 1$. Convolving it with the right-hand side f produces the same solution as applying the Laplace transform, which will possess $n-1$ continuous derivatives. (We leave it to the reader to complete the details.[2])

The preceding algorithm produces a solution with similar smoothness properties for a substantially wider class of the functions f on the right-hand side of the equation in section P1.13***. For example, this is so for all f that are piecewise continuously differentiable, equal to zero on $\{t < 0\}$,

[2] The fundamental solution possesses $n-2$ continuous derivatives, so convolving it with the right-hand side brings $n-1$ continuous derivatives, which may be established proceeding step by step as follows:

1. Let the functions f and g be identically equal to zero on, respectively, $\{t < 0\}$ and $\{t < a\}$ and continuous on, respectively, $\{t \geq 0\}$ and $\{t \geq a\}$, where $a \geq 0$; show that their convolution $h = f^*g$ equals zero identically on $t < a$ and $h(t) = \int\limits_a^t f(t-s)g(s)\,ds = \int\limits_0^{t-a} f(s)g(t-s)\,ds$ for $t \geq a$.
2. In addition, let g be continuously differentiable on $\{t \geq a\}$; prove that h is continuously differentiable on $\{t \geq a\}$, with $h'(t) = g(a+0)f(t-a) + \int\limits_a^t f(t-s)g'(s)\,ds$. (Use integration by parts.)
3. Show by induction on m that if $f(t)$ is an $m \geq 0$ times continuously differentiable function with $f(0) = \ldots = f^{[m]}(0) = 0$, then the function $F(t) = \begin{cases} \int\limits_a^t f(t-s)g(s)\,ds, & \text{for } t \geq a, \\ 0, & \text{for } t \end{cases}$, with g continuous on $\{t \geq a\}$, is m times continuously differentiable.
4. Combining the results of steps 2 and 3 (applied for g' in place of g), deduce by induction that if, under the conditions of step 2, f is m times continuously differentiable on $(-\infty,\infty)$ [so that $f(0) = \ldots = f^{[m]}(0) = 0$] and g is continuously differentiable on $\{t \geq a\}$, then f^*g is $m+1$ time continuously differentiable.

and grow at an exponential, or smaller, rate on $\{t \geq 0\}$. [The last feature means that $f(t) \leq Ae^{kt}$, with appropriate constants $A \geq 0$ and k.] This can be proved using the fundamental solution. (Provide the details.)

In turn, establishing the indicated characteristic property of the fundamental solution by the generalized differentiation [as, for instance, in Schwartz and Huet (1961)], which does not employ the Jacobi identities, and, on the other hand, obtaining this solution explicitly by the Laplace transform [which brings $x_0(t) = \sum_{i=1}^{n} \frac{e^{\lambda_i t}}{\prod_{j \neq i} (\lambda_i - \lambda_j)}$ for simple characteristic roots $\lambda_1, \ldots, \lambda_n$, in accordance with section P1.12**] provides the Jacobi identities in section P1.1**. with one more proof! (We leave it to the reader to work out the details.)

A Property of Recursive Sequences

Problems

P2.0

Preliminaries. A sequence $a_0,a_1,\ldots$ is called *recursive* if there exist a natural number k and a function f of k variables such that $a_{n+k} = f(a_n,\ldots,a_{n+k-1})$ for any n. (The preceding equation is called the *recursion relation*.) *Linear* recursive sequences [when f is linear, $a_{n+k} = m_1a_{n+k-1} + \ldots + m_ka_n$ ($m_k \neq 0$)] are the most important of all recursive sequences. Numerical linear recursive sequences are discrete analogs of solutions of linear autonomous ordinary differential equations, and they are used in applications of combinatorial analysis. Linear recursive sequences with polynomial elements (when a_n and m_i are polynomials), such as the well-known Chebyshev polynomials, also find multiple applications.

This short problem group is devoted to a beautiful elementary problem about numerical linear recursive sequences that was suggested by A.D. Sakharov. We open the group with Sakharov's original formulation of the problem which refers to the famous Fibonacci sequence (section P2.1*) and then ask readers to find its generalizations (sections P2.2* and P2.3*). We would like to emphasize that the problems of this group assume no familiarity with combinatorial analysis and can be solved by readers without experience.[1]

[1] Some other aspects of numerical linear recursive sequences are discussed in this problem book in sections S7.9 (the problem group "2 × 2 Matrices That Are Roots of Unity" below) and P11.0 (the problem group "Several Problems in Combinatorial Theory and Analysis That Are Explored with Generating Functions" below), and will be further discussed in the "Fibonacci Numbers, Continued Fractions, and Pell Equation" problem group in volume 2. Linear recursive sequences with polynomial elements related to Chebyshev polynomials are considered in sections P12.30*** and H12.30 (the "Least Squares and Chebyshev Systems" problem group below). To learn more about linear recursive sequences and their applications, consult the references from the aforementioned sections.

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_2,

P2.1[*]

(Sakharov 1996). *Prove* that the Fibonacci sequence $a_0, a_1, \ldots, a_n, \ldots$ defined by

$$a_0 = 1, \qquad a_1 = 1, \qquad a_{n+2} = a_{n+1} + a_n, \qquad n = 0, 1, \ldots$$

contains multiples of all integers.

P2.2[*]

Prove a stronger result – that the Fibonacci sequence contains infinitely many multiples of any integer; these multiples appear in the foregoing Fibonacci sequence with a definite period. (Of course, the shortest period lengths may be different for different integers.)

P2.3[*]

The Fibonacci recursion relation is not special. *Find* a wider class of recursive sequences with the foregoing property using the same method.

Hint

H2.2 and H2.3

In this section, we will call **appropriate** a sequence satisfying a polynomial recursion relation with integral coefficients. We suggest dividing the whole problem P2.2[*] and P2.3[*] into three smaller problems as follows.

(a) Prove that the remainders, modulo any integer, of the members of an appropriate sequence form a periodic sequence.
(b) The periodicity in (a), is, generally, "mixed" (periods do not necessarily start from the very beginning). Find a sufficient condition for a "pure" (starting from the very beginning) repetition of remainders from (a).
(c) For an appropriate sequence with "purely" repeated remainders, find a sufficient condition for zero being one of the remainders.

Explanation

E2.2 and E2.3

(a) "Mixed" repetition of remainders follows from recursion of their sequence, for the appropriate source sequence,[2] and the application of Dirichlet's box principle (also known as the pigeonhole principle) to the finite rings of remainders.
(b) The periodicity in (a) will always be "pure" when the appropriate source sequence is uniquely extendable, from any place, in the backward direction, with the extension being appropriate, more precisely, when there exists a polynomial g in m variables with integral coefficients such that $a_n = g(a_{n+m},\ldots,a_{n+1})$, $\forall n > 0$ (m may differ from k). A simple sufficient condition consists, of course, in requiring that the recursion formula $f(a_n,\ldots,a_{n+k-1})$ be linear in a_{n+k-1} and that its coefficient be ± 1.
(c) For the appropriate source sequence satisfying (b), an obvious sufficient condition for inclusion of zero in the set of remainders is that zero is included in the source sequence itself or in its backward extension.

Completing Solution

(a) For an appropriate sequence defined by a recursion relation $a_{n+k} = f(a_n,\ldots, a_{n+k-1})$, $N^k + 1$ successive sets, each consisting of k successive remainders

$$R_0 = \{a_n \bmod N, \ldots, a_{n+k-1} \bmod N\}, \quad \ldots,$$
$$R_{N^k} = \{a_{n+N^k} \bmod N, \ldots, a_{n+N^k+k-1} \bmod N\}$$

cannot all be different. On the other hand, every R_i uniquely determines all "future" sets $R_{i+1}, R_{i+2}, \ldots$. Therefore, if, say, $R_i = R_{i+p}$, then the sets

$$\{a_{n+j} \bmod N, \ldots, a_{n+j+p-1} \bmod N\}, \qquad \forall j \geq i \quad \text{and} \quad \forall n > 0$$

[2] If $a_{n+k} = f(a_n,\ldots,a_{n+k-1})$ for any n, where f is a polynomial with integral coefficients, the remainders $a_n,\ldots,a_{n+k-1} \bmod N$ determine the remainder $a_{n+k} \bmod N$ as $f\,(a_n \bmod N, \ldots, a_{n+k-1} \bmod N) \bmod N$.

form periods of the remainders modulo N. One should emphasize that periods appear not later than after N^k–1 steps.

(b) Let lengths of some periods of remainders of the source sequence and its backward extension be p, q, respectively. Consider the backward extension starting in a very distant "future." The required "pure" repetition of remainders of the source sequence will take place because the periods of remainders of the backward extension of length, say, pq, will be "antiperiods" for remainders of the source sequence.

A Combinatorial Algorithm in Multiexponential Analysis

Problems

P3.0

Introduction. This group contains real-life problems that arose while developing algorithms for deciphering, or multiexponential analysis of signals in a nuclear magnetic resonance (NMR) machine, which is necessary for a composite analysis (e.g., for separating oil fractions in well logging, detecting affected tissue in medicine). NMR signals are modeled by functions of the form $f(x) = \sum A_i e^{\omega_i x}$, which are linear combinations of exponential terms; in these f, the parameters ω_i correspond to different components, and A_i characterize their relative weights in the composite.[1] The number and value of the parameters A_i and ω_i, are not known, and so the multiexponential analysis consists in determining them (within some tolerance). This analysis relates to so-called **inverse problems**, usually characterized by a high degree of instability. This problem group contains real-life algebraic and combinatorial problems that arose in the development of a stable algorithm for multiexponential analysis (Itskovich et al. 1996, 1998).

P3.1[**]

Taking three successive elements f_k, f_{k+1}, f_{k+2} of any geometric progression $f_l = a \cdot q^{l-1}$ one has

[1] Readers interested in theoretical fundamentals and application-oriented aspects of NMR may refer to Abragam (1961) and references therein.

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_3,

$$\det\begin{pmatrix} f_k & f_{k+1} \\ f_{k+1} & f_{k+2} \end{pmatrix} = f_k \cdot f_{k+2} - {f_{k+1}}^2 = 0 \,.$$

Prove the generalization for $2n - 1$-element sequences. Let a sequence $f_1,\ldots,$ f_{2n-1} be a sum of m geometric progressions, $f_k = \sum_{1\le i\le m} a_i{\cdot}{q_i}^{k-1}$ ($q_i \ne q_j$ for $i \ne j$). Then the determinant

$$D(f) \;:= \det\begin{pmatrix} f_1 & \ldots & f_n \\ . & \ldots & . \\ f_n & \ldots & f_{2n-1} \end{pmatrix}$$

always equals zero for $m < n$; further, $D(f) \ne 0$ if all progressions are nonzero and

- either $m = n$
- or $m > n$, the progressions are real-valued and all a_i are either positive or negative simultaneously.

P3.2**

Show that, in fact, for $m \ge n$

$$D(f) = \sum_{1\le i_1<\ldots<i_n\le m} \prod_{v=1}^{n} a_{i_v} \cdot \prod_{1\le\mu<v\le n} \left(q_{i_v} - q_{i_\mu}\right)^2 .$$

P3.3**

Based on the results of section P3.1**, *develop* an algorithm for resolving the inverse problem of multiexponential analysis, formulated as follows.

Consider a function $y = f(x)$, which is a linear combination with positive coefficients of real-valued exponential terms,[2]

[2] An exponential function is a continuous version of a discrete geometric progression: for $q > 0$, $a{\cdot}q^{k-1} = A{\cdot}e^{\omega k}$, $\forall k{\in}\mathbb{Z}$, where $q = e^{\omega}$ and $a = A{\cdot}e^{\omega}$, or, equivalently, $\omega = \ln q$ and $A = a/q$. (Actually, this holds for any nonzero complex q, using a fixed complex branch of $\ln q$, but readers not experienced with complex numbers may consider only the case where q is a positive number.)

$$y = \sum_{i=1}^{m} A_i e^{\omega_i x}, \qquad \omega_i \in \mathbf{R}, \quad A_i > 0.$$

Assume that the values of y are known for any arguments x, but the number of exponential terms m as well as the "amplitudes" A_i and "frequencies" ω_i are unknown. *Find* m and *evaluate* (within some tolerance) A_i and ω_i.[3]

Hint

H3.1

The decomposition of the matrix into m summands corresponding to one-progression summands of the sequence f

$$\begin{pmatrix} f_1 & \dots & f_n \\ . & \dots & . \\ f_n & \dots & f_{2n-1} \end{pmatrix} = \sum_{i=1}^{m} \begin{pmatrix} f_1^i & \dots & f_n^i \\ . & \dots & . \\ f_n^i & \dots & f_{2n-1}^i \end{pmatrix}, \qquad \text{where} \quad f_k^i = a_i q_i^{k-1},$$

shows that the rank of this matrix cannot exceed m. (Why?) Therefore, $D(f) = 0$ for $m < n$ (Radzivilovsky, pers. Commun.). The sufficient conditions for $D(f) \neq 0$ can be derived from the explicit formula in section P3.2**.

H3.2

Simultaneous decomposition of $D(f)$ with respect to its rows and columns yields its representation as a linear combination of Vandermonde determinants of size n. Then convert that linear combination into the formula in section P3.2**.

H3.3

Let $f^{(n)}$ be the sequence of values $F(x_1),\ldots,F(x_{2n-1})$ on some $2n - 1$-point set

[3] In practice, the function f is known on a finite set of points ("time echoes" in NMR terminology) and, moreover, known to have an error. Usually, the error is random noise, which is assumed to be **normally distributed** (but a **systematic error** can also exist). Therefore, the correct formulation of the problem includes arguments of probability theory (Itskovich et al. 1996, 1998). We advise, however, ignoring the probabilistic aspect at this point and considering initially the resulting deterministic problem.

$$x_1, \quad x_2 = x_1 + h, \qquad \ldots, \qquad x_{2n-1} = x_1 + (2n-2)h.$$

The number of exponential terms, m, coincides with the index of the last nonzero element in a sequence $D(f^{(1)})$, $D(f^{(2)})$, ... The same holds for $f^{(n)}$, defined as integrals of $F(x)$ over $2n - 1$ sequential intervals of equal lengths. (Why?)[4] Evaluation (within a tolerance) of A_i and ω_i requires a combination of the results of section P3.1** with some other technique. Start from an observation that the value of the "longest" exponential term in the sum, corresponding to the maximal "frequency" ω_m, will dominate the total sum value for large arguments x. This means that for large x the total relative contribution of all the remaining terms is negligibly small. (The reader can make the appropriate estimates!) Therefore, the longest term may be identified, considering a domain restricted to large x, and using some optimization algorithm to determine A_m and ω_m (a commonly preferred algorithm is a linear regression with respect to log A_m and ω_m). This term is subtracted, and the procedure is iterated until the required level of approximation is achieved, within the tolerance. Unfortunately, inaccurate determination of a lower boundary for the "large x" domain leads to a fast error accumulation that can make the proposed method impractical. The results of section P3.1** proved to be fairly good in establishing proper lower boundaries for x in this method. Readers are asked to formulate such P3.1**-based lower boundaries for those domains where only the longest exponential term can be distinguished, where only the two longest terms can be distinguished, and so on.

Explanation

E3.1

The one-progression matrix summands have proportional rows; hence, they are of rank 1. On the other hand, the rank of a sum of matrices does not exceed the sum of their ranks. This can be proved using the formula for the dimension of the sum of subspaces of a vector space, $\dim(A_1 + A_2) = \dim A_1 + \dim A_2 - \dim(A_1 \cap A_2)$.[5]

[4] The use of integrals is much more practical since it decreases the impact of uncertainties related to the assignment of F.

[5] A generalization for m subspaces is the **principle of inclusion–exclusion**

$$\dim \sum A_i = \sum \dim A_i - \sum_{i<j} \dim (A_i \cap A_j) + \ldots + (-1)^{m-1} \dim \bigcap A_i.$$

E3.2

The desired representation of $D(f)$ as a linear combination of Vandermonde determinants is

$$D(f) = \sum_{1 \le i_1 < \ldots < i_n \le m} \left(\prod_{v=1}^{n} a_{i_v} \right) \cdot \sum_{\sigma \in S_n} \left[V_n \left(q_{i_{\sigma(1)}}, \ldots, q_{i_{\sigma(n)}} \right) \cdot \prod_{v=1}^{n} q_{i_{\sigma(v)}}^{v-1} \right],$$

where the inner sum is taken over all members σ of the group of permutations of an n-element set (the **symmetric group**), S_n. Prove the identities

$$\sum_{\sigma \in S_n} \left[V_n \left(q_{\sigma(1)}, \ldots, q_{\sigma(n)} \right) \cdot \prod_{v=1}^{n} q_{\sigma(v)}^{v-1} \right] = \prod_{1 \le i \le n} \left(q_j - q_i \right)^2$$

using the fact that the polynomials in $q_1, \ldots, q_n$ that appear on the left- and right-hand sides are homogeneous of equal degrees, symmetric, divisible by all $q_i - q_j$, and have equal coefficients at some monomial.

E3.3

We may assume that an $F(x)$ data series decays at infinity (otherwise multiply it by a proper decaying exponential function). Let x_∞ be a very large argument's value (at the end of the data list, in a practical situation). The goal is to define the lower bounds for domains in which only the longest exponential term can be distinguished, only the two longest terms can be distinguished, and so on. Starting from $x_1 = x_\infty$ decrease x_1 while $D(f^{(1)}) = 0$, within a predefined tolerance ε_1, where $f^{(1)}$ is the sequence of integrals of $F(x)$ over three sequential intervals of equal lengths[6]

$$[x_1, x_1 + h], \quad [x_1 + h, x_1 + 2h], \quad [x_1 + 2h, x_\infty], \quad h = \tfrac{x_\infty - x_1}{3}.$$

Iterating this procedure, start from $x_2 = x_1$ and decrease x_2 while $D(f^{(2)}) = 0$, within a predefined tolerance ε_2, where $f^{(2)}$ is the sequence of integrals of $F(x)$ over five sequential intervals of equal lengths

[6] Tolerances $\varepsilon_1, \varepsilon_2, \ldots$ can be derived by probabilistic methods from standard deviations of $F(x)$ (Itskovich et al. 1996). We advise, however, ignoring this subject upon first acquaintance.

$$[x_2,\ x_2+h],\quad [x_2+h,\ x_2+2h],\quad \ldots,\quad [x_2+4h,\ x_\infty],\qquad h=\tfrac{x_\infty-x_2}{5}\ .$$

And so on; the desired domains are, respectively, $[x_n, x_\infty]$.

Completing the Solution

S3.1

Proof that the rank of a sum of matrices does not exceed the sum of their ranks. Associate with matrices $\mathrm{M}_1, \ldots$ some linear operators M_i: $A \to A$. (These operators define the corresponding matrices if any fixed basis is selected in the vector space A.) For an operator's images A_i we will have dim A_i = rk M_i. But since im $\sum M_i \subseteq \sum A_i$ (why?), we will obtain

$$\mathrm{rk}\sum \mathrm{M}_i = \dim\ \mathrm{im}\sum M_i \le \dim\sum A_i \le \sum \dim\ A_i = \sum \mathrm{rk}\,\mathrm{M}_i.$$

QED.

S3.2

Rings of polynomials with coefficients from factorial rings (unitary commutative integral domains with the unique factorization property) themselves are factorial. [Readers may prove this themselves or find a proof in Lang (1965) or Van der Waerden (1971, 1967).] Hence, the rings R_n; $= \mathbb{Z}[q_1,\ldots, q_n]$ are factorial, and, since the quotient of R_n modulo its ideal spanned on q_n is isomorphic to R_{n-1}, monomials q_i are prime elements in R_n. (Furnish the details.) Therefore, a linear, over $\mathbb{Z}$, variable substitution $q_1, q_2, \ldots, q_n \mapsto q_1,\ q_2 - q_1, \ldots,\ q_n - q_1$ shows that $q_j - q_i$ with $i \ne j$ are prime elements in R_n. (Fill in the details.) They are distinct primes, in the sense that they are not divisible by one another, so they generate distinct prime ideals. (Why?) A polynomial $p(q_1,\ldots,q_n) \in R_n$ is symmetric with respect to q_n and q_1, in other words, is even with respect to a variable $q_n - q_1$; therefore, $p = (q_n - q_1)^2 \cdot p'$, where p' is a polynomial in $(q_n - q_1)^2$, with coefficients from R_{n-1}. [Use the fact that an even polynomial in x is a polynomial in x^2, $p(q) = p_1(x^2)$, and an odd polynomial has the form $x \cdot p_1(x^2)$,[7] which follows from the presentation as a

[7] A similar principle holds for the formal power series, and smooth even (odd) functions have even (resp. odd) Taylor series at the origin (even when those series diverge). Indeed, all derivatives of a smooth even (odd) function that have odd (resp. even) orders vanish at the origin. (Why?)

sum of its even and odd components.[8]] Hence, the left-hand side of the equality that is being proved is divisible by $\prod_i \left(q_j - q_i\right)^2$ in the ring R_n, in other words, the quotient must have integral coefficients! (Why?) Actually, the quotient is a number (a zero-degree polynomial) that is found by comparing the degrees of the dividend and the divisor. Finally, this quotient is determined by comparing the coefficients at monomials $\prod {q_i}^{2(i-1)}$ in the dividend and the divisor. (We leave it to the reader to provide the details.)

[8] For a function $f(x)$, this decomposition is $f(x) = \frac{f(x)+f(-x)}{2} + \frac{f(x)-f(-x)}{2}$ and is unique. Advanced readers know that the even component is f's average over a group of order 2, with the generator inverting the argument's sign. An analogous decomposition into symmetric and skew-symmetric components, $f(x_1, x_2) = \frac{f(x_1,x_2)+f(x_2,x_1)}{2} + \frac{f(x_1,x_2)-f(x_2,x_1)}{2}$, corresponds to a group with the generator permuting the arguments (for n arguments, this group has order $n!$). A generalization of this concept brings analogous averaging formulas for any finite group of symmetries and, furthermore, integral formulas corresponding to any compact group of symmetries. To practice, readers can prove the following direct generalizations of the above claim about even polynomials:

- *A polynomial on $\mathbb{R}^n$ is invariant with respect to a reflection in the origin if and only if it contains terms of even degrees only.*
- *A polynomial on $\mathbb{R}^n$ is invariant with respect to a group generated by reflections in the coordinate hyperplanes if and only if it is a polynomial in ${x_1}^2, \ldots, {x_n}^2$.*
- *A polynomial on $\mathbb{R}^n$ is invariant with respect to an orthogonal group if and only if it is a polynomial in $|x|^2 = \sum {x_i}^2$ (with numerical coefficients).*

(The similar claims hold for formal power series.) Readers may state and prove similar claims themselves using other groups of symmetries. Also, prove the following statement generalizing the fact that the sum of the coefficients of a polynomial p equals $p(1)$:

- *For a polynomial (or a power series of convergence radius exceeding 1) $p(x) = a_0 + a_1 x + \ldots$ of one variable with complex coefficients, the sum of coefficients $a_0 + a_m + a_{2m} + \ldots$ equals $m^{-1} \sum_{k=0}^{m-1} p(\xi^k)$, where ξ is a primitive root of degree m of 1.*

A Frequently Encountered Determinant

Problems

P4.0[*]

The following matrix appears in various problems connected with equidistants and envelopes in analysis, geometry, calculation of variations, and mathematical physics:

$$A = \begin{pmatrix} 1 + {a_1}^2 & a_1 a_2 & \dots & a_1 a_n \\ a_2 a_1 & 1 + {a_2}^2 & \dots & a_2 a_n \\ . & . & \dots & . \\ a_n a_1 & . & \dots & 1 + {a_n}^2 \end{pmatrix}$$

(i.e., $A_{ij} = \delta_{ij} + a_i a_j$). A calculation of determinants of this matrix and matrices of a more general form (sections P4.1[**] and P4.2[**]) is a nice exercise in linear algebra that does not require advanced knowledge on the part of readers (however, readers possessing such knowledge who are interested in applications will find in section P4.3[***] a typical application example and a brief follow-up discussion).

P4.1[**]

Show that the matrix from section P4.0[*] cannot degenerate, provided that a_i are real numbers; in fact, $\det A = 1 + \sum {a_i}^2$.

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_4, © Springer Science+Business Media, LLC 2013

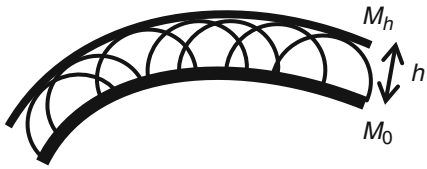


Fig. 1 An equidistant hypersurface

P4.2[**]

Show that for a matrix of a more general form, with entries $A_{ij} = c_i\delta_{ij} + a_ib_j$, $\det A = \prod c_i + \sum a_ib_i \prod_{j\neq i} c_j$. [In particular, for $c_1 = \ldots = c_n$ $\det A = {c_1}^{n-1}(c_1 + \sum a_ib_i)$.]

P4.3.[***]

For readers *familiar with differential geometry of hypersurfaces in Euclidean spaces* we can describe in more detail a typical situation when the matrix from section P4.0[*] appears. Consider a smooth (twice continuously differentiable) closed (compact and without boundary) hypersurface M_0 in a Euclidean space. A hypersurface M_h is referred to as **equidistant** from M_0 by h if the minimal distance of any point of M_h from the points of M_0 equals h. An elementary proof of the smoothness of M_h for small h uses the nonsingularity of the matrix from section P4.1[**]. M_h is an **envelope** of a family of spheres of radius h with centers at all points of M_0. (This means that M_h is tangent to each sphere; see Fig. 1.)

Locally, M_0 may, without loss of generality, be described by $x_n = F(x_1,\ldots, x_{n-1})$, so the spheres of the family are defined by the equation $(x_1 - {x'}_1)^2 + \ldots + (x_{n-1} - {x'}_{n-1})^2 + (x_n - F({x'}_1, \ldots, {x'}_{n-1}))^2 = h^2$ with free parameters ${x'}_1, \ldots, {x'}_{n-1}$. An equation for the envelope M_h can be derived via elimination of the free parameters from a system of n equations consisting of the aforementioned equation and its derivatives with respect to the free parameters. The elimination process uses the implicit function theorem, which can be applied because an $(n-1)\times(n-1)$ matrix of entries $\delta_{ij} + \frac{\partial F}{\partial x'_i}\cdot\frac{\partial F}{\partial x'_j} + (F - x_n)\cdot\frac{\partial^2 F}{\partial x'_i\partial x'_j} = \delta_{ij} + \frac{\partial F}{\partial x'_i}\cdot\frac{\partial F}{\partial x'_j} + O(h)\,(i,j = 1,\ldots,n-1)$ is nonsingular. (We leave it to the interested reader to furnish the details.)

Surfaces M_h (also called **fronts**) are smooth for small h but with growing h acquire singularities called **caustic points**. Those points also are known as **focal points** or the **centers of curvature**. (Readers are encouraged to draw the fronts of a planar ellipse for $-\infty < h < \infty$, where negative h corresponds to a different orientation of the distance with respect to the ellipse, to visualize the caustic points. *How* many topological metamorphoses do you count? See Fig. 2.) Equivalently, the caustic points may be described (for any Riemannian variety)

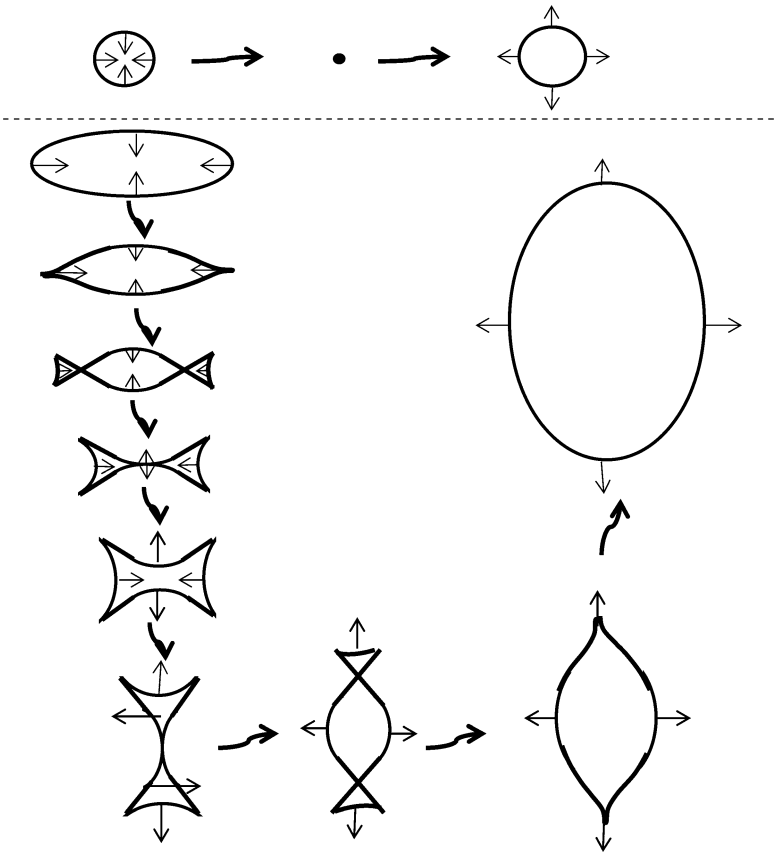


Fig. 2 The family of an ellipse's equidistants

as the points x for which the function on M_0 defined as $\left| x - y \right|^2$, $y \in M_0$ (the squared distance from x to the points of M_0) has a degenerate critical point (*which* one?). Every wave propagation process obeying **Huygens' principle** produces fronts with similar singularities. A full description of these features is far beyond the scope of the present book. It is made by methods usually incorporated into advanced branches of mathematics such as differential equations, calculus of variations, Morse theory, Hamiltonian mechanics, geometrical optics, and singularity theory. M_h are described as the level sets for a function $h = S(x)$, which is known as the **Hamiltonian action**, satisfying the **eikonal equation** $\left| \nabla S \right|^2 = 1$ (also referred to as a **Hamilton-Jacobi equation**). Readers will find detailed discussions and further developments of the topic in Milnor (1963), Arnol'd (1978, 1989), and Arnol'd et al. (1982, 1985) and multiple references therein.

Hint

H4.1

One can calculate det A by decomposing it into a sum of two addends, according to the decomposition of its first column

$$\begin{pmatrix} 1+{a_1}^2 \\ a_1a_2 \\ . \\ a_1a_n \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ . \\ 0 \end{pmatrix} + \begin{pmatrix} {a_1}^2 \\ a_1a_2 \\ . \\ a_1a_n \end{pmatrix}.$$

A nicer method of calculation is based on first finding the eigenvalues of matrix A.

H4.2

The same combinatorial method that was used in section H4.1 is also applicable here. For $c_1 = \ldots = c_n$, one also can find and use the eigenvalues.

Explanation

E4.1

The first addend in the sum given in section H4.1 is, by induction, equal to $1+\sum_{i>1}{a_i}^2$. The second one is equal to ${a_1}^2$. As for the eigenvalues of A, the reader is encouraged to calculate them for small n (say, $n = 2,3$), to formulate the result for the general case, and then try to prove it (mathematicians often find the answers in this way).

Completing the Solution

S4.1

The desired decomposition is

$$\det A = \det \begin{pmatrix} 1 & a_1a_2 & a_1a_3 & \dots & a_1a_n \\ 0 & 1+{a_2}^2 & a_2a_3 & \dots & a_1a_n \\ . & . & . & \dots & . \\ 0 & a_na_2 & a_na_3 & \dots & 1+{a_n}^2 \end{pmatrix} + \det \begin{pmatrix} {a_1}^2 & a_1a_2 & \dots \\ a_2a_1 & 1+{a_2}^2 & \dots \\ . & . & \dots \\ a_na_1 & a_na_2 & \dots \end{pmatrix} .$$

Carrying out ${a_1}^2$ from the first row and column of the second addend and then subtracting the first row, multiplied by a_i, from the ith row for all i one obtains

$$\det \begin{pmatrix} {a_1}^2 & a_1a_2 & \dots \\ a_2a_1 & 1+{a_2}^2 & \dots \\ . & . & \dots \\ a_na_1 & a_na_2 & \dots \end{pmatrix} = {a_1}^2 \det \begin{pmatrix} 1 & a_2 & \dots \\ a_2 & 1+{a_2}^2 & \dots \\ . & . & \dots \\ a_n & a_na_2 & \dots \end{pmatrix}$$

$$= {a_1}^2 \det \begin{pmatrix} 1 & a_2 & \dots & a_n \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = {a_1}^2,$$

which completes the combinatorial proof. The eigenvalues of matrix A are, respectively, $1 + \sum {a_i}^2$ and 1, with multiplicity $n - 1$, which can be proved as follows. Direct computation shows that ${}^t(a_1,\ldots,a_n)$ is an eigenvector, with the eigenvalue $1 + \sum {a_i}^2$. Searching for eigenvectors ${}^t(x_1,\ldots,x_n)$ with eigenvalue 1 leads to the system of equations

$$a_i \cdot \sum_{j=1}^{n} a_j x_j = 0, \qquad i = 1, \ldots, n \, .$$

Thus ${}^t(x_1,\ldots,x_n)$ is in the orthogonal complement of ${}^t(a_1,\ldots,a_n)$; hence, skipping a trivial case of $a_1 = \ldots = a_n = 0$, A has 1 as an eigenvalue, with multiplicity $n - 1$.

A Dynamical System with a Strange Attractor

Problems

P5.0

Preliminaries. This problem group contains real-life problems that can arise in the analysis of the dynamic behavior of a gradient algorithm for an iterative reconstruction in computerized tomography (CT) introduced in Lange and Fessler (1995). In most practical cases, the gradient algorithm apparently converges. However, the theory cannot exclude situations where its behavior becomes very complicated, even chaotic, depending on the parameters. We cannot launch into a complete theoretical investigation, being restricted to the scope of the present problem book; however, even in a simplified situation the gradient algorithm may show some striking behavior! The simplified situation is related to the recursion formula

$$x_0 > 0, \qquad x_{n+1} = x_n \cdot e^{-(x_n - \xi)}$$

($\xi \geq 0$ is the constant parameter). The asymptotic behavior of the sequence x_n is governed by the parameter ξ, as described in detail in sections P5.1*, P5.2, P5.3**, and P5.4***.

CT is widely used for nondestructive testing in medical diagnostic and industrial technology. The objects are reconstructed from their radiographs by special algorithms. The radiographs show absorption of the object's material with respect to the radiation since an observed photon count is the difference between expected and absorbed ones. The expected photon count is determined by the source of radiation. The absorbed photon count depends on the amount of matter on the photon's tracks; hence, that count depends on the track lengths and on the object's density distribution that must be reconstructed. In a practical model, the distribution takes a finite number of values $\mu = (\mu^i)$ related to the cells of a fixed finite partition of the space, and the radiation beam is quantized by a finite number of rays, so that the matrix of the lengths of a photon's tracks inside the cells $l = (l_j) = (l_j^i)$ (the length l_j^i corresponds to the jth radiograph and the ith spatial cell,

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_5,

respectively) can be determined before the reconstruction.[1] The gradient iterative algorithm corresponds to the recursion formula $\mu_{n+1} = A\mu_n$, where A is a nonlinear operator defined by $(A\mu)^i = \mu^i \cdot \frac{\sum_j I_j e^{-\langle l_j \cdot \mu \rangle} l_j^i}{\sum_j P_j l_j^i}$; in this expression, I_j and P_j are, respectively, the expected and observed photon counts in the jth radiograph, and $\langle l_j, \mu \rangle = \sum_i l_j^i \mu^i$. The simplified situation being considered here is related to the reconstruction of a one-cell distribution (#{i} = 1) from one radiograph (#{j} = 1) and, thus, is related to the previous recursive sequence in x-space using a dimensionless variable $x = l \cdot \mu$ and a constant parameter $\xi = \ln(I/P)$. [The parameter $I/P \geq 1$ characterizes the absorption of the material with respect to the radiation. An exponential dependence of the absorption on the length (for a monochromatic radiation) is due to a physical law; interested readers may learn about this subject in Feynman et al. (1963).]

P5.1[*]

Prove that if $0 < \xi \leq 2$, then x_n, starting from any initial point $x_0 > 0$, converges to a **fixed point** (which is always $x = \xi$). *Describe* the differences in the character of convergence when $0 < \xi \leq 1$, $1 < \xi < 2$, and $\xi = 2$.

P5.2

Definitions. A finite ordered set $C = \{x^{(0)}, \ldots, x^{(k-1)}\}$ is referred to as a **cycle** if the equality $x_n = x^{(i)}$ for some n and i implies that $x_{n+m} = x^{((i+m) \bmod k)}$, $\forall m = 0,1,\ldots$. (Therefore, the cycles are invariant sets, and the fixed points form 1-point cycles.) A cycle C is called (**asymptotically**) **stable** when the sequence x_n (as defined in section P5.0) converges to C if it starts in some fixed neighborhood of C.

[1] A systematic introduction to CT is beyond the scope of this problem book. Readers interested in learning more about its theoretical foundations (which are considered in integral geometry) and application-oriented aspects may consult Gelfand et al. (1959–1962), Smith et al. (1977), Herman (1980), Helgason (1980, 1984), Herman et al. (1987), and Kak and Slaney (1988) and references therein.

P5.3[**]

The result of section P5.1[*] guarantees stability (even global) of the fixed point $x = \xi$ when $0 < \xi \leq 2$. *Prove* that when ξ exceeds $\xi_0 = 2$, the fixed point will become unstable, but a stable 2-point cycle C_1 will appear: the elements of C_1 will split off from the fixed point, moving apart to a distance approximately proportional to the square root of the supercriticality $\varepsilon = \xi - 2$, provided that ε is small.

P5.4[***]

This section is addressed to readers familiar with the terminology of bifurcation theory. The described change in the asymptotic behavior of the sequences x_n in section P5.0, when $\xi_0 = 2$, is referred to as a **bifurcation of doubling of a cycle** (Arnol'd 1978). A famous **theorem by Feigenbaum** (Feigenbaum 1978, 1979) enables us to finalize the description of the asymptotic behavior. As ξ grows, an infinite series of **bifurcations of doubling of a cycle** will take place: when ξ exceeds a certain value ξ_n (with $\xi_0 = 2$), the 2^n-point cycle C_n will become unstable, but the stable 2^{n+1}-point cycle C_{n+1} will appear. The elements of C_{n+1} will split off from the elements of C_n in pairs, moving apart to a distance approximately proportional to the square root of the **supercriticality** $\varepsilon = \xi - \xi_n$, provided that ε is small. The increasing sequence $\xi_{n\,:\,n\,=\,0,1,\ldots}$ has a finite limit ξ_∞ (with exponentially decaying, as $n \to \infty$, differences $\xi_\infty - \xi_n$). When $\xi = \xi_\infty$ the periodicity will be replaced by a strange attractor. The attractor is an invariant subset of $[0,\infty)$ and consists of sequences exponentially scattering one from another. Topologically, the attractor is a **Cantor discontinuum**.

In 1943, a similar scenario was suggested by L.D. Landau as a model for arising turbulence in a laminar fluid flow (Landau and Lifshitz 1986, and references therein). The similar (to that in section P5.0) recursion formula (using the parameter $-\infty < \xi < \infty$) was in 1974–1976 applied in ecology for modeling reproduction processes taking into account competition. In particular, the series of bifurcations of doubling of a cycle was discovered by means of numerical investigation. Interested readers may find a discussion of this topic and references in the 2nd edition of Arnol'd (1978) (in Russian, Арнольд В.И. Геометрические методы в теории обыкновенных дифференциальных уравнений. Regular and Chaotic Dynamics, MCNMO, VMK NMU, 1999, or in its English translation).

Hint

H5.1

Consider the map

$$f = f_\xi : \quad [0,\infty) \to [0,\infty), \qquad \text{where} \qquad f_\xi(x) = x \cdot \mathrm{e}^{\xi - x}.$$

Show that the fixed points are $x = 0$, ξ. As a result, a sequence

$$\{x_n\} : \qquad x_0 > 0, \qquad x_{n+1} = f_\xi(x_n)$$

will converge to 0 or to ξ if it converges at all. (Why?) Draw a graph of f_ξ. [What are $f_\xi(0)$**,** the maximum, the inflection, and the asymptotic behavior while $x \to \infty$?] Indicate the intersections of this graph with a graph of the identity map $id_{[0,\infty)}$: $x \mapsto x$ in the following cases: **(a)** $\xi = 0$, **(b)** $0 < \xi < 1$, **(c)** $\xi = 1$, **(d)** $1 < \xi < 2$, and **(e)** $\xi = 2$. Show that:

- **In cases (a)–(c)**, starting from any $x_0 \geq 0$, x_1 will be confined to the segment [0,1], and x_{n+1} with $n \geq 1$ will be between ξ and x_n, so that x_n monotonically converge to ξ starting from any $x_0 > 0$ when $\xi > 0$ (why?);
- **In cases (d) and (e)**, starting from any $x_0 > 0$, x_n will sooner or later be confined to the segment $[1, e^{\xi - 1}]$, and consecutive terms will lie on different sides of ξ (why?).

Next, use the Lagrange intermediate value theorem to show that an estimate of the form

$$|f_\xi(x) - f_\xi(x')| \quad \leq \quad q \cdot |x - x'|,$$

with q depending only on ξ, holds:

- for x, x' next to ξ, with $q < 1$, when $0 < \xi \leq 1$;
- for x, $x' \geq 1$, with $q < 1$, when $1 < \xi < 2$;
- for x, $x' \geq 1$, with $q \leq 1$, when $\xi = 2$.

Deduce from this that when $0 < \xi < 2$, x_n converges no slower than at the rate of a geometric progression:

$$|x_{n+2} - x_{n+1}| = |f_\xi(x_{n+1}) - f_\xi\,(x_n)| \quad \leq \quad q \cdot |x_{n+1} - x_n|, \quad \forall n > n_0 \qquad (q<1)$$

(where n_0 depends on ξ and the starting point x_0). In particular, x_n also converge to ξ **in case (d)**. (Why?) If in the penultimate inequality x' tends to ξ, then we will find that x_n converges exponentially:

$$|x_{n+1} - \xi| = |f_\xi(x_n) - \xi| \quad \leq \quad q \cdot |x_n - \xi|, \qquad \forall n > n_0 \qquad (q < 1).$$

In the case of $\xi = 2$ the same reasoning provides a weaker recursive estimate,

$$|x_{n+1} - 2| = |f_2(x_n) - 2| \quad \leq \quad q \cdot |x_n - 2|, \qquad \forall n > n_0,$$

so that more ideas must be drawn upon for proving the convergence $x_n \to 2$. (And, of course, this convergence is slower than exponential.) Following Arnol'd (1978) we suggest considering the square of the map f_ξ, $f_\xi \circ f_\xi$, that is, the superposition with itself

$$f_\xi(f_\xi(x)) = x \cdot e^{2\xi - x\cdot(1+e^{\xi - x})} \qquad (\xi > 0),$$

because the fixed points of this map either are fixed points of f_ξ itself or, in pairs, form 2-point cycles of f_ξ (see the definition in section P5.2). The fixed points of $f_\xi \circ f_\xi$ in $(0,\infty)$ are defined by the equation

$$2\xi = x \cdot (1 + e^{\xi - x}).$$

Establish that this equation has in $(0,\infty)$ a single root when $0 < \xi \leq 2$, which is simple for $0 < \xi < 2$ and triple for $\xi = 2$, and three distinct simple roots when $\xi > 2$. (Create the figures. Do those roots correspond to the fixed points of f_ξ or its 2-point cycles? Remember that the root $x = \xi$ corresponds to the fixed points of f_ξ for any ξ.) Thus, $f_2 \circ f_2$ has the unique fixed point $x = 2$. From this and the preceding recursive estimate of the difference $x_n - 2$ deduce that subsequences of x_n of odd and even indices converge to the point $x = 2$ (they both become monotone for large indices).

H5.3

We have already found (section H5.1) that f_ξ possesses the (unique) 2-point cycle for $\xi > 2$. We have to show that this cycle is stable and also estimate its growth rate. Following Arnol'd (1978) we suggest proceeding using the **Poincaré method of normal forms**. Use the Taylor-series expansion of f_ξ at $x = \xi$. Keeping to the notations x, y in place of, respectively, $x - \xi$ and $y - \xi$ and denoting $\varepsilon = \xi - 2$, the function $y = f_\xi(x)$ is determined by the equation

$$y = (x + \varepsilon) \cdot e^{-x} - (\varepsilon + 2) = -(1 + \varepsilon) \cdot x + \tfrac{\varepsilon}{2} \cdot x^2 + \tfrac{1-\varepsilon}{6} \cdot x^3 + O(x^4).$$

Remove the term with x^2 on the right-hand side with the help of smooth local-near $x = 0$, $y = 0$, $\varepsilon = 0$ - changes in variables in the preimage and image spaces having the form $x \mapsto x + a_2(\varepsilon) \cdot x^2$, $y \mapsto y + a_2(\varepsilon) \cdot y^2$. Show that there is only one

way of doing this, namely, by setting $a_2 = \frac{\varepsilon}{2(1+\varepsilon)(2+\varepsilon)}$, and that the change brings the defining f_ξ equation to[2]

$$y = -(1+\varepsilon)\cdot x + c(\varepsilon)\cdot x^3 + O\big(x^4\big), \qquad \text{with} \quad c = \tfrac{1-\varepsilon}{6} + \tfrac{\varepsilon^2}{2(1+\varepsilon)(2+\varepsilon)}.$$

Note that near $\varepsilon = 0$, a_2 and c are smooth and $c > 0$.

A term of order $2n$ in the preceding expansion of f_ξ is removed by a similar change $x \mapsto x + a_{2n}(\varepsilon)\cdot x^{2n}$, $y \mapsto y + a_{2n}(\varepsilon)\cdot y^{2n}$, with a smooth, in the neighborhood of zero, a_{2n}.[3] The terms of odd orders (greater than 1) could be removed by similar changes, but with discontinuous, at $\varepsilon = 0$, a_{2n-1}. The distinction is that odd-order terms are **resonant** at $\varepsilon = 0$. These results arise from the famous **Poincaré normal form theorem** (Arnol'd 1978, and references therein).

Since $c > 0$, there exists a smooth, in ε, dilatation $x \mapsto k(\varepsilon)\cdot x$ (which exactly?) that turns a coefficient c at x^3 in this equation into 1. Therefore, for the notations x, y for the independent and dependent variables after the dilatation, the equation will become

$$y = -(1+\varepsilon)\cdot x + x^3 + O(x^4).$$

Check that in those variables the map $f_\xi \circ f_\xi$ is determined by the equation

$$z = [1 + 2\varepsilon(1+\ldots)]\cdot x - (2+\ldots)\cdot x^3 + O\big(x^4\big),$$

where the dots replace the terms of the order of ε, so that the fixed points are determined from the equation

$$z - x = 0 \qquad \Leftrightarrow \qquad 2\varepsilon(1+\ldots)\cdot x - (2+\ldots)\cdot x^3 + O(x^4) = 0.$$

A root $x = 0$ corresponds to the fixed point of f_ξ. The fixed points of $f_\xi \circ f_\xi$, which are points of the 2-point cycle of f_ξ, correspond to the roots of the equation obtained as a result of division by x,

$$\varepsilon = (1+\ldots)\cdot x^2 + O\big(x^3\big),$$

which has a solution with respect to x $x_\pm = \pm\sqrt{\varepsilon} + O(\varepsilon)$. (Why?) The derivatives with respect to x at those fixed points are

$$\frac{\partial z}{\partial x}(x_\pm(\varepsilon)) = 1 + 2\varepsilon - (6 + O(\varepsilon))\cdot x_\pm^2 + O\big(\varepsilon^2\big) = 1 - 4\varepsilon + O\Big(\varepsilon^{\frac{3}{2}}\Big).$$

Show that in the source coordinates (before all changes), $x_\pm$ are determined via ε in a similar, up to proportionality, way, $x_\pm = \pm k\sqrt{\varepsilon} + O(\varepsilon)$ $(k>0)$, and also that

[2] With the preceding notations x, y for the new coordinates in the preimage and image spaces.

[3] We will not perform those changes for $n>1$.

the derivatives of $f_\xi \circ f_\xi$ at those points remain unchanged under all preceding changes in the variables. Lastly, prove that the stability of the 2-point cycle is guaranteed when the derivatives of $f_\xi \circ f_\xi$ at the cycle's points are, by absolute value, less than 1. This will complete the solution of section P5.3**.

Explanation

E5.1

The range of values of f_ξ lies in the half-segment

$$(0, \max f_\xi] \quad = \quad (0,\ e^{\xi-1}] \quad \subseteq \quad (0, 1].$$

For $1 < \xi \leq 2$, not all of x_n can be confined to (0,1], otherwise this sequence would grow to a fixed point from this half-segment (why?), whereas $\xi > 1$ is the unique fixed point of f_ξ in $(0,\infty)$. Show that $f_\xi([1, e^{\xi-1}]) \subseteq [1, e^{\xi-1}]$ for $1 \leq \xi \leq 2$ (in fact, for $1 \leq \xi \leq 2.2564\ldots$, as the end values are the roots of the equation $2\xi = 1 + e^{\xi-1}$). For $1 \leq \xi \leq 2$, hence, x_n will be confined to $[1, e^{\xi-1}]$ after entering this segment. Next, the Lagrange intermediate value theorem yields the estimate

$$|f_\xi(x) - f_\xi(x')| \quad \leq \quad q \cdot |x - x'|, \qquad \text{with} \quad q = \max_{\theta \in [x,x']} |f'_\xi(\theta)|.$$

Show that q from the foregoing inequality has the following property:

- $q < 1$ for $0 < \xi \leq 1$ and θ close to ξ, and
- $q \leq e^{\xi-2}$ for $1 \leq \xi \leq 2$ and $\theta \geq 1$, so that for $\theta \geq 1$: $q < 1$ when $1 \leq \xi < 2$ and $q \leq 1$ when $\xi = 2$.

To find the number of fixed points of $f_\xi \circ f_\xi$, determine the inflection tangent's slope and the inflection value of the function $y = x \cdot (1 + e^{\xi - x}) - 2\xi$ depending on parameter ξ. [The number of roots of the equation $y = 0$ changes in the same way as in a one-parameter family $y = x^3 - (\xi - 2) \cdot x$. Create figures!]

E5.3

Substituting x, y into the defining f_ξ equation by, respectively, $x + a_2x^2$, $y + a_2y^2$ yields the equation

$$y + a_2y^2 = -(1+\varepsilon) \cdot x + \left[-(1+\varepsilon)a_2 + \tfrac{\varepsilon}{2}\right] \cdot x^2 + \left(\tfrac{1-\varepsilon}{6} + \varepsilon\, a_2\right) \cdot x^3 + O\left(x^4\right).$$

Next, find $a_2(\varepsilon)$ such that the solution does not contain a term with x^2, that is, it is in the form

$$y = -(1+\varepsilon)\cdot x + c(\varepsilon)\,.x^3 + O\big(x^4\big).$$

Substituting it in place of y in the equation defines $a_2(\varepsilon)$ and $c(\varepsilon)$, as required. **QED**.

Next, the dilatation coefficient must be $k = \pm\sqrt{c}$. To estimate the growth rate of the 2-point cycle for f_ξ, write down an equation, $\varepsilon = x^2[(1+O(\varepsilon)) + O(x)]$, that determines the cycle in terms of a local, near $x = 0$, $\varepsilon = 0$, variable $x' := x \times \sqrt{(1+O(\varepsilon)) + O(x)}$. We will obtain $\varepsilon = (x')^2$, or $x'_\pm = \pm\sqrt{\varepsilon}$. Returning to the variable x, which is expressed through x' and ε as $x = (1+O(\varepsilon))\cdot x' + O\Big((x')^2\Big)$, brings $x_\pm = \pm\sqrt{\varepsilon} + O(\varepsilon)$. **QED**.

To prove that the formula $x_\pm = \pm k\sqrt{\varepsilon} + O(\varepsilon) \quad (k>0)$ remains, up to a value of k, invariant under all changes in the preceding coordinates, write down an equation defining the map $f_\xi \circ f_\xi$ in the form $F(x,y,\varepsilon) = 0$. Then the fixed points are determined by the equation

$$\Phi(x,\ \varepsilon) : = F(x,x,\varepsilon) = 0.$$

A local change $x \mapsto \varphi_\varepsilon(x), \quad y \mapsto \varphi_\varepsilon(y)$, with φ smooth in x and ε, that has the form $\varphi_\varepsilon(x) = [k_0 + O(\varepsilon)]\cdot x + O(x^2)$ $(k_0 \neq 0)$ brings this equation to

$$\Phi(\varphi_\varepsilon(x),\ \varepsilon) \quad (= \quad F(\varphi_\varepsilon(x),\ \varphi_\varepsilon(x),\ \varepsilon)) = 0,$$

hence the transformation φ maps the solution of the first of the equations onto the solution of the second one. As we have already found, a fixed set is a union of two smooth curves defined as, respectively, $x = 0$ and $\varepsilon = x^2\,[1 + O(\varepsilon)] + \ldots$ after the last changes in variables. Using the source variables, these curves will be defined as, respectively,

$$\varphi_\varepsilon^{-1}(x) = 0 \Leftrightarrow x = 0 \quad \text{and} \quad \varepsilon = \big(\varphi_\varepsilon^{-1}(x)\big)^2[1+O(\varepsilon)] + \ldots \Leftrightarrow$$
$$\varepsilon = \left(\tfrac{x}{k_0}\right)^2[1+O(\varepsilon)] + \ldots,$$

so that $x_\pm = \pm|k_0|\sqrt{\varepsilon} + O(\varepsilon)$. **QED**.

To show the invariance of the derivatives of $f_\xi \circ f_\xi$ at x_+ and x_-, use the fact that the map $f\colon X \to X$ is brought by a change in variables (the same change in variables for the preimage and image spaces) $\varphi\colon X \to X'$ to the superposition $f' = \varphi \circ f \circ \varphi^{-1}$: $X' \to X'$. (What follows from that?)

To establish stability, take Taylor-series expansions of $f_\xi \circ f_\xi$ at x_+ and x_-. Keeping to the notations x, y in place of, respectively, $x - x_\pm$ and $y - x_\pm$, these are

$$y = f_\xi(f_\xi(x)) = A_\pm x + o(x), \quad |A_\pm| < 1.$$

Therefore, for small x we will have

$$|y/x| < |A_{\pm}| + o(1) < q < 1,$$

so the iterations of small x tend to zero at the rate of a geometric progression.

Completing the Solution

S5.1

Concerning the invariance of $[1, e^{\xi-1}]$ under the map f_ξ for $1 \le \xi \le 2$: first, the image of f_ξ is within $[0, \max f] = [0, e^{\xi-1}]$; second, since $f_\xi(x)$ decreases for $x \ge 1$,

$$\min_{1 \le x \le e^{\xi-1}} f_\xi(x) = f_\xi\left(e^{\xi-1}\right) = e^{2\xi-\left(1+e^{\xi-1}\right)}.$$

Deduce from this the invariance of $[1, e^{\xi-1}]$ using the inequality $2\xi \ge 1 + e^{\xi-1}$, which holds for $1 \le \xi \le 2.2564\ldots$.[4]

Estimating the convergence rate with the help of the Lagrange intermediate value theorem, find that $f_\xi'(\theta) = e^{\xi-\theta}(1-\theta)$, from which it follows immediately that $\left|f_\xi'(\theta)\right| \le q < 1$ for $0 < \xi \le 1$ and θ close enough to ξ. (Why?) The function $f_\xi'(x)$ has for $x \ge 0$ a unique local extremum (which is a minimum) $f_\xi'(2) = e^{\xi-2}$ (also, it has a unique zero at $x = 1$ and a unique inflexion at $x = 3$; create the graph!). Therefore, $\left|f_\xi'(\theta)\right| \le e^{\xi-2} < 1$ when $\xi < 2$ and $\theta \ge 1$ and $\left|f_\xi'(\theta)\right| \le 1$ for $\xi = 2$ and $\theta \ge 1$.

The fixed points of $f_\xi \circ f_\xi$ on $(0,\infty)$ are determined by the equation $x\,(1 + e^{\xi-x}) = 2\xi$. The derivative on the left-hand side equals $f_\xi'(x) + 1$, and since $-e^{\xi-2}$ is greater than, equal to, or less than -1 when ξ is, respectively, less than, equal to, or greater than 2, the graph of this derivative then has no common points with the abscissas' axis, or one such point (of multiplicity two, without crossing), or two distinct such points (simple, with transversal crossing). (Create a figure.) Hence, the graph on the left-hand side of the equation itself (draw it!) has a (unique) inflection for $x = 2$, and with this

- for $\xi = 2$, the tangent at the inflection point is horizontal,
- for $\xi > 2$, a local maximum on the left side of the inflection and a local minimum on the right-hand side of it arise.

[4] The endpoints are the roots of the equation $2\xi = 1 + e^{\xi-1}$; the inequality can be proved using the property of the graph of a strictly convex function to have at most two common points with any straight line; see details in section S9.8, the problem group "Convexity and Related Classical Inequalities."

Such a graph has one simple common point for $\xi < 2$, one simple or one triple common point for $\xi = 2$, and one simple, or one simple and one double, or three simple common points for $\xi > 2$ with a horizontal straight line. Therefore, we must verify that the graph $y = x\,(1 + e^{\,\xi - x})$ has just three distinct common points with the horizontal line $y = 2\xi$ ($\xi > 2$), and the point $x = \xi$, which corresponds to a fixed point of f_ξ, is the middle one (so the two extreme points will correspond to its 2-point cycle). For this, use the negativity of the derivative at a point $x = \xi$: $y'(\xi) = f_\xi'(\xi)\ +1 = 2 - \xi < 0$ for $\xi > 2$. (We leave it to the reader to furnish the details.)

S5.3

A map $\varphi:\ x \mapsto x\sqrt{1+\kappa}$, with $|\kappa| \le r < 1$, may be determined via the absolutely convergent binomial series as $\varphi = x\,(1 + \kappa/2 - \kappa^2/(4{\cdot}2!) + \ldots)$. For small κ it has the form $\varphi = x{\cdot}(1 + \kappa_1)$, using respectively small κ_1.[5] Verify that for a smooth (that is, sufficiently many times or infinitely differentiable) function $\kappa = f(x,\varepsilon)$, disappearing when $(x,\varepsilon) \to 0$, φ has derivatives of the same orders as f, with $(\partial\varphi/\partial x)(0,0) = 1$. Therefore, using the implicit function theorem shows that the map $\varphi(.,\varepsilon)$ is, for small ε, uniquely invertible on a neighborhood of the origin in the x-space[6] and brings the inversion formula to $x = \varphi \cdot [1 + g(\varphi,\varepsilon)]$, with a smooth function g (having derivatives of the same orders as f) disappearing as $(\varphi,\varepsilon) \to 0$. Generally, knowing a Taylor-series expansion $\varphi(x,\varepsilon) = x \cdot (a_0(\varepsilon) + a_1(\varepsilon)x + \ldots)$ ($a_0 \neq 0$), one will find the expansion for the inverse map, $\varphi^{-1}(\varphi,\varepsilon) = \varphi \cdot (b_0(\varepsilon) + b_1(\varepsilon)\varphi + \ldots)$, by solving step by step the following system of equations:

$$
\begin{aligned}
&a_0 b_0 = 1,\\
&a_0 b_1 + a_1 b_0 = 0,\\
&a_0 b_2 + 2a_1 b_0 + a_2 b_0 = 0,\\
&a_0 b_3 + a_1\left(\mathrm{b}_1{}^2 + 2b_0\mathrm{b}_2\right) + 3a_2 b_0 \mathrm{b}_1 + a_3 b_0 = 0,
\end{aligned}
$$

and so on. (Fill in the details.) For $\varphi = x{\cdot}[1 + O(\varepsilon) + O(x)]$ [when $a_0(\varepsilon) = 1 + O(\varepsilon)$] we will have $b_0 = 1/a_0 = 1 + O(\varepsilon)$ (why?), so the inverse coordinate substitution has the form $x = \varphi{\cdot}\cdot[1 + O(\varepsilon) + O(\varphi)]$.

[5] The same holds for $\varphi = x{\cdot}(1 + \kappa)^\alpha$, with any real exponent α, and, moreover, for $\varphi = x{\cdot}\Phi(1 + \kappa)$, where Φ is any differentiable, with a bounded first derivative, function on a neighborhood of 1, such that $\Phi(1) = 1$. (Complete the details.)

[6] Which means that $x \mapsto \varphi(x,\varepsilon)$ is a (local) coordinate substitution: φ is the new (local) coordinate (which depends on parameter ε).

The fact that the derivatives of $f_\xi \circ f_\xi$ at x_+ and x_- do not change if we apply an equal coordinate substitution for the preimage and image spaces can be proved in a general situation: for any map f: $X \to X$, the map $f' = \varphi \circ f \circ \varphi^{-1}$: $X' \to X'$ obtained from it by a change in variables φ: $X \to X'$, and a fixed point x_0, $df(x_0) = df'(\varphi(x_0))$. [Use two identities $df'(\varphi(x)) = d\varphi\,(f(x)) \circ df(x) \circ d\varphi^{-1}(\varphi(x))$, $d\varphi^{-1}(\varphi(x)) = (d\varphi(x))^{-1}$ and the fact that the point x_0 is fixed, $f(x_0) = x_0$. We leave it to the reader to come up with the details.]

Polar and Singular Value Decomposition Theorems

Problems

P6.0

Preliminaries. "There exists a very powerful set of techniques for dealing with sets of equations or matrices that are either singular or else numerically very close to singular. In many cases where Gaussian elimination and LU[1] decomposition fail to give satisfactory results, this set of techniques, known as **singular value decomposition**, or **SVD**, will diagnose for you precisely what the problem is. In some cases, SVD not only diagnoses the problem, it will also solve it, in the sense of giving you a useful numerical answer... SVD is also the method of choice for solving most linear least-squares problems..." [quoted from Press et al. (1992)]. Usually SVD is applied to matrices with the number of rows, m, greater than or equal to the number of columns, n. *For* $m \geq n$, a **real-valued SVD** of $m \times n$ (real-valued) matrix A is defined as a factorization

$$A = U \circ D \circ V,$$

where U is an $m \times n$ matrix with orthonormal columns, ${}^tU_i \circ U_j = \delta_{ij}, i, j = 1,\ldots, n$, D is a diagonal $n \times n$ matrix with nonnegative diagonal elements, and V is an $n \times n$ orthogonal matrix, ${}^tV_i \circ V_j = \delta_{ij}, i,j = 1,\ldots, n$ ($\Leftrightarrow$ ${}^tV \circ V = V \circ {}^tV = E_{n\times n}$). In a **complex-valued version of SVD**, the orthogonality is replaced with the unitary property: ${}^t\overline{U}_i \circ U_j = \delta_{ij}$, ${}^t\overline{V} \circ V = V \circ {}^t\overline{V} = E_{n\times n}$. *For* $m < n$, the definition of SVD is almost the same, with the difference that in U the last $n - m$ columns are zero, and 0 in D the last $n - m$ diagonal elements are zero. **Polar decompositions**,

[1] **L**ower- and **U**pper-triangular.

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_6, © Springer Science+Business Media, LLC 2013

or **PDs** (also known as **canonical factorization**), are defined as factorizations of square matrices

$$A = F_1 \circ S_1 = S_2 \circ F_2,$$

where S_1, S_2 are symmetric (self-adjoint) matrices with nonnegative eigenvalues and F_1, F_2 are orthogonal (resp. unitary) (Gelfand 1971; Kostrikin and Manin 1980). *PDs generalize a complex number's factorization by its modulus and phase factor*.

The existence of SVD implies the existence of both forms of PD, as we may take $S_1 = V^{-1} \circ D \circ V$, $S_2 = U \circ D \circ U^{-1}$, $F_1 = F_2 = U \circ V$. We will see that the SVD of any matrix is always realizable. In fact, there are practically effective SVD procedures that provide efficient usage of SVD in numerical analysis; this topic is beyond the scope of this book; interested readers may consult sources such as Press et al. (1992) (with no proofs) and references therein. Applications of SVD are not restricted to numerical analysis; some other application examples are presented below (sections P6.8***, P6.9**, P6.10**, and P6.11**). More advanced readers who wish to familiarize themselves with infinite-dimensional versions of SVD and PD and their various applications may refer to Riesz and Nagy (1972), Dunford and Schwartz (1963), and Reed and Simon (1972).

*P6.1***

Show that for $m \geq n$, matrices determine their diagonal SVD factors up to arbitrary permutations of diagonal elements, and for $m < n$ up to arbitrary permutations of first m diagonal elements. In fact, D^2 is similar to ${}^tA \circ A$ (or, in the complex-valued version, to ${}^t\overline{A} \circ A$).

*P6.2***

Show that the PD symmetric (self-adjoint) factors of a square matrix are determined uniquely. (This can be done as follows: $S_1^2 = {}^tA \circ A$, $S_2^2 = A \circ {}^tA$.) Therefore, A is determined by the product ${}^tA \circ A$ (or $A \circ {}^tA$) up to an arbitrary left (resp. right) orthogonal (resp. unitary) factor (in the complex-valued version, ${}^t\overline{A}$ is used). *Show* that the PD orthogonal (unitary) factors of a square nondegenerate matrix also are determined, and with this, $F_1 = F_2$.

*P6.3***

Show that there exist SVDs of all matrices if there exist PD of all square matrices.

P6.4**

Prove the existence of both forms of PD for nondegenerate matrices. (*This problem is much easier than the next one!*)

P6.5**

Prove the existence of both forms of PD of arbitrary (square) matrices. *Are* the orthogonal (unitary) factors uniquely determined if a matrix is degenerate?

P6.6**

Show that for a square matrix A, ${}^tA \circ A$ and $A \circ {}^tA$ (in the complex-valued version, ${}^t\overline{A}$ is used) are similar matrices. [***Warning***: matrices $B \circ A$ and $A \circ B$ may be not similar (if both A and B degenerate), although they always have equal characteristic polynomials; *prove* it!]

P6.7**

Why does the previous problem concern the similarity of ${}^tA \circ A$ and $A \circ {}^tA$ (${}^t\overline{A} \circ A$ and $A \circ {}^t\overline{A}$) but not ${}^tA$ (resp. ${}^t\overline{A}$) and A themselves? *Prove* that matrices in $\mathbb{R}^n$ are similar to their adjoints, ${}^tA = T^{-1} \circ A \circ T$. *Does* the analogous theorem hold for $\mathbb{C}^n$, that is, are operators ${}^t\overline{A}$, A similar?

P6.8***

This problem is for readers familiar with elements of general topology. A topological space is **connected** when it is not a disjunctive union of its nonempty open subsets; otherwise, a maximal connected subset is called a **connected component**. (The space is a disjunctive union of its connected components, which in turn are closed subsets; *prove*.) The group of all invertible matrices $GL(n, \mathbb{F})$ ($\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$) is an (open) subset of the space of all $n \times n$ matrices (*why*?), so that this group gets a topology making it a topological space itself. *Prove* that $GL(n,\mathbb{C})$ is connected, whereas $GL(n,\mathbb{R})$ has two connected components (Fig. 1).

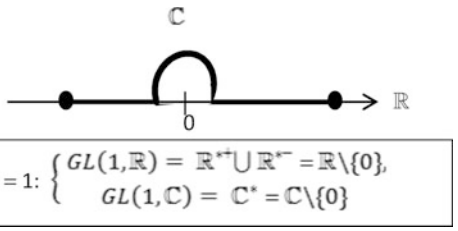


Fig. 1 A route in $GL(1,\mathbb{C})$ between two elements of $GL(1,\mathbb{R})$

*P6.9***

Prove using SVD that a nonzero linear operator on $\mathbb{R}^n$ ($n \geq 2$) commutes with any **rotational** (also called **special orthogonal**: orthogonal and having the determinant equal to one, i.e., orientation preserving) linear operators if and only if it is a **scalar operator** (also called a **homothetic transformation**, or a **dilatation**, i.e., proportional to the identity operator) when $n \geq 3$ and a composition of scalar and rotational operators when $n = 2$. (*Why* is the answer more complicated for $n = 2$?) *Derive* from the foregoing statements that the linear operators on $\mathbb{R}^2$ commuting with the rotations form a commutative two-dimensional subring of the four-dimensional ring of all linear operators that is isomorphic to the field $\mathbb{C}$.

*P6.10***

A linear operator A in a Euclidean (complex Hermitian) space is called **normal** if it commutes with its adjoint operator: $A^* \circ A = A \circ A^*$, where $A^* = {}^tA$ (resp. $A^* = {}^t\overline{A}$).[2] Obviously, a linear operator is normal if and only if its adjoint operator is normal. Symmetric and orthogonal (resp. self-adjoint and unitary) linear operators are normal. However, these are not all of the normal operators, and, conversely, not every linear operator is normal (unless the space is one-dimensional), as will be made clear from the spectral theorem below.[3]

Show that for a normal operator A in a finite-dimensional space, the subspaces $\ker A$ and $\operatorname{im} A$ are orthogonal complements of each other, and therefore $\ker A = \ker A^*$ and $\operatorname{im} A = \operatorname{im} A^*$.

Prove that a linear operator on a finite-dimensional space is normal if the factors of some of its PD commute with each other, and only if this holds for any of its PD. *Show* that the kernel and the image of the normal operator are invariant subspaces for all of its PD factors.

---

[2] Here we do not distinguish between the operators and their matrix representations with respect to the standard coordinate orthobasis.

[3] We consider normal operators on finite-dimensional spaces only. More advanced readers wishing to familiarize themselves with the properties of infinite-dimensional normal operators may refer to Riesz and Nagy (1972) and Dunford and Schwartz (1963).

Derive from the preceding statements the following spectral theorem for normal operators on finite-dimensional spaces, which generalizes spectral theorems for symmetric (self-adjoint) and orthogonal (resp. unitary) operators:

A linear operator on a Hermitian space is normal if and only if it is diagonalizable in an appropriate orthobasis. Therefore, a linear operator having only real eigenvalues is normal if and only if it is self-adjoint.

A linear operator on a Euclidean space is normal if and only if its matrix has a block-diagonal form in an appropriate orthobasis, with one- or two-dimensional blocks being, respectively, λ *and* $|\lambda| \begin{pmatrix} \cos \arg \lambda & -\sin \arg \lambda \\ \sin \arg \lambda & \cos \arg \lambda \end{pmatrix}$ *for real eigenvalues* λ *and complex conjugate pairs of eigenvalues* $\lambda, \overline{\lambda}$ $(\arg \lambda \neq 0, \pi)$*. Therefore, a linear operator having only real eigenvalues is normal if and only if it is symmetric (i.e., is diagonalizable in an appropriate orthobasis).*

P6.11[**]

(*The problems in this section assume no familiarity with normed vector spaces.*) The norm of linear operator A in a finite-dimensional Euclidean (Hermitian) vector space is a nonnegative number defined, with respect to the Euclidean (resp. Hermitian) modulus $|\cdot|$ in the source space, as $\|A\| := \sup_{x \neq 0} |Ax|/|x| = \sup_{|x|=1} |Ax|$. Because of the continuity of both the function $x \mapsto |x|$ and finite-dimensional linear operators and the compactness of a finite-dimensional sphere, we might use the Weierstrass theorem to substitute "sup" with "max" in the preceding formulation. (We leave it to the reader to fill in the details.)

Verify that $\| A \circ B \| \leq \|A\| \cdot \|B\|$ for any linear operators A, B. (Therefore, $\|A^n\| \leq \|A\|^n$, $\forall n = 2,3,\ldots$)

What is the norm of an orthogonal (unitary) operator? Prove that $\| A \circ B \| = \| B \circ A \| = \|A\|$ if B is orthogonal (unitary). Therefore, for PD factors S_1, S_2 of a linear operator A, $\|S_1\| = \|S_2\| = \|A\|$.

Prove that the norm of a normal operator (section P6.10[**]) equals the maximal modulus of its eigenvalue. Therefore, for normal operators $\|A^n\| = \|A\|^n$, $\forall n = 2,3,\ldots$. (Do the similar claims hold for a linear operator when it is not the normal one?)

Derive from the preceding statements that for any linear operator A, $\| A^* \circ A \| = \| A \circ A^* \| = \|A\|^2$, and therefore $\|A\| = \|A^*\|$ (A^* is the adjoint operator).

Hint

In the cases where real- and complex-valued versions of sections H6.1, H6.2, H6.3, H6.4, H6.5, and H6.6 are parallel and have similar proofs we consider only the first version.

H6.1

An $n \times n$ matrix ${}^tU \circ U$ is either identity (E_n) or $\begin{pmatrix} E_m & 0 \\ 0 & 0 \end{pmatrix}$ for, respectively, $m \geq n$ and $m < n$. From that we get

$$ {}^tA \circ A = ({}^tV \circ D \circ {}^tU) \circ (U \circ D \circ V) = {}^tV \circ D^2 \circ V.$$

H6.2

The equalities $S_1^2 = {}^tA \circ A$, $S_2^2 = A \circ {}^tA$ are obtained by direct computation. The right-hand sides are symmetric matrices with nonnegative eigenvalues, so that one can extract unique square roots S_1, S_2 of nonnegative eigenvalues. (Why?) They do not degenerate if A does not, and then the orthogonal PD factors are uniquely determined as $F_1 = A \circ S_1^{-1}$, $F_2 = S_2^{-1} \circ A$. Next, representing A as

$$A = F_1 \circ S_1 = \left(F_1 \circ S_1 \circ F_1^{-1}\right) \circ F_1$$

we will have, due to the PD uniqueness, $S_2 = F_1 \circ S_1 \circ F_1^{-1}$, $F_2 = F_1$.

H6.3

In what follows we will identify matrices with linear operators that these matrices define using fixed coordinate bases. [An $m \times n$ matrix $W = \begin{pmatrix} w_{11} & \dots & w_{1n} \\ \vdots & & \vdots \\ w_{m1} & \dots & w_{mn} \end{pmatrix}$ defines the operator W that maps n-dimensional column vectors $e_i = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \leftarrow i$ onto, respectively, $\begin{pmatrix} w_{1i} \\ \vdots \\ w_{mi} \end{pmatrix}$, $i = 1, \dots, n$, or, equivalently, W acts on vectors $x = \sum x_i e_i = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ in accordance with the usual matrix multiplication rule:

$$Wx = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & & \vdots \\ w_{m1} & \cdots & w_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \sum w_{1j}x_j \\ \vdots \\ \sum w_{mj}x_j \end{pmatrix}.$$ To establish SVD when $m \geq n$, consider the usual (by coordinates) inner products in spaces $\mathbb{R}^m$, $\mathbb{R}^n$, take an orthogonal decomposition $\mathbb{R}^m = L^n \bigoplus M^{m-n}$, where an n-dimensional subspace L contains the image of A, and then define a linear operator ${}^tW : \mathbb{R}^m \to \mathbb{R}^n$ with kernel M, which maps an orthobasis of L onto an orthobasis of $\mathbb{R}^n$. Show that the operator $W \circ {}^tW$ [where the adjoint to the ${}^tW$ operator W corresponds to the transposed matrix ${}^t({}^tW) = W$] maps L onto itself and, moreover, acts on it as identity operator id_L. Using this, derive SVD of A from PD of a matrix ${}^tW \circ A$:

$$A = W \circ {}^tW \circ A = W \circ S \circ F = (W \circ {}^tT) \circ D \circ (T \circ F),$$

where D is a diagonal form of S in proper Cartesian coordinates and T is the corresponding $n \times n$-orthogonal matrix. To establish SVD for $m < n$, consider the orthogonal decomposition $\mathbb{R}^n = \mathbb{R}^m \bigoplus \mathbb{R}^{n-m}$ and a linear operator ${}^tW$: $\mathbb{R}^m \to \mathbb{R}^n$ orthogonally mapping the first summand onto itself. $W \circ {}^tW$ acts on $\mathbb{R}^m$ as identity operator, $W \circ {}^tW = E_m$. (Why?) Using this information, derive the SVD of A from the PD of ${}^tW \circ A$, as was done previously. Next, every element in the last n–m rows of the matrix ${}^tW$ has zero value; therefore, the same holds for the matrix $S = {}^tW \circ A \circ F$. Hence, S, due to its symmetry, has the form $S = \begin{pmatrix} S_m & 0 \\ 0 & 0 \end{pmatrix}$, where S_m is a symmetric $m \times m$ matrix . Consequently, S is diagonalized as $S = {}^tT \circ D \circ T$, with $D = \begin{pmatrix} D_m & 0 \\ 0 & 0 \end{pmatrix}$ and $T = \begin{pmatrix} T_m & 0 \\ 0 & E_{n-m} \end{pmatrix}$. Lastly, every element in the last n–m rows of the matrix W has zero value, so the same holds for the matrix $U = W \circ {}^tT$.

H6.4

Diagonalizing the symmetric operator ${}^tA \circ A$ in proper Cartesian coordinates,

$${}^tA \circ A = {}^tT \circ D \circ T, \qquad D = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix} \text{is diagonal,} \quad T \text{ is orthogonal,}$$

define $S_1 := {}^tT \circ \sqrt{D} \circ T$, with $\sqrt{D} = \begin{pmatrix} \sqrt{d_1} & & \\ & \ddots & \\ & & \sqrt{d_n} \end{pmatrix}$, and $F_1 := A \circ {S_1}^{-1}$. Check by direct computation that $F_1 \circ {}^tF_1 = E$. We have obtained the first PD

$A = F_1 \circ S_1$. The second one, $A = F_2 \circ S_2$, arises as in section H6.2 by defining $S_2 = F_1 \circ S_1 \circ {F_1}^{-1}$, $F_2 = F_1$.

H6.5

It is sufficient to establish only one PD form, for example, the form $A = S \circ F$, since these forms produce one another: $S \circ F = F \circ ({}^tF \circ S \circ F)$. Based on the lemmas from section E6.3, take a spectral orthobasis for $A \circ {}^tA$ as a union of orthobases in the spaces im A and ker ${}^tA$. In this basis, operators A and $S = \sqrt{A \circ {}^tA}$ have block matrices $\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} s & 0 \\ 0 & 0 \end{pmatrix}$, respectively ($s$ is invertible on im A). Prove the existence of orthogonal matrices $F = \begin{pmatrix} x & y \\ z & t \end{pmatrix}$, satisfying the equation $\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} s & 0 \\ 0 & 0 \end{pmatrix} \circ \begin{pmatrix} x & y \\ z & t \end{pmatrix}$, and find a way to enumerate them.

H6.6

The similarity of ${}^tA \circ A$ and $A \circ {}^tA$ follows from the PD of A. Next, for arbitrary square matrices A, B, if A (or B) is invertible, then $A^{-1} \circ (A \circ B) \circ A = B \circ A$ [resp. $B \circ (A \circ B) \circ B^{-1} = B \circ A$]. Analyze the following simplest example when $A \circ B$ is not similar to $B \circ A$: $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ $\left(A \circ B = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \right.$ $\left. B \circ A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \right)$. Next, turning to the characteristic polynomials $\chi_C(\lambda) = \det(\lambda E - C)$, we suggest proving the equality $\chi_{A \circ B} = \chi_{B \circ A}$ proceeding by any of the following three methods.

(1) Using continuous continuation: The left- and right-hand sides can be approximated with any precision by, respectively, $\chi_{A' \circ B'}$ and $\chi_{B' \circ A'}$, where A', B', or even both are invertible. (How does this help to prove the equality?)

(2) Using analytic continuation: Coefficients of the polynomial $\chi_C(\lambda) = \lambda^n + \ldots$ are polynomials in entries of matrix C, so they are analytical functions of these entries. The equality being proved holds on an open subset of the space of $2n^2$ entries of A, B (which one?); therefore, in accordance with the **analytic continuation principle** (Bochner and Martin 1948; Dieudonné 1960), it will hold on all of the space (as the space is connected). Applying a slight modification of this argument, coefficients of $\chi_C(\lambda)$ are analytic functions on the space of unordered (taken up to arbitrary permutations of the elements) sets of eigenvalues, and the

desired equality for all points in this space follows by analytic continuation from an open subset.

(3) Using the Jordan canonical form: It is sufficient to establish that "elementary factors" $(\lambda-\lambda_i)^{n_i}$ $(\lambda_i\neq 0)$ of the factorization $\chi_{A\circ B}(\lambda)=\lambda^{n_0}\prod_{\lambda_i\neq 0}(\lambda-\lambda_i)^{n_i}$, which is taken over the eigenvalues of $A\circ B$, divide $\chi_{B\circ A}(\lambda)$. (Why?) Considering the complexified space ${}^{\mathbf{C}}\mathbb{R}^n=\mathfrak{L}\otimes_{\mathbf{R}}\mathbb{R}^n=\mathbb{R}^n\oplus i\mathbb{R}^n$ and complexified linear operators reduces the real-valued case to the complex one.[4] Let L be the sum of all root subspaces of $A\circ B$ that correspond to the eigenvalue $\lambda\neq 0$, so that it acts on a Jordan basis of L as follows:

$$\begin{aligned}&(A\circ B)e_1^1=\lambda\, e_1^1,\qquad .....,\quad (A\circ B)e_{k_1}^1=e_{k_1-1}^1+\lambda\, e_{k_1}^1,\\ &\quad\dots\\ &(A\circ B)e_1^N=\lambda\, e_1^N,\quad .....,\quad (A\circ B)e_{k_N}^N=e_{k_N-1}^N+\lambda\, e_{k_N}^N.\end{aligned}$$

Since $\lambda\neq 0$, $L\cap\ker B=\{0\}$, in other words, B does not identify points of L. (Why?) Denoting $g_l^j:=B\,e_l^j$, application of B to the previous series of equalities yields

$$\begin{aligned}&(A\circ B)e_1^1=\lambda\, e_1^1,\qquad \dots..,\qquad (A\circ B)\,e_{k_1}^1=e_{k_1-1}^1+\lambda\, e_{k_1}^1,\\ &\quad\dots\\ &(A\circ B)e_1^N=\lambda\, e_1^N,\qquad \dots..,\qquad (A\circ B)\,e_{k_N}^N=e_{k_N-1}^N+\lambda\, e_{k_N}^N.\end{aligned}$$

These equalities show that B maps the sum of the root subspaces of $A\circ B$ corresponding to the eigenvalue λ onto a sum of root subspaces of $B\circ A$ corresponding to the same eigenvalue, completing the proof. [The same method applies to matrices with entries from any integral domain (a commutative and associative unitary ring, free of zero divisors).]

H6.7

Adjoint operators on $\mathbb{C}^n$ are generally not similar. For example, similarity with a scalar (homothety) operator zE is equivalent to equality with it, whereas conjugation is not (unless $z\in\mathbb{R}$). Prove that the linear operators A and ${}^t\overline{A}$ in $\mathbb{C}^n$ are similar if and only if A has a real-valued matrix with respect to some basis. Use the Jordan canonical form. For $\mathbb{R}^n$, use the Jordan form over the reals. (***Warning***: By duality, A and ${}^t\overline{A}$ are similar if and only if ${}^t\overline{A}$ has a real-valued matrix with respect to some

[4] The complexified linear operator has the same matrix in a $\mathbb{C}$-basis of the complexified space as a source operator in the same basis, considered as a $\mathbb{R}$-basis of a source space.

basis. These bases may however not be orthogonal, and so A and ${}^t\overline{A}$ may have real-valued matrices with respect to different bases. Also, the similitude between A and ${}^tA$ in $\mathbb{R}^n$ cannot always be realized using an orthogonal matrix.)

H6.8

The union of two intersecting connected subsets is connected so that distinct connected components do not intersect, and the space is a disjunctive union of the components. The components are closed since the closure of a connected set is also connected. We can investigate connected components of $GL(n,.)$ by employing **arcwise connectedness**. A continuous map on a segment [0,1] to a topological space T is called an **arc in** T connecting $\gamma(0)$ and $\gamma(1)$. T is **arcwise connected** when any two points are connected by an arc; otherwise a maximal arcwise connected subset is called an **arcwise connected component**. (A topological space is a disjunctive union of its arcwise connected components. Why? Must the components be closed subsets?) Prove the following:

Lemma *Arcwise connectedness implies connectedness.*

(Does connectedness imply arcwise connectedness?) Show that the symmetric operators in $GL(n,\mathbb{R})$ [and self-adjoint ones in $GL(n,\mathbb{C})$] with positive eigenvalues form an arcwise connected and, therefore by virtue of this lemma, connected subset. Find, using the lemma, connected components of orthogonal and unitary groups. Using this and employing PD, find connected components of $GL(n,\mathbb{R})$ and $GL(n,\mathbb{C})$.

Also, we can establish the arcwise connectedness of $GL(n,\mathbb{C})$ by the following method: for $M_0, M_1 \in GL(n,\mathbb{C})$ consider a polynomial $f(z) := \det M_z$, $M_z := (1 - z) M_0 + zM_1$. f's zero set does not divide $\mathbb{C}$; that is, there exists an arc $t \mapsto z(t)$ in $\mathbb{C}$ with endpoints 0 and 1, avoiding those zeros, or - equivalently - the arc $\gamma : t \mapsto M_{z(t)}$ between M_0 and M_1 will lie completely in $GL(n,\mathbb{C})$. **QED**. (With this, we can obtain $|\operatorname{Im} z(t)| < \varepsilon$ for an arbitrarily given $\varepsilon > 0$.)

H6.9

The special orthogonal group (or the group of all rotations) $SO(n)$ is commutative (Abelian) if and only if $n \leq 2$, which explains the difference in the answers for $n = 2$ and $n \geq 3$. Readers who have not worked with $SO(2)$ can establish its commutativity with the help of an identity such as $\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \circ \begin{pmatrix} \cos\Psi & -\sin\Psi \\ \sin\Psi & \cos\Psi \end{pmatrix} = \begin{pmatrix} \cos(\varphi+\Psi) & -\sin(\varphi+\Psi) \\ \sin(\varphi+\Psi) & \cos(\varphi+\Psi) \end{pmatrix}$, or using the complex-valued representation (isomorphism)

$$SO(2) \cong U(1) : \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \leftrightarrow e^{i\varphi},$$

or using the isomorphism $SO(2) \cong \mathbf{S}^1 \cong \mathbb{R} \bmod (2\pi\mathbb{Z})$ and the fact that a homomorphic image of an Abelian group also is Abelian, or proceeding by a different method of the readers' preference. In turn, noncommutativity of $SO(n)$ for $n \geq 3$ can be shown as follows: two rotational operators, either of which rotates the vectors lying on a two-dimensional plane and acts as the identity operator on its $n-2$-dimensional orthogonal complement,[5] in general do not commute since they do not have the same plane of rotation. (We leave it to the reader to come up with examples.)

A proof of sufficiency: this is evident taking into account the commutativity of $SO(2)$.

A proof of necessity:

$n = 2$. Use SVD to factorize a linear operator A by $A = F \circ D \circ G$, where F and G are rotational and D is a diagonal (with respect to the standard coordinate basis) operator. If A commutes with any rotations, then, by virtue of the commutativity of $SO(2)$, D commutes with any rotations (why?), and therefore it is a dilatation (why?). Finally, $A = T \circ D$, where $T = F \circ G \in SO(2)$.

$n \geq 3$. In this case PD will not help us much since $SO(n)$ is not an Abelian group. We suggest proceeding by the following method. Operator A commutes with any rotations, in particular, with any linear operator T rotating the vectors on some two-dimensional plane Π and acting as the identity operator on its $n-2$-dimensional orthogonal complement $\Pi^{\perp}$; therefore, A retains invariant (as a whole) each $n-2$-dimensional linear subspace in $\mathbb{R}^n$. (Why?) For $n = 2$ this is a trivial fact, but for $n \geq 3$ it implies retaining as invariant (as a whole) any linear subspace (a straight line passing via the origin); indeed, for $n \geq 3$ that subspace is an intersection of $n-2$ vector subspaces, each having dimension $n-2$. Deduce from the preceding statements that A is a dilatation.

A different proof using arguments of analysis is discussed in section S6.9 (the "Completing the Solution" section).

A proof of the fact that all linear operators on $\mathbb{R}^2$ that are compositions of rotations and dilatations form a subring of the ring of all linear operators that is isomorphic to the field $\mathbb{C}$. Let us employ a matrix representation of linear operators with respect to the standard basis. Readers can immediately verify that,

$$\lambda E \circ R_\theta = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} = aE + bI, \text{ using } a = \lambda \cdot \cos\theta = |\lambda| \cdot \begin{cases} \cos\theta & \text{when } \lambda \geq 0, \\ \cos(\theta+\pi) & \text{otherwise,} \end{cases}$$
$$b = \lambda \cdot \sin\theta = |\lambda| \cdot \begin{cases} \sin\theta & \text{when } \lambda \geq 0, \\ \sin(\theta+\pi) & \text{otherwise,} \end{cases}$$

[5] In fact, these operators generate the group $SO(n)$ (see section P10.27**, in the "One-Parameter Group of Linear Transformations" problem group below).

where $R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ and $I = R_{\pi/2} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, and, conversely, $aE + bI = rE \circ R_\theta$, where $r = \sqrt{a^2 + b^2}$, $\theta = \begin{cases} \operatorname{arctg}(b/a) & \text{when } a \geq 0, \\ \operatorname{arctg}(b/a) + \pi & \text{otherwise;} \end{cases}$ hence, the foregoing compositions in fact form the set as $\{aE + bI\colon a,b \in \mathbb{R}\}$. A one-to-one correspondence $aE + bI \leftrightarrow a + bi$ supports the ring structure because of the relations $E \circ I = I \circ E = I$, $E^2 = -I^2 = E$; this completes the proof. (Furnish the details.)

H6.10

A proof of the facts about ker A and im A for a normal operator A: use Lemma 1 in section E6.3 ("Explanation") and its Hermitian analog.

Characterization of the normal operator via the commutativity of PD factors follows from the uniqueness of both self-adjoint PD factors and explicit formulas for these factors (section P6.2**) and from Lemma 1 in section E6.3 and its Hermitian analog.

Proof of spectral theorem for normal operators.

Sufficiency. Let the indicated orthobasis exist. A straightforward computation, which we leave to the reader, shows that $A^* \circ A$ and $A \circ A^*$ have the same matrix in this basis, namely diagonal, with the diagonal entries $|\lambda|^2$ [two-dimensional blocks of matrix A corresponding to λ, $\overline{\lambda}$ (arg $\lambda \neq 0,\pi$) give two diagonal elements $|\lambda|^2$].

Necessity. It is sufficient to prove that for the restriction $A|_{\mathrm{im}\ A}$ instead of A. If all of its eigenvalues belong to the field, as always happens in the Hermitian case and in the Euclidean case where all eigenvalues are real, we may for an eigenvalue λ replace A by an operator $A - \lambda E$ that is normal if A is normal and finalize the proof by induction on the spatial dimension. Complex conjugate eigenvalues in the Euclidean case can be processed by a similar method using complexification (we leave it to the reader to furnish the details of this approach). However, we suggest making use of the commutativity of PD factors and applying spectral theorems for symmetric (self-adjoint) and orthogonal (unitary) operators, which will enable one to complete the proof in one step.

H6.11

Verification of inequality $\|A \circ B\| \leq \|A\| \cdot \|B\|$. By definition of the norm, $|Ax| \leq \|A\| \cdot |x|$ for $x \neq 0$. If $x = 0$, then $|Ax| = 0 = \|A\| \cdot |x|$. Therefore, $|Ax| \leq \|A\| \cdot |x|$, $\forall x$, so we have $|A(Bx)| \leq \|A\| \cdot |Bx| \leq \|A\| \cdot \|B\| \cdot |x|$, $\forall x$, and, $|A(Bx)|/|x| \leq \|A\| \cdot \|B\|$, $\forall x \neq 0$, which implies the desired inequality.

An orthogonal (unitary) operator preserves the modulus of a vector, and so its norm equals one. The equalities $\| A \circ B \| = \| B \circ A \| = \|A\|$ with orthogonal (unitary) B follow from the modulus preservation property of B, but they can also be derived from the inequality $\| A \circ B \| \leq \|A\| \cdot \|B\|$ and the fact that $\| B\| = \| B^{-1}\| = 1$. (Provide the details.)

Computation of normal operator's norm. Prove that if a space is decomposed into an orthogonal sum of invariant subspaces L_i of a linear operator A, then $\|A\| = \max_i \|A|_{L_i}\|$. Therefore, one may calculate the maximum for the blocks as described by the spectral theorem for normal operators (section P6.10**). Note that in the Euclidean case a two-dimensional block represents a composition of a diagonal operator with two diagonal elements $|\lambda|$ and an orthogonal operator (a counterclockwise rotation by the angle $\arg \lambda$). The claim that is being proved is now evident.

$\|A\|$ may be greater than the maximal modulus of the eigenvalue, and $\|A\|^n$ may be greater than $\|A^n\|$ if a linear operator A is not normal. In particular, if a linear operator on a two-dimensional vector space is not normal (i.e., either is diagonalizable in a nonorthogonal basis or possesses only one eigendirection), then $\|A\|$ is greater than the maximal modulus of the eigenvalue and $\|A\|^n > \|A^n\|$, which can be easily established with the help of PD.

Proof of equalities $\| A^* \circ A \| = \| A \circ A^* \| = \|A\|^2$. Since S_1 and S_2 are symmetric (self-adjoint), so are the normal operators $\|A\|^2 = \|S_1\|^2 = \|S_1{}^2\| = \| A^* \circ A \|$ and $\|A\|^2 = \|S_2\|^2 = \|S_2{}^2\| = \| A \circ A^* \|$.

Also, one of the foregoing equalities, $\| A^* \circ A \| = \|A\|^2$, can be established using the fact that the norm of $A^* \circ A$ equals its maximal eigenvalue and diagonalizing a quadratic form $x \mapsto \langle Ax, Ax \rangle = \langle (A^* \circ A)x, x \rangle$ with respect to the scalar product $x \mapsto \langle x, x \rangle$, i.e., determining critical points and critical values of this form on the hypersphere $\{x : \langle x, x \rangle = 1\}$ (which can be done either by the usual means of linear algebra or with Lagrange multipliers). We leave it to the interested reader to come up with the details.

An equality $\|A\| = \|A^*\|$ follows from the equalities $\| A^* \circ A \| = \| A \circ A^* \| = \|A\|^2$ substituting A by A^* and taking into account the involution property $A^{**} = A$.

The equality $\|A\| = \|A^*\|$ can be established also via duality, without PD. Let us consider a duality mapping of the source space onto its adjoint (or dual) space, D: $x \mapsto Dx(.) = \langle x, . \rangle$. D transforms the adjoint operator A^* into an operator $D \circ A^* \circ D^{-1}$: $x^* \mapsto x^* \circ A$ in the dual space (x^* denote elements of the dual space).[6] Defining the dual scalar product in the dual space as $\langle x^*, y^* \rangle^* = \langle D^{-1}x^*, D^{-1}y^* \rangle$ we have $|Dx|^* = |x|$, which implies the equality $\|D \circ A^* \circ D^{-1}\|^* = \|A^*\|$, and $|x^*|^* = \max_{|x|=1} |x^*(x)|$, which implies the equality $\|D \circ A^* \circ D^{-1}\|^* = \|A\|$. (Fill in all the details.)

[6] Note that the operator $D \circ A^* \circ D^{-1}$ is completely determined by A and does not depend on a scalar product that we used to define it!

Explanation

In cases where real- and complex-valued versions of sections E6.1, E6.2, E6.3 and E6.5 are parallel and have similar proofs, we consider only the first version.

E6.1

Eigenvalues of D are square roots of eigenvalues of ${}^tA \circ A$. Since diagonal elements of D are its eigenvalues, the diagonal set as a whole is determined uniquely. But any permutation, as prescribed in section P6.1**, is allowed. Indeed, after a permutation we obtain the diagonal matrix $D' = {}^tT \circ D \circ T$, with orthogonal T, so that $A = U \circ D \circ V = U' \circ D' \circ V'$ ($U' = U \circ T$, $V' = {}^tT \circ V$). For $m < n$, if only the first m diagonal elements of D can be permuted, then T has the form $\begin{pmatrix} T_m & 0 \\ 0 & E_{n-m} \end{pmatrix}$, with orthogonal $m \times m$ matrix T_m, so that the first m columns of U' are orthonormal and the last $n - m$ ones are zero (as with U). **QED**.

E6.2

The unique existence of symmetric, with nonnegative eigenvalues, square roots of ${}^tA \circ A$, $A \circ {}^tA$ follows from the following lemma:

Lemma *For a symmetric linear operator with nonnegative eigenvalues C, $\mathbb{R}^n \to \mathbb{R}^n$, there exists the unique symmetric linear operator with nonnegative eigenvalues S: $\mathbb{R}^n \to \mathbb{R}^n$ such that $S^2 = C$.*

Proof Consider the orthogonal decomposition $\mathbf{R}^n = \sum_{\lambda \in Spec(C)} {}^{\oplus} L_{C,\lambda}$, where $L_{C,\lambda}$ are the eigenspaces of C corresponding to eigenvalues λ and the sum is taken over the λ. A linear operator S, which acts on $L_{C,\lambda}$ by the homothety $\sqrt{\lambda} \cdot id_{L_{C,\lambda}}$, is a square root of C. To prove the uniqueness, consider the similar decomposition for a symmetric, with nonnegative eigenvalues, square root S of C, $\mathbf{R}^n = \sum_{\mu \in Spec(S)} {}^{\oplus} L_{S,\mu}$. We have $L_{C,\mu^2} = L_{S,\mu}$, $\forall \mu \in Spec(S)$. (Why?) Thus, the uniqueness of the preceding decomposition for C implies the uniqueness of S.

E6.3

Recall that for a linear operator $W\colon \mathbb{R}^n \to \mathbb{R}^m$ of two Euclidean spaces with the usual coordinatewise inner (or scalar) products (denoted by $\langle\rangle$ for both spaces), the adjoint operator W^* (*or* ${}^tW$): $\mathbb{R}^m \to \mathbb{R}^n$ is defined by a formula

$$\langle Wx, y\rangle_{\mathbf{R}^m} = \langle x, {}^tWy\rangle_{\mathbf{R}^n}, \quad \forall x \in \mathbf{R}^n, \quad \forall y \in \mathbf{R}^m,$$

the second conjugation returns to the source: ${}^t({}^tW) = W$, and the matrices of ${}^tW$ and W with respect to the coordinate bases of $\mathbb{R}^n$ and $\mathbb{R}^m$ are transposed[7]. The following two assertions explain the claims in section H6.3 concerning adjoint operators.

Lemma 1 *For a pair of adjoint operators* $W\colon \mathbb{R}^n \to \mathbb{R}^m$, ${}^tW\colon \mathbb{R}^m \to \mathbb{R}^n$ *there are orthogonal decompositions* $\mathbf{R}^n = \operatorname{im}{}^tW \oplus \ker W$, $\mathbf{R}^m = \operatorname{im} W \oplus \ker {}^tW$. *Therefore,* W (${}^tW$) *maps* $\operatorname{im}{}^tW$ *(resp.* $\operatorname{im} W$*) isomorphically onto* $\operatorname{im} W$ *(resp.* $\operatorname{im}{}^tW$*); in particular*

$$\ker W = \ker {}^tW \circ W, \qquad \operatorname{im}{}^tW = \operatorname{im}{}^tW \circ W,$$
$$\ker {}^tW = \ker W \circ {}^tW, \qquad \operatorname{im} W = \operatorname{im} W \circ {}^tW,$$

and therefore, $\operatorname{rk}{}^tW = \operatorname{rk}{}^tW \circ W = \operatorname{rk} W$ and $\operatorname{rk} W = \operatorname{rk} W \circ {}^tW = \operatorname{rk}{}^tW$, *so that all of these ranks are equal.*

Lemma 2 *The following claims about a pair of adjoint operators* $W\colon \mathbb{R}^n \to \mathbb{R}^m$, ${}^tW\colon \mathbb{R}^m \to \mathbb{R}^n$ *are equivalent:*

(i) *W maps some orthobasis of* $\operatorname{im}{}^tW$ *onto an orthobasis of* $\operatorname{im} W$.
(ii) *W maps any orthobasis of* $\operatorname{im}{}^tW$ *onto an orthobasis of* $\operatorname{im} W$.
(iii) *The restriction of* ${}^tW \circ W$ *to subspace* $\operatorname{im}{}^tW$ *is equal to* $id_{\operatorname{im}{}^tW}$
(iv) – (vi) *Similar claims,* with ${}^tW$ *and W switched.*

Prove these lemmas.

E6.5

The matrix equation in section H6.5 determines blocks x, y as follows:

$$x = s^{-1} \circ a, \qquad y = s^{-1} \circ b,$$

[7] Some advanced readers may object, saying that the adjoint operators act in the adjoint spaces (spaces of linear functionals), $W^*\colon \mathbb{R}^{m*} \to \mathbb{R}^{n*}$, to be exact $(W^*y^*)x = y^*(Wx)$, so use of inner products is not necessary [as for instance, "orthogonality" between $x \in \mathbb{R}^n$ and $x^* \in \mathbb{R}^{n*}$ means that $x^*(x) = 0$, and so on]. However, we use an equivalent approach because an inner product in a vector space defines the isomorphism of this space onto its adjoint $x \mapsto x^*$ such that $x^*(x') = \langle x, x\prime\rangle$, $\forall x'$.

and it allows arbitrary blocks z, t. Show that z and t can be found such that the matrix $\begin{pmatrix} s^{-1}\circ a & s^{-1}\circ b \\ z & t \end{pmatrix}$ becomes orthogonal. Prove that the orthogonality is equivalent to a system of equations

$$z\circ {}^t a + t\circ {}^t b = 0, \qquad z\circ {}^t z + t\circ {}^t t = E.$$

Thus rows of the block $(z\ t)$ form an orthonormal basis of the subspace of $\mathbb{R}^n$ orthogonal to the rows of the block $(a\ b)$. Finally, if $\ker A = 0$ ($\Leftrightarrow \ker {}^t A = 0$), then A possesses a unique orthogonal PD factor; otherwise there is a one-to-one correspondence between these factors and orthobases of $\ker {}^t A$.

E6.7

Operators A and B are similar if and only if the sets of matrices of A and B with respect to the spatial bases are the same, subject to a different order only. In particular, this holds for the Jordan matrices. Consequently, in $\mathbb{C}^n$, A and B are similar if and only if they have the same (considering multiplicity) eigenvalues and the direct sum of root subspaces of A corresponding to any its eigenvalues λ is isomorphic to that of B. Next, if A has matrix ς with respect to a basis $e_1,\ldots$, then ${}^t\overline{A}$ has matrix ${}^t\overline{A}$ with respect to the dual basis $e_1^*, \ldots$ (the duality means that $\langle e_i^*,\ e_j\rangle = \delta_{ij}$). Deduce from these statements the following condition of similarity of A, ${}^t\overline{A}$: it occurs if and only if the Jordan boxes of A corresponding to a complex conjugate pair of its eigenvalues can be combined in double boxes such as $\begin{pmatrix} \lambda & & 1 & & & \\ & \overline{\lambda} & & 1 & & \\ & & \lambda & & 1 & \\ & & & \overline{\lambda} & & \ddots \\ & & & & \ddots & \end{pmatrix}$. Any real-valued matrix has the complex Jordan canonical form satisfying this condition. Conversely, the indicated box is brought to the Jordan box over reals $\begin{pmatrix} \mathrm{Re}\lambda & -\mathrm{Im}\lambda & 1 & & & \\ \mathrm{Im}\lambda & Re\,\lambda & & 1 & & \\ & & \mathrm{Re}\lambda & -\mathrm{Im}\lambda & 1 & \\ & & \mathrm{Im}\lambda & Re\,\lambda & & \ddots \\ & & & & \ddots & \end{pmatrix}$ by a basis substitution

$$e_1,\ e'_1,\ e_2,\ e'_2,\ \ldots \mapsto \frac{e_1+e'_1}{2},\ \frac{e'_1-e_1}{2i},\ \frac{e_2+e'_2}{2},\ \frac{e'_2-e_2}{2i},\ \ldots.$$

In turn, a matrix of real entries can be brought to the preceding form by means of a similitude transformation over the reals. Deduce from it the similarity of A and ${}^t A$ taking into account that if A has matrix ς with respect to basis $e_1,\ldots$, then ${}^t A$ will have matrix ${}^t\varsigma$ with respect to the dual basis $e_1^*, \ldots$ and the equality

$$\begin{pmatrix} 1 & & & & \\ & -1 & & & \\ & & 1 & & \\ & & & -1 & \\ & & & & \ddots \end{pmatrix} \circ \begin{pmatrix} a & -b & 1 & & \\ b & a & & 1 & \\ & & a & -b & 1 \\ & & b & a & \ddots \\ & & & & \ddots \end{pmatrix} =$$

$$\begin{pmatrix} a & b & 1 & & \\ -b & a & & 1 & \\ & & a & b & 1 \\ & & -b & a & \ddots \\ & & & & \ddots \end{pmatrix} \circ \begin{pmatrix} 1 & & & & \\ & -1 & & & \\ & & 1 & & \\ & & & -1 & \\ & & & & \ddots \end{pmatrix}.$$

This completes the proof of both real- and complex-valued cases.

E6.8

Obviously, arcwise connectedness is a kind of binary equivalence relation on a topological space so that distinct components do not intersect, and the space is a disjunctive union of the components. These components may be not closed, as appears from the following example (Gelbaum and Olmsted 1964). Find arcwise connected components of a planar set

$$T := \{(x,y) : \ x>0, \ y = \sin(1/x)\} \cup \{(x,y) : \ x = 0, -1 \le y \le 1\}.$$

The same example shows the failure of the converse of the lemma in section H6.8. (Why?)

We can prove this lemma in three steps as follows.

(1) A real segment is connected. Let us assume that $[0,1] = U_1 \cup U_2$, where U_i are open in [0,1] and do not intersect and U_1 is nonempty. U_1 is a disjunctive union of intervals. The endpoints of the intervals, which are internal for [0,1], belong to U_2. Hence, there are no such points. (Why?) Therefore, $U_1 = [0,1]$ and U_2 is empty.
(2) A continuous image of a connected space is connected. (Why?)
(3) Readers must formulate and complete this part by themselves.

Thus, we will show that symmetric operators in $GL(n,\mathbb{R})$ [and self-adjoint ones in $GL(n,\mathbb{C})$] with positive eigenvalues form an arcwise connected subset if we present an arc in this subset connecting its arbitrary element with the identity operator. Based on a diagonal presentation $S = U^{-1} \circ D \circ U$, $D = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix}$, a desired arc is

$$\gamma(t) := U^{-1} \circ \begin{pmatrix} t+(1-t)d_1 & & \\ & \ddots & \\ & & t+(1-t)d_n \end{pmatrix} \circ U, \quad t \in [0,\,1].$$

With a similar technique, using the spectral theorem for orthogonal (resp. unitary) operators, prove that the unitary groups are connected and the orthogonal groups cannot have more than two connected components.[8] To be precise, the subsets of the orthogonal group consisting, respectively, of the operators that

[8] The unitary spectral theorem states that a unitary $n \times n$-matrix is diagonalizable in a proper orthonormal basis in $\mathbb{C}^n$ and that its eigenvalues are located on the unit circle in $\mathbb{C}$. In turn, the orthogonal spectral theorem claims that an orthogonal matrix has a block-diagonal form in an appropriate orthonormal basis, with one- or two-dimensional blocks being, respectively, ± 1 and rotations as $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$. Note that a pair of "$-1$" blocks is equivalent to a two-dimensional block, which is a rotation by 180°, so that orthogonal matrices that preserve (invert) orientation have spectral representations with no (resp. one) "-1" block.

The orthogonal spectral theorem is reduced to the unitary one using complexification. In turn, the unitary spectral theorem may be proved by the following general method. Let a group G act on a vector space L over a field F by linear operators; that is, there is a group homomorphism (referred to as a **linear representation**, or just **representation** for short) $\rho\colon G \to GL(L)$, so that the action $g\colon L \to L$ ($g \in G$) is equated to the linear operator $\rho(g)$. A subspace $K \subseteq L$ is called a ***G*-invariant subspace**, or ***G*-subspace**, when $gK \subseteq K$, $\forall g \in G$. The representation is called **irreducible** if space L does not have proper (not equal to ether $\{0\}$ or itself) G-subspaces. The following claim is evident.

Proposition 1 *A finite-dimensional representation (*$\dim L < \infty$*) has a G-subspace K, irreducible with respect to the representation* $g \mapsto \rho(g)\,|_K$.

For two representations $\rho_i\colon G \to GL(L_i)$ $(i = 1,2)$ a ***G*-operator** is a linear operator $A\colon L_1 \to L_2$, which makes the diagrams $\begin{matrix} A: & L_1 & \to & L_2 \\ & \downarrow g & & \downarrow g \\ A: & L_1 & \to & L_2 \end{matrix}$ commutative, $\forall g \in G$.

Lemma 1 *Kernels and images of G-operators are G-subspaces.* (Why?)

An invertible G-operator is also called an **equivalence** of the two representations.

Lemma 2 *If a basic field F is algebraically closed,* then *an equivalence of irreducible finite-dimensional representations is defined uniquely, within proportionality.*

Indeed, let $A\colon L_1 \to L_2$ and $B\colon L_1 \to L_2$ be equivalences. A polynomial $\chi(\lambda) := \det(A - \lambda B)$ has a root $\lambda_0 \in F$. A G-operator $A - \lambda_0 B$ is not invertible; therefore, it is zero, $A = \lambda_0 B$. (Why?) **QED**.

Corollary *If a basic field F is algebraically closed, then irreducible finite-dimensional representations of Abelian groups are one-dimensional.*

Indeed, consider the representation $\rho\colon G \to GL(L)$. The operators $\rho(g)$ ($g \in G$) are (auto-) equivalences (why?), so they must be proportional to id_L, that is, be homothetic transformations. Hence, all subspaces of L are invariant, and therefore $\dim L = 1$. **QED**.

So far, we have enumerated common results for all representations. A special result of unitary representations [operators $\rho(g)$, $\forall g \in G$ are unitary] is as follows.

Proposition 2 *For a unitary representation, the orthogonal complements to G-subspaces are G-subspaces.*

(Why?) Considering the self-representation [$\rho(g) = g$, $\forall g \in G$] of the cyclic group G generated by a unitary operator U, and using this proposition in combination with the previous one and the preceding corollary, we will be able to prove the spectral theorem for U by induction on the spatial dimension.

preserve orientation and those that invert it are arcwise connected. Using PD conclude that $GL(n,\mathbb{C})$ are connected and $GL(n,\mathbb{R})$ cannot have more than two connected components. Finalize the description of the connected components of $GL(n,\mathbb{R})$ (and orthogonal groups) by showing that these groups are not connected. For this, consider a continuous map (which is in fact a group homomorphism) $M \mapsto \det M$. If $GL(n,\mathbb{R})$ were connected, so would the image of this map be (why?); but it is not!

E6.9

On the proof of necessity in section H6.9.

$n = 2$:

Use of SVD for factorization as described in section H6.9. A group homomorphism $M \mapsto \det M$ onto $\mathbb{Z}(2) = \{1, -1, *\}$ (see section E6.8) shows that $SO(n)$ is a subgroup of index two in $O(n)$ for any n; hence, for any fixed orientation-reversing operator, say a diagonal (with respect to the standard coordinate orthobasis) operator $R = \begin{pmatrix} -1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$ (which geometrically corresponds to the reflection in the coordinate hyperplane spanned on the second,..., nth axes), we have $O(n) = R \circ SO(n) = SO(n) \circ R$. This yields an equivalent form of SVD as $A = F \circ X \circ D \circ Y \circ G$, where D is diagonal (still with nonnegative diagonal elements, but we will not use that), $F,G \in SO(n)$ and each of X, Y is either R or the identity operator. Obviously, $X \circ D \circ Y$ also is a diagonal operator, and we will denote it by D.

D commutes with any rotations if A does. Indeed, for a rotational operator T we have

$$T^{-1} \circ A \circ T = A \quad \Rightarrow \quad F \circ T^{-1} \circ D \circ T \circ G = F \circ D \circ G,$$

and therefore $T^{-1} \circ D \circ T = D$. (Furnish the details.)

D is a dilatation. Indeed, a straightforward computation shows that a diagonal operator commutes with $\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$, where $\varphi \neq 0, \pi \bmod 2\pi$ if and only if the diagonal elements are equal.

$n \geq 3$:

A retains $\Pi^{\perp}$ invariant if A commutes with T. We have for $x \in \Pi^{\perp}$ $Ax = A(Tx) = T(Ax)$, and so $Ax \in \Pi^{\perp}$ because the fixed set of T is exactly $\Pi^{\perp}$.

A is a dilatation if it retains every one-dimensional subspace invariant. For two vectors $x,y \in \mathbb{R}^n$ we have, $Ax = \lambda_x \cdot x$, $Ay = \lambda_y \cdot y$ and $A(x + y) = \lambda_{x+y} \cdot (x + y)$; thus for any linearly independent x and y we derive $\lambda_x = \lambda_{x+y} = \lambda_y$. **QED**. (We leave it to the reader to fill in the details.)

E6.10

On the proof of the facts about ker A and im A for a normal operator A. Using Lemma 1 from section E6.3 or its Hermitian analog we have

$$\ker A = (\operatorname{im} A^*)^\perp = [\operatorname{im}(A^* \circ A)]^\perp = [\operatorname{im}(A \circ A^*)]^\perp = (\operatorname{im} A)^\perp = \ker A^*,$$
$$\operatorname{im} A = (\ker A)^\perp = \operatorname{im} A^*.$$

On the characterization of the normal operator via commutativity of PD factors.

Sufficiency. Both PD forms are brought by the same factors in different orders: $A = F \circ S = S \circ F$. Hence, by virtue of the uniqueness of S_1 and S_2, $S_1 = S = S_2$, and using explicit formulas in section P6.2**, $A^* \circ A = {S_1}^2 = {S_2}^2 = A \circ A^*$. **QED**.

Necessity. By virtue of the explicit formula in section P6.2** and Lemma 1 in section E6.3 or its Hermitian analog, for any linear operator A, its PD factor S_1 acts as a zero operator on ker A and maps its orthogonal complement im A^* isomorphically onto itself; therefore, F_1 has to map im A^* onto im A, and since it is an orthogonal (unitary) operator, it will also map their orthogonal complements onto one another, ker A onto ker A^*. Similarly, S_2 acts as a zero operator on ker A^* and maps its orthogonal complement im A isomorphically onto itself; therefore, F_2 has to map ker A onto ker A^*, and since it is an orthogonal (unitary) operator, it will also map their orthogonal complements onto one another, im A^* onto im A. Thus, both F_1 and F_2 map ker A onto ker A^* and im A^* onto im A.

Hence, for a normal operator A, S_1 $(= S_2)$ maps a subspace im A isomorphically onto itself, and F_1 and F_2 retain subspaces ker A $(=$ ker $S_1)$ and im A $(=$ im $S_1)$ invariant. Apply now the uniqueness of an orthogonal (unitary) factor for a nondegenerate operator (section P6.2**) to show that S_1 commutes with F_1 and F_2.

Alternatively, readers can prove the commutativity of PD factors by a short deduction using that $S_1 = S_2$.

On the proof of the spectral theorem for normal operators.

Necessity. Prove the following simple lemma.

Lemma *The total eigenspace corresponding to the same eigenvalue of a linear operator is an invariant subspace for any linear operator commuting with that one.*

Decompose the space into an orthogonal sum of the eigenspaces corresponding to the distinct eigenvalues of the PD factor S_1 $(= S_2)$. Each summand is invariant for

F_1 (= F_2), and we are able to apply the orthogonal (unitary) spectral theorem for that summand. This will yield the desired representation of the diagonal blocks, and, obviously, the modulus and phase factors of the eigenvalues of A will be equal to the corresponding eigenvalues of S_1 and F_1.

E6.11

On the proof of the equalities $\| A \circ B \| = \| B \circ A \| = \|A\|$ with orthogonal (unitary) B. We have $\| A \circ B \| \leq \|A\| \cdot 1 = \|A\|$ and $\|A\| = \| A \circ B \circ B^{-1} \| \leq \| A \circ B \| \cdot 1 = \| A \circ B \|$. The second equality can be proved similarly.

On the computation of the normal operator's norm. It is clear that $\|A|_L\| \leq \|A\|$ for the restriction of A to any subspace L. If the space is decomposed into an orthogonal sum of invariant subspaces L_i, then for any vector x we have $x = \sum x_i$ ($x_i \in L_i$, $\forall i$) and $(|Ax|/|x|)^2 = (\sum |Ax_i|^2)/(\sum |x_i|^2)$. Hence, the inverse inequality $\|A\|^2 \leq \max_i \|A|_{L_i}\|^2$ immediately follows from an elementary lemma.

Lemma *If* $a_1, \ldots, a_p, b_1, \ldots, b_p > 0$, then $\min (a_i/b_i) \leq \sum a_i / \sum b_i \leq \max (a_i/b_i)$, *and these are strict inequalities unless* $a_1/b_1 = \ldots = a_p/b_p$.

Prove this lemma.

A proof of the facts that for a linear operator on a two-dimensional Euclidean (Hermitian) space that either is diagonalizable but not in an orthogonal basis or has one eigendirection only, the norm is greater than the maximal modulus of its eigenvalue and $\|A^n\| \leq \|A\|^n$, $\forall n = 2,3,\ldots$. Consider the PD $A = F_1 \circ S_1$. We may assume that $S_1 \neq 0$ (otherwise the claim that is being proved is trivial), and so S_1 has distinct eigenvalues $\mu_1 > \mu_2$ (≥ 0). Indeed, if S_1 is proportional to the identity operator, then A has a common eigenspace with F_1; since the spatial dimension is two and F_1 has an eigenvector, F_1 is diagonalizable and, as it is an orthogonal (unitary) operator, it is diagonalizable in an orthogonal basis; but A is diagonalizable in the same basis, which contradicts the assumption. Let M_i be the eigenspaces corresponding to μ_i ($i = 1,2$). M_1 cannot be an eigenspace of A, as otherwise it would also be an eigenspace of F_1, so both M_i are eigenspaces of F_1, so both M_i are eigenspaces of A, which contradicts the assumption. (The same argument shows that M_2 is not an eigenspace of A if $\mu_2 > 0$, but we do not utilize this fact in this proof.) Now, (1) for any vector x $|S_1 x| = |Ax|$ and (2) the value of $|S_1 x|$ achieves its maximum on $\{x: |x| = 1\}$ if and only if $x \in M_1$. Thus, since M_1 is not an eigenspace of A, for unit vectors e_1 and e, such that $e_1 \in M_1$ and e belongs to an eigenspace of A, $\|A\| = |Ae_1| > |Ae| = |\lambda|$, where λ is the corresponding eigenvalue of A. Since $F_1 e_1 \notin M_1$ (as M_1 is not an eigenspace of F_1), for any unit vector e and $n = 2,3,\ldots$ we have

$$|A^n e| = |(F_1 \circ S_1 \circ \ldots \circ F_1 \circ S_1)e|^n < |S_1 e_1|^n = \|A\|^n,$$

and therefore $\|A^n\| = \max |A^n e| < \|A\|^n$, which finalizes the proof. (Furnish all details.)

Completing the Solution

S6.3

Proof of Lemma 1. We have, taking into account the nondegeneracy of the inner (scalar) product, $<Wx,y> = 0, \forall y \Leftrightarrow x \in \ker W$; in turn, $<x,{}^tWy> = 0, \forall y \Leftrightarrow x \perp \text{im } {}^tW$, which yields the orthogonal decomposition $\mathbb{R}^n = \text{im } {}^tW \bigoplus \ker W$,[9] and, swapping ${}^tW$ and W and taking into account the equality ${}^t({}^tW) = W$,[10] yields the dual orthogonal decomposition $\mathbb{R}^m = \text{im } W \bigoplus \ker {}^tW$.

The equalities for the ranks follow from the preceding decompositions by virtue of equalities $\text{rk } A = \dim \text{im } A = \dim \text{dom } A - \dim \ker A$ (for any linear operators on finite-dimensional spaces). **QED**.

In this proof, the equality $\text{rk } {}^tW = \text{rk } W$ was derived without employing matrix representations of linear operators. *Therefore, the equality of matrix row and column ranks was not used*. This makes it possible to derive that equality from the equality $\text{rk } {}^tW = \text{rk } W$ because for the by-coordinate inner (scalar) product in $\mathbb{R}^n$ and $\mathbb{R}^m$, the matrices for W and ${}^tW$ with respect to the coordinate bases are transposed with each other. (We leave it to the reader to complete the details.)[11]

Proof of Lemma 2. (ii) ⇒ (i) is trivial.

(i) ⇒ (iii). Let, for an orthobasis $e_1,\ldots$ of im ${}^tW$, $We_1,\ldots$ be an orthobasis of im W. We must verify that ${}^tWWe_i = e_i, \forall i$, or, in other words, that $< {}^tWWe_i,e_j > = \delta_{ij}$. But we have

$$\left\langle {}^tWWe_i, e_j \right\rangle = \left\langle We_i, We_j \right\rangle = \delta_{ij}.$$

QED.

(iii) ⇒ (ii). Let $e_1,\ldots$ be an orthobasis of im ${}^tW$. We must verify that $We_1,\ldots$ is an orthobasis of im W. But, assuming (iii), we have

$$\left\langle We_i, We_j \right\rangle = \left\langle {}^tWWe_i, e_j \right\rangle = \delta_{ij}.$$

QED.

[9] Taking into account that a subspace im ${}^tW$ coincides with its closure, due to the finite dimensionality.

[10] Which holds due to the finite dimensionality also.

[11] A similar proof of the equality of row and column ranks is valid for the matrices over any field and is generalized for wider classes of rings. (Which ones?) A different proof is discussed in the problem group "A Property of Orthogonal Matrices" (section E8.11).

S6.7

The proof is completed taking into account the similarity between the matrices

$\begin{pmatrix} \lambda & 1 & & 0 \\ 0 & \ddots & \ddots & \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda \end{pmatrix}$ and $\begin{pmatrix} \lambda & 0 & \cdots & 0 \\ 1 & \ddots & \ddots & \vdots \\ & \ddots & \ddots & 0 \\ 0 & & 1 & \lambda \end{pmatrix}$. (They are similar since an operator defined by one of them using an ordered basis $e_1,\ldots,e_m$ has the second one as its matrix with respect to the same basis in the inverse order $e_m,\ldots,e_1$.)

S6.8

On the proof of the lemma.

(1). Endpoints of the intervals, which are internal for [0,1], enter U_2 with their neighborhoods. On the other hand, these endpoints are limits for internal points of those intervals, and therefore U_2 intersects with U_1.

(2). A continuous image of a connected space is connected because a map between topological spaces is continuous if and only if preimages of any open sets themselves are open sets. (Provide the details.)

(3). Any arc in a topological space is connected (as proved in the previous step). But if a disconnected topological space $U_1 \cup U_2$ (where U_i are nonintersecting non-empty open subsets) were arcwise connected, then an arc γ between two points $x_i \in U_i$ ($i = 1,2$) would be disconnected (because it can be decomposed into two sets im $\gamma \cap U_1$ and im $\gamma \cap U_2$ open in γ). (Fill in the details.)

On the number of connected components of the orthogonal and unitary groups. A diagonal matrix D of diagonal elements on the unit circle is connected with the identity matrix by an arc $t \mapsto D(t)$ entirely contained in $U(n)$. Indeed, a complex number $e^{i\varphi}$ is connected with 1 by an arc of the unit circle $t \mapsto e^{i\varphi\,(1-t)}$ ($t \in [0,1]$). This yields an arc between an arbitrary unitary operator and E as $t \mapsto C^{-1} \circ D(t) \circ C$, and thus $U(n)$ is connected. Similarly, the orthogonal spectral theorem provides arcs between arbitrary orthogonal operators and one of $E = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & -1 \end{pmatrix}$, entirely contained in $O(n)$. [Furnish the details using the fact

that a multiplication by the second of these matrices maps the arcs in $SO(n) = \{T \in O(n)\colon \det T = 1\}$ onto the arcs in $\{T \in O(n)\colon \det T = -1\}$ and vice versa.] Hence, $O(n)$ may have at most two connected components; actually it has exactly two (and these two operators belong to different components, as follows from the second step in the foregoing proof of the lemma applied to a continuous map "det" (taking the determinant) on the topological space $O(n)$.

S6.9

A description of the linear operators on $\mathbb{R}^n$ commuting with $SO(n)$ using arguments of analysis. If a linear operator A commutes with $SO(n)$, then a quadratic form $f(x) = \langle Ax, x\rangle$ on $\mathbb{R}^n$ (where $\langle .,. \rangle$ is the scalar product) is invariant with respect to $SO(n)$:

$$\langle A(Tx), Tx\rangle = \langle T(Ax), Tx\rangle = \langle Ax, x\rangle, \qquad \forall T \in SO(n)$$

[in other words, f is constant on any orbit of $SO(n)$]. The orbits are concentric hyperspheres with their centers at the origin (the *regular* orbits) and the origin itself (a *singular* orbit), so that the orbits in a **general position** are $n - 1$-dimensional. Let us pick a subvariety of $\mathbb{R}^n$ having a complementary dimension equal to 1 that intersects with each orbit; a rectilinear ray with the vertex at the origin satisfies these conditions. It is a manifold with a boundary, and f is analytic on it including the boundary. Using the uniqueness of a Taylor-series expansion of f at the origin and the fact that f is homogeneous of degree two, we deduce that on the ray $f(x) = \lambda \cdot \langle x, x\rangle$, $\lambda = const$. (We could arrive at the same result without analyticity and Taylor-series expansions using the homogeneity of f and $\langle .,. \rangle$.) The preceding equality with the same λ will hold everywhere because of the invariance of f and $\langle .,. \rangle$ with respect to $SO(n)$: $f(Tx) = f(x) = \lambda \cdot \langle x, x\rangle = \lambda \cdot \langle Tx, Tx\rangle$, $\forall T \in SO(n)$.

Using a symmetric linear operator (symmetrization) $S(B) = \frac{1}{2}\,({}^tB + B)$, we have $\langle Bx, x\rangle = \langle S(B)x, x\rangle$ for any linear operator B. Applying the *symmetric operator spectral theorem* and taking into account the fact that $\langle .,. \rangle$ is nondegenerate, we will obtain that $S(B) = 0$ (i.e., B is skew-symmetric) if $\langle Bx, x\rangle$ equals zero identically. Thus we have established the skew symmetry of $B = A - \lambda E$. By the *skew-symmetric operator spectral theorem*, A has the following block-matrix representation with respect to an appropriate orthobasis (which is a *Darboux basis* for B; see section H10.27, in the problem group "One-Parameter Groups of Linear Transformations" below):

$$A = \begin{pmatrix} \lambda & \mu_1 & & & & & & \\ -\mu_1 & \lambda & & & & & & \\ & & \ddots & & & & & \\ & & & \lambda & \mu_k & & & \\ & & & -\mu_k & \lambda & & & \\ & & & & & \lambda & & \\ & & & & & & \ddots & \\ & & & & & & & \lambda \end{pmatrix}.$$

Now we must prove that there are no two-dimensional blocks for $n \geq 3$. If any such block exists, then $\mu_1 \neq 0$. Denote e_i the ith element of our basis $e_i = {}^2(0, \ldots, 0, \underset{i}{1}, 0, \ldots, 0)$. Defining a rotation T so that $Te_2 = e_3$, $Te_3 = -e_2$, and $Te_i = e_i$, $\forall i \neq 2{,}3$, we will obtain $A(Te_1) = \lambda e_1 - \mu_1 e_2$, $T(Ae_1) = \lambda e_1 - \mu_1 e_3$, and therefore $A \circ T \neq T \circ A$, which finalizes the proof. (We leave it to the reader to fill in the details.)

A similar method could be applied to many other situations. For example, such a method makes it possible to prove a nontrivial theorem used in elasticity theory. Let us consider a vector space S_n of all symmetric bilinear (or, equivalently, quadratic) forms on the Euclidean space $\mathbb{R}^n$ ($n \geq 2$), which we will not distinguish from their matrices with respect to the standard coordinate orthobasis. S_n is a Euclidean space, as a subspace of the space of all $n \times n$-matrices with a scalar product $\langle U, V \rangle = \operatorname{tr}({}^tU \circ V) = \sum_{i,j\,=\,1,\ldots,\,n} U_{ij}V_{ij}$ [which for symmetric matrices looks like $\operatorname{tr}(U \circ V)$]. The scalar product is $O(n)$-similitude-invariant: $\langle U, V \rangle = \langle T^{-1} \circ U \circ T, T^{-1} \circ V \circ T \rangle$, $\forall T \in O(n)$, and also, $\mathcal{S}_n$ is $O(n)$-similitude invariant: $T^{-1} \circ U \circ T = {}^tT \circ U \circ T \in \mathcal{S}_n$, $\forall T \in O(n)$, $\forall U \in \mathcal{S}_n$. We will call symmetric linear operators on $\mathcal{S}_n$ *Hooke's operators*, as in the elasticity theory these operators are related to elastic materials and are employed for the formulation of **Hooke's law**. [Hooke's operators form a ½ $N(N + 1)$-dimensional vector space, where $N = \dim \mathcal{S}_n = ½\, n(n + 1)$.[12]] We also say that a Hooke's operator H is *isotropic* if it commutes with the $O(n)$-similitude action, i.e., $T^{-1} \circ (HU) \circ T = H(T^{-1} \circ U \circ T), \forall T \in O(n), \ \forall U \in \mathcal{S}_n$.

Theorem *A Hooke's operator is isotropic if and only if there exist real numbers* λ *and* μ *such that*[13]

$$HU = \lambda \cdot \operatorname{tr} U \cdot E + 2\mu \cdot U, \qquad \forall U \in \mathcal{S}_n$$

(Landau and Lifshitz (1987). Therefore, for any $n \geq 2$ the isotropic operators form a two-dimensional subspace in the ½$N(N + 1)$-dimensional space of all Hooke's operators.

Proof of theorem.

Sufficiency. A linear operator defined by the preceding equation evidently commutes with the similitude action of any linear operator, and its symmetry is verified directly:

$$\langle HU, V \rangle = \lambda \cdot \operatorname{tr} U \cdot \langle E, V \rangle + 2\mu \cdot \langle U, V \rangle = \lambda \cdot \operatorname{tr} U \cdot \operatorname{tr} V + 2\mu \cdot \langle U, V \rangle.$$

[12] In the most frequently encountered case, $n = 3$, the space of Hooke's operators has a dimension of 21.

[13] In the elasticity theory, λ and μ are referred to as **Lamé constants**.

Necessity. We will proceed using arguments that are similar to the arguments used previously in the analytical proof. For an isotropic Hooke's operator H, let us consider the following quadratic form on $\mathcal{S}_n$: $\mathcal{F}(U) = ½ \langle HU, U\rangle$.[14] H produces the gradient of $\mathcal{F}$, $HU = \nabla\mathcal{F}(U)$; hence, we must prove that

$$\mathcal{F}(U) = {}^1\!/_2\lambda \cdot (\mathrm{tr}U)^2 + \mu \cdot \langle U, U\rangle, \qquad \forall U \in S_n \ \ (\lambda, \mu = const). \qquad (*)$$

$\mathcal{F}$ is invariant with respect to the $O(n)$-similitude action on $\mathcal{S}_n$:

$$\begin{aligned}\mathrm{tr}\big(\big[H\big(T^{-1} \circ U \circ T\big)\big] \circ \big(T^{-1} \circ U \circ T\big)\big) &= \mathrm{tr}\big(\big[T^{-1} \circ (HU) \circ T\big] \circ \big(T^{-1} \circ U \circ T\big)\big) \\ &= \mathrm{tr}\big(T^{-1} \circ (HU) \circ U \circ T\big) = \mathrm{tr}((HU) \circ U)\end{aligned}$$

[in other words, $\mathcal{F}(U)$ is constant on any orbit of this action], and the same is true of tr U and $\langle U, U\rangle$. Therefore, it is sufficient to establish (*) for all U of a subset of $\mathcal{S}_n$ intersecting with each orbit (perhaps several times). Let us take for that subset a vector subspace of $\mathcal{S}_n$ having the least possible dimension $N - \dim O(n) = ½\, n\,(n + 1) - ½\, n(n - 1) = n$. By the spectral theorem for symmetric matrices, any $O(n)$-similitude orbit in $\mathcal{S}_n$ contains a diagonal matrix, which is determined within arbitrary permutations of the diagonal elements; thus, the subspace $\mathcal{D}_n \subset \mathcal{S}_n$ consisting of all diagonal matrices satisfies our conditions, and we need only establish (*) for $U \in \mathcal{D}_n$.

$\mathcal{F}(U)$ is a polynomial of degree two in the U entries, so that for $U \in \mathcal{D}_n$ it is a polynomial of degree two in the diagonal elements $\omega_1, \ldots, \omega_n$. Since $\mathcal{F}$ is invariant with respect to the permutations of ω_i, it is a polynomial in the set of elementary symmetric polynomials $\sigma_1(\omega_1, \ldots, \omega_n), \ldots, \sigma_n(\omega_1, \ldots, \omega_n)$ (by the symmetric polynomial theorem).

In the preceding formulation, $\sigma_1, \ldots, \sigma_n$ may be substituted by any symmetric polynomials $p_1(\omega_1, \ldots, \omega_n), \ldots, p_n(\omega_1, \ldots, \omega_n)$ of degrees $1, \ldots, n$, respectively, such that the expression of each p_k via $\sigma_1, \ldots, \sigma_n$ contains a term proportional to σ_k with a nonzero coefficient of proportionality. (Why?)

The polynomials $p_k = \sum \omega_i^{\,k}$ have the required form; the term of p_k proportional to σ_k is $(-1)^{k-1}k{\cdot}\sigma_k$. To establish that, readers can proceed by the same method as in section S1.1, from the "Jacobi Identities and Related Combinatorial Formulas"

[14] In the elasticity theory, $\mathcal{F}$ is referred to as the **Helmholtz potential** (or **free**) **energy**.

problem group above (in section S1.1, this method was applied in a special case $k = n$).[15]

In fact, $\mathcal{F}$ must depend only on p_1, p_2; and $\mathcal{F}(U) = ½\,\lambda \cdot [p_1(\omega_1,\ldots,\omega_n)]^2 + \mu \cdot p_2(\omega_1,\ldots,\omega_n)$ ($\lambda,\mu = const$) if p_1 and p_2 are homogeneous since $\mathcal{F}(U)$ is homogeneous of degree two.

We have $\operatorname{tr} U^k = \sum \omega_i^{\,k}$ ($U^k = U \circ \ldots \circ U_k, \quad k = 1,\ldots,n$), which is evident for $U \in \mathcal{D}_n$, and putting $p_k = \operatorname{tr} U^k$ in $\mathcal{F} = ½\,\lambda \cdot {p_1}^2 + \mu \cdot p_2$ yields (*), which completes the proof. (Furnish all the details.)

We leave it to the reader to prove the corollaries as formulated below that are important for the elasticity theory. $\mathcal{S}_n$ is an orthogonal sum of two subspaces consisting, respectively, of matrices with zero traces and dilatations, $\mathcal{S}_n = R_n \bigoplus^{\perp} \mathcal{C}_n : U = R + C$ ($C = (\operatorname{tr} U/n)\cdot E$, $R = U - C$).[16] Any isotropic Hooke's operator has R_n and $\mathcal{C}_n$ as its eigenspaces, with the eigenvalues, respectively, 2μ and nK, where $K = \lambda + 2\mu/n$.[17] Thus, an isotropic Hooke's operator is invertible if and only if μ and K are distinct from zero, and in this case, the inverse operator can be determined by the formula $H^{-1}S = \frac{1}{2\mu}\left[S - \frac{1}{nK} \cdot (\operatorname{tr} S)E\right]$. H^{-1} maps a diagonal matrix

[15] An explicit determination of coefficients of proportionality is unnecessary; to prove their distinctness from zero, one can proceed as follows. A map P: $(\omega_1,\ldots,\omega_n) \mapsto (\sum \omega_i,\ldots,\sum {\omega_i}^n)$ is a composition of Vieta's map V: $(\omega_1,\ldots,\omega_n) \mapsto (\sigma_1,\ldots,\sigma_n)$ and a map S: $(\sigma_1,\ldots,\sigma_n) \mapsto (\sum \omega_i,\ldots,\sum {\omega_i}^n)$. The Jacobi matrix S_* is polynomial and triangular, with constant diagonal elements equal to the desired coefficients. The Jacobi matrix P_* is a composition of matrices V_* and S_*. P_* is nondegenerate for (and only for; why?) distinct $\omega_1,\ldots,\omega_n$, and hence S_* is nondegenerate for distinct $\omega_1,\ldots,\omega_n$; but since the diagonal elements of S_* are constant, they are distinct from zero. **QED**. (We leave it to the reader to fill in the details.) Also, we have determined the critical set of Vieta's map (the set of all critical, or singular, points of V, i.e., the points that V_* degenerates). (We leave it to the reader to formulate this result. A different proof of this result uses the explicit form of the Jacobian of Vieta's maps, $\det V_* = \det\left(\sigma_{i-1}(\omega_1,\ldots,\widehat{\omega_j},\ldots,\omega_n)\right) = \prod_i (\omega_i - \omega_j)$, that can be derived using section S1.4 from the "Jacobi Identities and Related Combinatorial Formulas" problem group. In addition, readers familiar with elements of complex analysis can prove the foregoing result using the implicit function theorem's inversion, which states that a bijective holomorphic map $\mathcal{U} \to \mathcal{V}$, where $\mathcal{U}, \mathcal{V}$ are open sets in $\mathbb{C}^n$, is biholomorphic ($\Leftrightarrow$ nonsingular), and the facts about polynomials with complex coefficients – that a polynomial of degree n has exactly n complex roots and all roots of the polynomial with the leading coefficient equal to one will remain bounded when all coefficients remain bounded. Furthermore, a point $(\omega_1,\ldots,\omega_n)$ with distinct complex coordinates ω_i is in an open polydisk $\mathcal{U} = \mathcal{D}_1 \times \ldots \times \mathcal{D}_n$, where $\mathcal{D}_i$ are disjoint open disks in $\mathbb{C}$ with centers ω_i, respectively. $\mathcal{U}$ intersects at most once with any orbit of the group of all permutations of coordinates (the **symmetric group**) S_n; hence, the restriction $V|_u$ is injective. Next, $V(\mathcal{W})$ are open for open $\mathcal{W}$: if for $\omega \in \mathcal{W}$ the point $\sigma = (\sigma_1,\ldots,\sigma_n) = V(\omega)$ does not possess a neighborhood in $\mathbb{C}^n$ every point of which has a preimage in $\mathcal{W}$, then we can find a convergent sequence $\omega^{(v)} \to \omega'$ so that the points $\sigma^{(v)} = V(\omega^{(v)})$ do not have preimages in $\mathcal{W}$ and $\sigma^{(v)} \to \sigma$ as $v \to \infty$; but since $A\omega' = \omega$ for an appropriate transformation $A \in S_n$, $A\omega^{(v)} \to \omega$, and so $A\omega^{(v)} \in \mathcal{W}$ for all sufficiently large v; however – contrary to the assumption – $A\omega^{(v)}$ (along with $\omega^{(v)}$) are preimages of $\sigma^{(v)}$. Thus, $\mathcal{U}$ is bijectively mapped onto an open set $\mathcal{V} = V(\mathcal{U})$. We leave it to interested reader to fill in the details.)

[16] In the elasticity theory, R and C are related to the **shear tensor** and the **compression tensor**, respectively.

[17] In this connection, μ is also called the **shear modulus** and K the **bulk elastic modulus**.

$$S = \begin{pmatrix} \ddots & & & & \\ & 0 & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & \ddots \end{pmatrix} \quad \text{onto} \quad U = \frac{1}{E}\begin{pmatrix} \ddots & & & & \\ & -\sigma & & & \\ & & 1 & & \\ & & & -\sigma & \\ & & & & \ddots \end{pmatrix}, \quad \text{where}$$

$E = \frac{2n^2\mu K}{n(n-1)K+2\mu}$, $\sigma = \frac{nK-2\mu}{n(n-1)K+2\mu} = \frac{n\lambda}{n(n-1(n-1))K+2\mu}$. [18] Expressions of λ, μ, and K via and σ are $\lambda = \frac{E\sigma}{(1+\sigma)[1-(n-1)\sigma]}$, $\mu = \frac{E}{2(1+\sigma)}$, $K = \frac{E}{n[1-(n-1)\sigma]}$. [19] For positive μ and K, σ lies between -1 and $(n-1)^{-1}$, taking these boundary values for $K = 0$ and $\mu = 0$, respectively.[20] Also for positive μ and K [more exactly, for a positive n $(n-1)K + 2\mu$], λ is positive if and only if σ is positive.[21] Interested readers will find a further development and discussions of related aspects in Landau and Lifshitz (1987) and references therein.

S6.10

On the characterization of the normal operator via commutativity of PD factors. Necessity. We must prove the commutativity of S_1 $(= S_2)$ with F_1 and F_2 on $\ker A$ and $\operatorname{im} A$. On $\ker A$ it is evident, and on $\operatorname{im} A$ it follows from equalities of the PD factors, as $S_1|_{\operatorname{im} A} = S_2|_{\operatorname{im} A}$ and $F_1|_{\operatorname{im} A} = F_2|_{\operatorname{im} A}$, for a nondegenerate normal operator $A|_{\operatorname{im} A}$. (We leave it to the reader to furnish the details.)

On the proof of the spectral theorem for normal operators. A proof of the lemma from section E6.10. Let X commute with Y and L be the total eigenspace of X corresponding to an eigenvalue λ. For $x \in L$, $X(Yx) = Y(Xx) = Y(\lambda x) = \lambda Yx$, and therefore $Yx \in L$ since L is the total eigenspace corresponding to λ.

S6.11

A proof of lemma from section E6.11. Assume that $a_1/b_1 \leq \ldots \leq a_p/b_p$. Sum the inequalities

[18] In the elasticity theory, these parameters are called, respectively, **Young's modulus** and the **Poisson coefficient**.

[19] In elasticity-theoretic considerations, λ, μ, K, and have the dimensionality of pressure (energy per unit volume) and σ is dimensionless.

[20] Values of σ close to $(n-1)^{-1}$ (occurring for $\mu/K \ll 1$) are characteristic for rubber and materials with a similar elastic behavior.

[21] The positiveness of σ is usual for natural materials; a material having $\sigma < 0$ would thicken under the axial extension!

$$\frac{a_1}{b_1}b_1 = a_1 \le \frac{a_p}{b_p}b_1, \quad \frac{a_1}{b_1}b_2 \le a_2 \le \frac{a_p}{b_p}b_2, \quad \ldots, \quad \frac{a_1}{b_1}b_p \le a_p = \frac{a_p}{b_p}b_p,$$

and then divide all by $b_1 + \ldots + b_p$. The resulting inequalities will be strict unless all of a_i/b_i are equal.

Establishing the equality $\|A\| = \|A^*\|$ using duality.

An operator $D \circ A^* \circ D^{-1}$ has the form $x^* \mapsto x^* \circ A$. Denoting by B the operator $x^* \mapsto x^* \circ A$ yields a diagram,

$$x \overset{D}{\mapsto} \langle x, . \rangle \overset{B}{\mapsto} \langle x, A(.) \rangle = \langle A^*x, \ . \rangle \overset{D^{-1}}{\mapsto} A^*x,$$

or, $A^* = D^{-1} \circ B \circ D$.

$|Dx|^* = |x| \Rightarrow \|D \circ A^* \circ D^{-1}\|^* = \|A^*\|$. Indeed,

$$\left\|D \circ A^* \circ D^{-1}\right\|^* = \max_{|x^*|^*=1} \left|\left(D \circ A^* \circ D^{-1}\right)x^*\right|^* = \max_{|x|=1} \left|(D \circ A^*)x\right|^* = \max_{|x|=1} |A^*x| = \|A^*\|.$$

$|x^*|^* = \left|D^{-1}x^*\right| \Rightarrow |x^*|^* = \max_{|x|=1} |x^*(x)| \left(= \max_{x \ne 0} |x^*(x)|/|x|\right)$. Considering an orthogonal sum $x = \alpha e + x'$, where e is a unit vector collinear with $D^{-1}x^*$,

$$|x^*(x)|/|x| = \left|\left\langle D^{-1}x^*, x\right\rangle\right|/|x| = \left|\alpha D^{-1}x^*\right| \Big/ \sqrt{\alpha^2 + |x'|^2},$$

which has the maximal value $|D^{-1}x^*|$ (obtained when $x' = 0$).

$|x^*|^* = \max_{|x|=1} |x^*(x)| \Rightarrow \|D \circ A^* \circ D^{-1}\|^* = \|A\|$. A deduction similar to the previous one shows that $|x^*|^* = \max_{|x|=1} |x^*(x)| \Rightarrow |x^*(x)| \le |x^*|^* \cdot |x|, \forall x$, and therefore $|(x^* \circ A)\,x| \le |x^*|^* \cdot |Ax| \le |x^*|^* \cdot \|A\| \cdot |x|$; hence, $|x^* \circ A|^* \le |x^*|^* \cdot \|A\|$, so $\|D \circ A^* \circ D^{-1}\|^* \le \|A\|$. The inverse inequality can be proved as follows. Let x be a unit vector such that $|Ax| = \|A\|$, and let e be a unit vector collinear with Ax and $f = De$. We have $|f|^* = 1$ and $|f(Ax)| = |\langle e, Ax\rangle| = |Ax| = \|A\|$; therefore, denoting the operator $D \circ A^* \circ D^{-1}$ as B, we have

$$||B||^* = ||B||^* \cdot |f|^* \cdot |x| \ge |Bf|^* \cdot |x| \ \ge \ |(Bf)x| = |f(Ax)| = ||A||.$$

Readers familiar with infinite-dimensional Euclidean (Hermitian) spaces (usually referred to as Hilbert spaces) can modify the foregoing proof to make it valid for continuous linear operators on these spaces. In this case, "sup" cannot be substituted by "max" because a unit vector x such that $|Ax| = \|A\|$ may not exist, but in any case there exists a sequence of unit vectors x_n such that $|Ax_n| \to \|A\|$ as $n \to \infty$, so we can take unit vectors e_n collinear with Ax_n, respectively. If the space is complete, then the duality mapping D defined similarly to the one defined in section H6.11 brings an algebraic and a topological isomorphism between this and the adjoint space, as stated by the **Riesz theorem**.[22] An incomplete space can be completed with respect to the Euclidean (Hermitian) metric that will not affect the adjoint space. [To prove this statement, we need to prove that a continuous linear functional f is uniquely continuously extendable to the completed

[22] The adjoint space is defined as a space of all *continuous* linear functionals, which for a finite-dimensional space is the same as the space of all linear functionals.

domain and that this extension has the same norm as f. Let a sequence x_n in the source space converge to a point of the completion. $f(x_n)$ is a Cauchy (fundamental) sequence because $|f(x) - f(y)| \leq \|f\| \cdot |x - y|$. Clearly, defining $f_{ext}(x) = \lim f(x_n)$ when $x = \lim x_n$ is the only option for a continuous extension of f to the completion of the space, and this definition yields a continuous linear functional on the completion that has the same norm as f.] We leave it to the interested reader to fill in the details.

Also for a finite- or infinite-dimensional vector space with a non-Euclidean (non-Hermitian) norm, the preceding arguments show that if A is a continuous linear operator, an operator $x^* \mapsto x^* \circ A$ in the adjoint space is continuous and possesses the same norm as A. Indeed, by the **Hahn-Banach theorem** the linear functionals f_n of the unit norm defined on the one-dimensional subspaces spanned, respectively, on Ax_n [so that $|f_n(Ax_n)| \to \|A\|$ as $n \to \infty$] are extendable to the whole space as continuous linear functionals of the unit norm.

2 × 2 Matrices That Are Roots of Unity

Problems

P7.0

Preliminaries. As readers know, a polynomial equation of degree n has at most n roots (considering multiplicity) in a number field containing its coefficients. How many roots does a polynomial equation have in a matrix ring? First, how many roots of degree n of $1 = E$, that is, matrices X, such that X^n=$X \circ \ldots \circ X_n = E$, are there? And how would one enumerate them? These and related questions that can be answered given a relatively modest level of knowledge on the reader's part are discussed in this problem group. That is, we will characterize the roots of E in a ring of 2×2 real-valued matrices (sections P7.1**, P7.2**, P7.3**, P7.4**, and P7.5**) and show some applications to other mathematical topics: matrix norm estimation (sections P7.6***, P7.7***, and 7.8***) and spectral analysis of three-diagonal Jacobi matrices (sections P7.9** and P7.10**).

In what follows, we will often not distinguish between linear operators and their matrices with respect to a fixed basis.

P7.1**

Obviously, n orthogonal operators on the Euclidean plane $\mathbb{R}^2$

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \qquad \theta = \theta_k^n = 2\pi k/n, \quad k = 0, \ldots, n-1,$$

which are rotations by angles θ_k^n, are roots of degree n of E. *Prove* that for odd n no other orthogonal roots of E exist. *Does* a similar statement also hold for even n?

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_7,

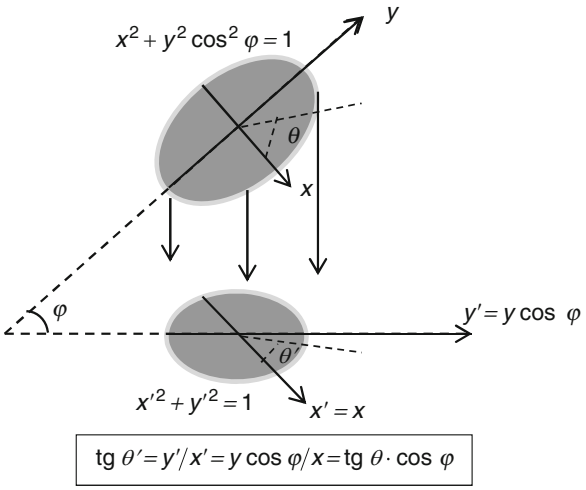


Fig. 1 Projection makes rotations irregular

*P7.2***

Matrices similar to roots of degree n of E also are such roots. *Prove* that any root of E in a ring of linear operators on a Euclidean (complex Hermitian) space is similar to an orthogonal (resp. unitary) root, or, which is the same thing, is itself orthogonal (resp. unitary) with respect to an appropriate Euclidean (resp. Hermitian) quadratic form (in other words, an inner, or scalar, product).

Rotations, relative to some scalar product on a plane, look like "irregular rotations" with respect to a different scalar product. This corresponds to the following geometrical picture: when a point rotates in a plane in the Euclidean space $\mathbb{R}^3$, its projection onto a nonparallel plane in the same space will rotate in an "irregular" way (Fig. 1). In practical terms, such situations are encountered in computerized tomography (CT) when radiographs and reconstruction are applied in different (nonparallel) planes.[1]

*P7.3***

Prove that a real-valued 2×2 matrix with nonreal eigenvalues of modulus 1 is orthogonal with respect to some scalar products in $\mathbb{R}^2$. *Show* that for a fixed matrix all these products are proportional to one another.

[1] A short glossary (including *radiograph*) and literature references related to CT are in section P5.0 (the "Dynamical System with a Strange Attractor" problem group above). The use of different planes for raying and for reconstruction can be imposed by equipment restrictions (Robinson and Roytvarf 2001).

P7.4[**]

Describe minimal polynomials of the roots of an odd degree n of E in a ring of 2×2 matrices with real entries. *Derive* from the preceding task the following characterization of the set of these roots: geometrically, it is a disjoint (or disconnected) union of one single point and $(n-1)/2$ hyperboloids of two sheets $\left(\text{in the space } \mathbf{R}^4_{\alpha,\beta,\gamma,\delta} = \left\{ \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \right\}\right)$; this single point corresponds to $R_{\theta_0^n} = E$, and $n-1$ vertices of those sheets correspond to $R_{\theta_k^n}$, $k = 1,\ldots, n-1$;[2] similitude transformations $A \mapsto C^{-1} \circ A \circ C$, with the orientation-preserving linear operators C, transpose elements inside the hyperboloids' sheets, whereas transformations with the orientation-reversing C transpose the sheets of each hyperboloid.[3]

P7.5[**]

Make for even degree n descriptions analogous to those given in section P7.4[**] for odd n.

P7.6[***]

This and the next two problems are addressed to readers familiar with normed vector spaces. A Euclidean or Hermitian space is normed since the triangle inequality $|x+y| \le |x| + |y|$ follows from the Cauchy-Schwarz-Bunyakovskii inequality $|\langle x,y\rangle|^2 \le \langle x,x\rangle \cdot \langle y,y\rangle$. (These inequalities are discussed in more detail in the "Convexity and Related Classic Inequalities" problem group below). The **norm of linear operator** A in a normed vector space is a nonnegative number defined, with respect to the norm $|\cdot|$ in the source space, as

$$\|A\| := \sup_{x \neq 0} |Ax|/|x| = \sup_{|x|=1} |Ax|$$

[2] The property of being a rotation and the locations of sheet vertices are related to a scalar product. An observer connected to a different scalar product might refer to different roots of E as rotations and, in just the same way, different points of the sheets as their vertices.

[3] Readers preferring "scientific" terminology may reformulate this claim as follows: the group $GL(2,\mathbb{R})$ acts by similitude transformations on the space of linear operators on $\mathbb{R}^2$; the roots of degree n of E form an invariant subset; for odd n, the hyperboloids coincide with all regular orbits (one to one), and a single point forms a singular orbit. $GL(2,\mathbb{R})$ consists of two connected components of orientation-preserving and orientation-reversing operators (problem P6.8[***]); thus, the similitude transformations realized by the elements of the first (second) component preserve (as a whole) (resp. transpose) the sheets.

(which is equivalent to the equation $\|A\| = \inf\{C>0 : \ |Ax| \le C|x|, \forall x\}$; *verify* this equivalence). *Verify* that, in fact, this definition makes the vector space of linear operators (matrices) normed (with the norm $\|\cdot\|$). In particular, the operator's norm is continuous: $\|A_n\| \to \|A\|$ as $A_n \to A$. (*Prove* this continuity using the finite dimensionality of the space of operators). Next, since $|Ax| \le \|A\|\cdot|x|$, $\forall x$, we have $\| A \circ B \| \le \|A\|\cdot\|B\|$ for any two linear operators A, B. (*Prove* this last inequality. Also see section H6.11 in the "Polar and Singular Value Decomposition Theorems" problem group above). From this continuity, roots of E cannot have norms less than 1 (using the operator's norm defined with respect to any norm in the source vector space; *why*?). *Deduce* from these that linear operators that are orthogonal with respect to some scalar product cannot have norms less than 1 (using the operator's norm defined with respect to any norm in the source vector space). However, a root of E (orientation-preserving or -reversing) can have its norm arbitrarily large; *give* examples to show this is true.

*P7.7****

According to the previous problem, a linear operator's properties of having a norm equal to 1 and having it greater than 1 are not similitude-invariant. Also, the property of having a norm less than 1 is not similitude-invariant (*give* examples). If $\|A\| < 1$ (using the operator's norm defined with respect to any norm in the source vector space), $\lim\limits_{n\to\infty} A^n = 0$ in the operator's space. (*Prove*.) Therefore, sequences of iterations $A^n x$ starting from any x tend to zero (and vice versa). And obviously, $\lim\limits_{n\to\infty} A^n = 0 \Rightarrow \lim\limits_{n\to\infty} B^n = 0$ for any operator B similar to A, regardless of B's norm. Express the property "$\lim\limits_{n\to\infty} A^n = 0$" in terms of A's eigenvalues.[4] Also, *show* that this property is equivalent to $\|B\| < 1$ for some linear operator B similar to A.

*P7.8****

Prove **Gelfand's formula** for a linear operator on a finite-dimensional normed vector space: $\lim\limits_{n\to\infty} \sqrt[n]{\|A^n\|} = \max\limits_{\lambda \in Spec(A)} |\lambda|$. [5]

[4] The eigenvalues are similitude-invariant; moreover, the numbers of eigenvalues located inside and outside the unit disk in $\mathbb{C}$, and the eigenvalues themselves that lie on the unit circle, are topological invariants (Arnol'd 1975, 1978).

[5] The inequality $\lim\limits_{n\to\infty} \sqrt[n]{\|A^n\|} \le \|A\|$ becomes an equality for a normal operator on a Euclidean space (since in this case $\|A\| = \max\limits_{\lambda\in Spec(A)} |\lambda|$) but may be strict for a nonnormal one (see section P6.11**, the "Polar and Singular Value Decomposition Theorems" problem group above). For an infinite-dimensional version of Gelfand's formula see Belitskii and Lubich (1984) and references therein.

P7.9[**]

Find the eigenvalues of the $n \times n$ matrix $\begin{pmatrix} 0 & 1 & & & \\ 1 & 0 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 0 & 1 \\ & & & 1 & 0 \end{pmatrix}$. *Derive* from it the inequality $x_1x_2 + \ldots + x_{n-1}x_n \le \cos\frac{\pi}{n+1} \cdot \sum_{i=1}^{n} x_i^2$ $(\forall x_1,\ldots,x_n \in \mathbb{R})$ (which is exact, in the sense that $\cos\frac{\pi}{n+1}$ cannot be replaced by a smaller factor that would depend only on n).

Establish the identities $\prod_{k=1}^{n} \cos\frac{k\pi}{2n+1} = 2^{-n}$ and $\prod_{k=1}^{n} \cos\frac{k\pi}{2n+2} = 2^{-n}\sqrt{n+1}$, $\forall n = 1, 2, \ldots$.

P7.10[**]

Prove that all eigenvalues of the $n \times n$ matrix $\begin{pmatrix} 0 & 1 & & & -1 \\ 1 & 0 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 0 & 1 \\ -1 & & & 1 & 0 \end{pmatrix}$ (entries not filled with numbers or dots are zeros) belong to the set $\left\{2\cos\frac{\pi}{n}, \ldots, 2\cos\frac{\pi(n-1)}{n}, -2\right\}$. *Derive* from it the inequality $x_1x_2 + \ldots + x_{n-1}x_n - x_nx_1 \le \cos\frac{\pi}{n} \cdot \sum_{i=1}^{n} x_i^2$ $(\forall x_1,\ldots,x_n \in \mathbb{R})$. *Show* that this matrix really has an eigenvalue of -2 if and only if n is odd; this eigenvalue is simple, and the corresponding eigenspace is spanned on the vector $(1,-1,1,-1,\ldots,1)$.

One can prove that $2\cos\frac{\pi}{n}$ really is the eigenvalue (of multiplicity 2, with a two-dimensional eigenspace) with the help of the well-known **Frobenius-Perron theorem** about matrices with nonnegative entries. (This theorem states that if $\lambda > 0$ is the maximal absolute value of an eigenvalue, then λ is an eigenvalue and, in addition, possesses a nonnegative eigenvector.[6] Apply this theorem to a matrix from Sect. P7.9** substituting $n-1$ for n.) Thus, the preceding

[6] Very often the Frobenius-Perron theorem is formulated with an extra condition producing such implications that λ is a simple (nonmultiple) eigenvalue, all elements of the corresponding eigenvector are positive, $-\lambda$ is not an eigenvalue, and so on. (The matrix from section P7.9** does not satisfy that extra condition since $-\lambda$ is an eigenvalue together with λ, but this matrix has n distinct, simple eigenvalues). The Frobenius-Perron theorem and its generalizations are connected with Markov chains, the theory of groups, and the theory of graphs. These theorems are widely applied to stochastic processes, differential equations, functional analysis, convex analysis, theory of games, mathematical economics, and other areas. Readers will find further discussions of the Frobenius-Perron theorem with generalizations and related subjects in Gantmacher (1967), Bellman (1960), Beckenbach and Bellman (1961), Horn and Johnson (1986), Belitskii and Lubich (1984), and multiple references therein.

inequality is exact. We leave it to interested readers to determine the remaining eigenvalues, their multiplicities, and all eigenspaces. (This may require some advanced technical elements as discussed in section E12.30, the "Least-Squares and Chebyshev Systems" problem group below).

Hint

H7.1

The orientation-preserving and orientation-reversing elements of the orthogonal (for a fixed scalar product) group in $\mathbb{R}^2$ are, respectively, R_θ and $Q_\theta := \begin{pmatrix} \sin\theta & \cos\theta \\ \cos\theta & -\sin\theta \end{pmatrix}$. All Q_θ are square roots of E and, hence, roots of any even degree of E. (They are not roots of any odd degree of E; why?) At the same time, for any n, all R_θ that are roots of degree n of E are $R_{\theta_k^n}$; prove it.

H7.2

$(C^{-1} \circ A \circ C)^{\,n} = C^{-1} \circ A^{\,n} \circ C = E$ when $A^n = E$. To prove the inversion, show, for the root of E, that root subspaces in the complex case and the complexified root subspaces in the real case have (complex) dimension 1. Therefore, the Jordan canonical form of the root of E will be unitary (resp. orthogonal) with respect to the coordinate scalar product and to the Jordan basis e_1,... (that is, $\langle e_i, e_j \rangle = \delta_{ij}$).

H7.3

The orthogonality of a real-valued 2×2 matrix, of nonreal eigenvalues of modulus 1, with respect to some scalar product is proved by arguments similar to those in section H7.2. To prove the second statement, use a simultaneous diagonalization of two quadratic forms and the fact that a linear operator having matrix K in some basis is orthogonal with respect to a bilinear form having matrix M in the same basis if and only if ${}^t\mathrm{K} \circ \mathrm{M} \circ \mathrm{K} = \mathrm{M}$;.

H7.4 and H7.5

The minimal polynomial of a root of degree n of E in a ring of 2×2 real-valued matrices should divide $x^n - 1$ and have real coefficients and degree of at most

2 (because it divides the characteristic polynomial). Therefore, which could it be? (***Warning***: cases of odd and even n differ slightly).

H7.6

(A). The continuity of an operator's norms can be proved as follows. The space of the linear operators on a finite-dimensional space itself is finite-dimensional; thus, the norm is a convex function in a whole finite-dimensional space; therefore, it is continuous (which is discussed in the "Convexity and Related Classic Inequalities" problem group below).[7]

[7] For a norm $v = |\cdot|$, the balls $B_{v,k} := \{v \le k\}$ are homothetic convex sets, symmetric relative to the origin and containing αe with a coefficient $\alpha > 0$ for any vector e. (In a finite-dimensional space, by virtue of the norms' continuity, they are closed sets. Why?) Conversely, let a set H possess the aforementioned properties of convexity, symmetry, etc. For any basis e_1, e_2,... H contains an "octahedron" K that is the convex hull of a set $\{\beta e_1, -\beta e_1, \beta e_2, -\beta e_2, \ldots\}$ for some $\beta > 0$. (Why?) In turn, K contains an open neighborhood of the origin. Therefore, H contains αH $(0 \le \alpha < 1)$ in its interior, $\alpha H \subseteq \text{int } H$. [Indeed, $\alpha H + (1 - \alpha)H \subseteq \text{int } H$, while $(1 - \alpha)H$ contains a neighborhood of the origin. How does this help?] Also, the closure $\overline{H}$ coincides with the intersection $\bigcap_{\alpha>1} \alpha H$. (If $x \in \overline{H}$, $x + \varepsilon H$ intersects with H for all $\varepsilon > 0$, so that $x \in H + \varepsilon H$, then $\subseteq \alpha H$ for $\alpha > 1$ when $\varepsilon \le 1 - \alpha$. If $x \notin \overline{H}$, a spherical neighborhood of x, does not intersect with H, then $x \notin \alpha\overline{H}$ for some $\alpha > 1$; provide a figure!) Lastly, a function defined as

$$v = v_H(x) := \inf\{k>0 : x \in kH\}$$

is a norm and $B_{v,k} = k\overline{H}$. (The inclusion $k_1\overline{H} + k_2\overline{H} \subseteq (k_1 + k_2)\overline{H}$ provides the triangle inequality; why?) The norms' continuity implies a topological **equivalence of** any **norm** v_1, v_2 in a finite-dimensional space, which means that $\sup_{v_2(x)=1} v_1(x)<\infty$, $\sup_{v_1(x)=1} v_2(x)<\infty$. (Another formulation is $C_1 v_1(x) \le v_2(x) \le C_2 v_1(x), \forall x\ (0 < C_1 \le C_2 < \infty)$. In particular, any two Euclidean norms in a finite-dimensional space are bound by such an inequality (note: this can be proved using much simpler means; how?). In fact, for any norm v_2 there exists a Euclidean norm v_1 that satisfies this inequality with $C_2/C_1 \le \sqrt{n}$, where n is the spatial dimension [Jordan's theorem (Schmidt 1980; Hörmander 1983)]. Geometrically, the equivalence means that $\alpha B_{v,1} \subseteq B_{v,1} \subseteq \beta B_{v,1}$ for some $0 < \alpha \le \beta < \infty$. (In particular, the balls of all norms in a finite-dimensional space are bounded sets; why?) Indeed, define for a norm v a polyhedron K contained in $B_{v,1}$, as above. Since v is a continuous function, it has a minimum on the boundary ∂K (as this is a compact set). This minimum is positive. (Why?) Therefore, $K \subseteq B_{v,1} \subseteq \alpha K$ for some $\alpha \ge 1$ (why?), so that norms v and v_K are equivalent. Since the norms' equivalence is obviously a kind of equivalence relation, this implies the equivalence of all norms. **QED**. Lastly, a function $v_0(x) := \max_i |x_i|$, where x_i are coordinates with respect to some fixed basis, is a norm. Therefore, from the norms' equivalence we obtain the following statement: a sequence $\{x^{(n)}\}$ in a finite-dimensional space tends to zero when $x^{(n)} \to 0$ for some norm. (From the norms' continuity, $x^{(n)} \to 0 \Rightarrow |x^{(n)}| \to 0$ for any norm).

(B). From the inequality for the norm of a composition we have, for $R = E^{1/n}$,

$$1 = ||E|| = ||R^n|| \quad \leq \quad ||R||^n,$$

so that $||R|| \geq 1$.

(C). Show that the roots of E form an everywhere dense subset of the set of orthogonal linear operators on a Euclidean plane $\mathbb{R}^2$. Generalize this result proving a similar statement for a Euclidean space $\mathbb{R}^n$; this can be achieved using the orthogonal spectral theorem over the real numbers (see footnote in section E6.8).

From (A) – (C), norms of orthogonal linear operators on a Euclidean space cannot be less than 1. (Why?) At the same time, the distant points of the hyperboloids from sections P7.4** and P7.5** correspond to roots of E with large norms. Here are some of those roots of norms greater than or equal to N: $R^n_{N,k} = \begin{pmatrix} 0 & -N^{-1} \\ N & 2\cos\theta^n_k \end{pmatrix}$, $Q_N = \begin{pmatrix} 0 & N^{-1} \\ N & 0 \end{pmatrix}$. (The norms are not less than N as $\left|R^n_{N,k}x\right| = N$, $|Q_N x| = N$ for some $x \in \mathbb{R}^2$ of $|x| = 1$, which ones? Also, readers can find the norms' exact values).

H7.7

An orthogonal root of E, R in the ring of linear operators on the Euclidean plane $\mathbb{R}^2$ has a (Euclidean) norm value of 1. However, according to section P7.6***, a similar operator $C^{-1} \circ R \circ C$ can have an arbitrary norm $N > 1$. Hence, $||N^{-1}R|| < 1$, whereas $||C^{-1} \circ (N^{-1}R) \circ C|| = 1$. Also, $\left\| \frac{N-\varepsilon}{N} R \right\| < 1$, $\left\| C^{-1} \circ \left(\frac{N-\varepsilon}{N} R \right) \circ C \right\| > 1$ for $0 < \varepsilon < N - 1$. Next, the property $\lim_{n\to\infty} A^n = 0$ is equivalent to the following statement: all eigenvalues of A are located inside an (open) unit disk in X. Prove the foregoing statement and the last claim of section P7.7*** using Jordan's canonical form.

H7.8

For eigenvector x in the complexification of the vector space corresponding to eigenvalue λ, $|A^n x| = |\lambda|^n \cdot |x|$, so $||A^n|| \geq |\lambda|^n$ and $||A^n||^{1/n} \geq |\lambda|$. Establish the inverse inequality using Jordan's canonical form.

H7.9

The answer is: $2\cos\frac{\pi k}{n+1}$: $k = 1, \ldots, n$. To calculate the characteristic polynomial, apply the usual reduction to a triangular form:

$$\chi(\lambda) = \det \begin{pmatrix} \lambda & -1 & & & \\ -1 & \lambda & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \lambda & -1 \\ & & & -1 & \lambda \end{pmatrix} = \det \begin{pmatrix} \lambda & -1 & & & \\ & \lambda - \frac{1}{\lambda} & -1 & & \\ & & \lambda - \frac{1}{\lambda - \frac{1}{\lambda}} & -1 & \\ & & & \ddots & \ddots \end{pmatrix}.$$

When the characteristic equation $\chi(\lambda) = 0$ is solved using an approach from the theory of **continued fractions,** 2×2 matrices with eigenvalues $\cos\frac{\pi k}{n+1} \pm i \sin\frac{\pi k}{n+1}$ arise. (Does this tell the reader something?)

Next, the inequality holds because it is equivalent to the statement that the maximal eigenvalue of the quadratic form on its left-hand side is equal $\cos\frac{\pi}{n+1}$.

The first trigonometric identity from section P7.9** is obtained with the help of the relation "*determinant = the product of all eigenvalues,*" as applied to the matrix from section P7.9** for even n. The second identity can be obtained by a similar application to another three-diagonal matrix. (Which one?)

H7.10

Denoting the matrix from section P7.10** as B, we have

$$B = \begin{pmatrix} 0 & 1 & \cdots & -1 \\ 1 & \lceil & & \rceil \\ & & A & \\ -1 & \lfloor & & \rfloor \end{pmatrix} = \begin{pmatrix} \lceil & & \rceil & -1 \\ & A & & \\ \lfloor & & \rfloor & 1 \\ -1 & & 1 & 0 \end{pmatrix},$$

where A is the $(n-1) \times (n-1)$ matrix from the previous problem section P7.9**. A direct computation brings, for eigenvalue λ of B and the corresponding eigenvector ${}^t(x_1, \ldots, x_n)$,

$$(\lambda - A)\begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ 0 \\ -x_1 \end{pmatrix}, \qquad (\lambda - A)\begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \end{pmatrix} = \begin{pmatrix} -x_n \\ 0 \\ x_n \end{pmatrix},$$

$$x_2 - x_n = \lambda x_1, \qquad -x_1 + x_{n-1} = \lambda x_n.$$

Matrix A is symmetric with respect to the reflection in a secondary diagonal and, hence, is permutable with a linear transformation T: $\mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ transposing the coordinates, which are equally spaced with respect to the ends. Therefore, T preserves eigenspaces of A, and in addition, applying a linear operator $E + T$ to the pair of the preceding matrix equations, we obtain that either λ is an eigenvalue of A or $\begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix} = -\begin{pmatrix} x_n \\ \vdots \\ x_2 \end{pmatrix}$ and $\begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \end{pmatrix} = -\begin{pmatrix} x_{n-1} \\ \vdots \\ x_1 \end{pmatrix}$. (Fill in the details). In the first case apply section P7.9** to make a decision concerning λ. In the second case, an inductive computation of the relationship among x_i using the pair of preceding scalar equations yields all possible values $\lambda = \pm 2$. Eigenvalue -2 will exist and eigenvalue 2 will not exist. (Furnish the details).

Readers familiar with the **Frobenius-Perron theorem** may show that $2\cos\frac{\pi}{n}$ really is an eigenvalue of B (of multiplicity 2, with the root subspace that actually is an eigenspace since B is asymmetric matrix). From this theorem we conclude that, the eigenvector ${}^t(x_2,\ldots,x_n)$ corresponding to an eigenvalue $\lambda = 2\cos\frac{\pi}{n}$ is nonnegative and the equality $T\begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix} = -\begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix}$ is impossible because this eigenvalue has the maximal absolute value of all eigenvalues of A and all eigenvalues of A are distinct and, hence, simple. Consequently, $T\begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix}$ as T preserves both eigenspaces of A and Euclidean norms and $x_2 = x_n$. Now a direct verification shows that $e_1 = {}^t(0, x_2,\ldots, x_n)$ is an eigenvector of B corresponding to the eigenvalue $\lambda = 2\cos\frac{\pi}{n}$ Similarly, B has an eigenvector $e_2 = {}^t(x_1,\ldots, x_{n-1},0)$ (with the same $x_2,\ldots,x_{n-1}$) corresponding to this eigenvalue. We have $x_1 \neq 0$ and $x_n \neq 0$ (why?), so that vectors e_1, e_2 are linearly independent. An eigenvector $e = {}^t(x'_1, x'_2, \ldots, x'_{n-1}, x'_n)$ of B with $x'_1 = 0$ is proportional to e_1, and if $x'_1 \neq 0$, then $e - (x'_1/x_1)e_2$ is proportional to e_1 (why?), so that the eigenspace corresponding to $\lambda = 2\cos\frac{\pi}{n}$ is spanned on e_1, e_2.

A proper modification of the preceding arguments using the description of eigenspaces of the matrix from section P7.9** (see section E12.30 in the "Least Squares and Chebyshev Systems" problem group) allows us to establish that the eigenvalues of B are $\lambda_k = 2\cos\varphi_k$, $k = 1,\ldots, \left[\frac{n+1}{2}\right]$ ([] denotes the integral part of a number), where $\varphi_k = (2k - 1)\pi/n$, which have eigenvectors ${}^t\left(\sin\frac{\varphi_k}{2}, \sin\frac{3\varphi_k}{2}, \ldots, \sin\left(\left(n - \frac{1}{2}\right)\varphi_k\right)\right), {}^t\left(\cos\frac{\varphi_k}{2}, \cos\frac{3\varphi_k}{2}, \ldots, \cos\left(\left(n - \frac{1}{2}\right)\varphi_k\right)\right)$ Hence, the eigenvalues are at least double if $k \leq [n/2]$; so these are double, whereas the eigenvalue -2 corresponding to

$k = \frac{n+1}{2}$ (which exists if and only if n is odd) is simple (nonmultiple). [Also, we can see once again that the eigenspace corresponding to -2 is spanned on $(1,-1,1,-1,\ldots,1)$.] Interested readers may fill in the details.

Eigenvalues and eigenvectors of multiple three-diagonal and related matrices can be found proceeding by the foregoing method. For example, readers can establish that the $n \times n$ matrix T

$$= \begin{pmatrix} 0 & 1 & & & 1 \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ 1 & & & 1 & 0 \end{pmatrix}$$ (entries not filled with numbers or dots are zeros) has the eigenvalues $\tau_k = 2\cos(2\pi k/n)$, $k = 0,\ldots,[n/2]$, and

- $\tau_0 = 2$ is simple and its eigenspace is spanned on ${}^t(1,\ldots,1)$;
- For $k = 1,\ldots,[(n-1)/2]$, τ_k is double and its eigenspace is spanned on ${}^t(\sin(\pi k/n),\ldots,\sin((2n-1)\pi k/n))$ and ${}^t(\cos(\pi k/n),\ldots,\cos((2n-1)\pi k/n))$;
- $\tau_{n/2}$ (existing for even n) is simple and its eigenspace is spanned on ${}^t(1,-1,\ldots,1,-1)$.

This description makes it possible to formulate inequalities similar to those in sections P7.9** and P7.10**. The first of them is quite simple, $x_1x_2 + \ldots + x_nx_1 \le \sum {x_i}^2$ (it can immediately be deduced from the Cauchy-Schwarz-Bunyakovskii inequality), but the second one much more interesting:

$$x_1x_2 + \ldots + x_nx_1 \quad \le \quad \cos(2\pi/n) \cdot \sum {x_i}^2 \qquad \text{when} \qquad x_1 + \ldots + x_n = 0;$$

it follows from the fact that the quadratic form $x_1x_2 + \ldots + x_nx_1$, as restricted to the orthogonal complement of ${}^t(1,\ldots,1)$, has a maximal eigenvalue of $\cos(2\pi/n)$ with respect to the quadratic form $\sum {x_i}^2$. Readers should be able to write out the third, fourth, ... inequalities from this list by themselves.

The preceding description also makes it possible to compare two difference linear operators on $\mathbb{R}^n$, defined by the matrices $\Delta^{(1)} = h^{-2} \cdot \begin{pmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & & \cdots & & \\ & & 1 & -2 & 1 \\ 1 & & & 1 & -2 \end{pmatrix}$ and $\Delta^{(2)} = h^{-2} \cdot \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & & \cdots & & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}$.

These operators are uniform grid digitalizations of the one-dimensional Laplace operators $\Delta = d^2/dt^2$ in spaces of functions of $t \in [0,l]$ using the **digitalization order** n and the **spacing** $h = l/n$. We have $\Delta^{(1)} = h^{-2}(T - 2E)$ and $\Delta^{(2)} = h^{-2}(A - 2E)$, where A is the matrix from section P7.9**. The spectral properties of these operators prove to be quite different! To be more specific:

- $\Delta^{(2)}$ is negative definite; $\Delta^{(1)}$ is negative semidefinite, with the kernel spanned on ${}^t(1,\ldots,1)$;
- Arranging nonzero eigenvalues by their magnitudes $(-{\lambda_1}^{(1)} < -{\lambda_2}^{(1)} < \ldots, -{\lambda_1}^{(2)} < -{\lambda_2}^{(2)} < \ldots)$, $-{\lambda_k}^{(1)}/(2\pi k/l)^2 \to 1$, and $-{\lambda_k}^{(2)}/(\pi k/l)^2 \to 1$ as $k/n \to 0$;
- All eigenvalues ${\lambda_k}^{(2)}$ are simple; ${\lambda_0}^{(1)}$ is simple; ${\lambda_k}^{(1)}$ with $k > 0$ are double (except for the eigenvalue of the maximal magnitude, ${\lambda_{n/2}}^{(1)}$, which exists for even n).[8]

[8] Similar claims were first proved by J.L. Lagrange in the 1750s.

That is, $\lambda_k^{(1)} = -4\ h^{-2}\sin^2(\pi k/n)$ and have the same eigenspaces as τ_k, and $\lambda_k^{(2)} = -4\ h^{-2}\sin^2(\pi k/2(n+1))$ and have the same eigenspaces as the eigenvalues of matrix A, which are spanned on vectors ${}^t(\sin(\pi k/(n+1)), \ldots, \sin(\pi nk/(n+1)))$, respectively. (We leave it to the interested reader to furnish all details).

Readers experienced in the theory of ordinary linear differential equations know that the **Sturm-Liouville boundary eigenvalue problems** using operators Δ defined above have different solutions for the spaces of functions on $t\in[0,l]$ in the following two cases: (1) having arbitrary equal values at both endpoints (**periodic boundary condition**) and (2) having zero values at both endpoints (**homogeneous boundary condition of the first kind**). In the first case, Δ has the eigenvalues $-(2\pi k/l)^2$ ($k = 0,1,\ldots$), simple for $k = 0$ and double for $k > 0$ [the eigenspaces are spanned on $\sin(2\pi kt/l)$ and $\cos(2\pi kt/l)$]; in the second case it has the eigenvalues $-(\pi k/l)^2$ ($k = 1,2,\ldots$), all of which are simple [the eigenspaces are spanned on $\sin(\pi kt/l)$].[9] Thus, difference operators $\Delta^{(1)}$ and $\Delta^{(2)}$ correspond to digitalizations of Δ under different boundary conditions – respectively, the periodic condition and the homogeneous condition of the first kind.

$\Delta^{(1)}$ is a special case of a symmetric matrix of the form $\begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \cdots & \alpha_2 & \alpha_1 \\ \alpha_1 & \alpha_0 & \alpha_1 & \alpha_2 & \cdots & \alpha_2 \\ \alpha_2 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \alpha_2 \\ \alpha_2 & \cdots & \alpha_2 & \alpha_1 & \alpha_0 & \alpha_1 \\ \alpha_1 & \alpha_2 & \ldots & \alpha_2 & \alpha_1 & \alpha_0 \end{pmatrix}$ [the first row looks either like $(\alpha_0,\alpha_1,\ldots, \alpha_{(n-1)/2}, \alpha_{(n-1)/2},\ldots, \alpha_1)$ or $(\alpha_0,\alpha_1,\ldots, \alpha_{(n-2)/2}, 0, \alpha_{(n-2)/2},\ldots, \alpha_1)$ – respectively, for odd and even n, and the remaining rows are obtained by circular shifts of the first row]. These matrices can be expressed by polynomials $c_0E + \ldots + c_qT^q$ in T of degrees smaller than or equal to $q = [(n-1)/2]$ with numerical coefficients c_i; each coefficient can be determined from an upper-triangular system of $q+1$ linear equations with, respectively, $\alpha_0,\ldots, \alpha_q$ on their right-hand sides. (Readers can explicitly determine the left-hand side of the system by induction on n using Pascal-like triangles; the cases of even and odd n should be treated separately). Such matrices having $\alpha_i \geq 0$ and $\alpha_0 + 2\sum_{i>0}\alpha_i = 1$ are used to digitize symmetric **blur** operators. Blur is inevitably caused by several factors that are inherent in every physical system. For example, in X-ray (tomographic) technology the main causes of blur stem from the scintillator, the optics and the CCD, X-ray dispersions, and the finite size of the focal spot; there are some secondary causes of blur as well. Blurring makes reconstruction difficult because it rounds edges and corners and practically removes small geometrical details such as toothed surfaces: mathematically, the value of $2\sum_{i>0}\alpha_i$ may be comparable with α_0. Accurate reconstruction requires deblurring the data, i.e., removing the impact of blurring. Deblurring (deconvolution) consists of inversion of the blur operator. As one is able to prove, in general (when $2\sum_{i>0}\alpha_i$ is comparable with α_0), the minimal modulus of the digital blur operator's eigenvalue on the orthogonal complement to its kernel tends to zero as a digitalization order $n \to \infty$. Because of that, unlimited increasing of this order raises noncontrollable artifacts, which makes common inversion algorithms entirely impractical for deblurring. This problem and its solution using the **Tikhonov-Phillips regularization** will be discussed in the "Method of Steepest Descent and Deblurring" problem group in volume 2 of this book.

[9] In fact, all differential operators of the second order that are symmetric with respect to a properly weighted integral scalar product in the space of functions have similar spectral properties.

Explanation

E7.1

Q_θ cannot be the roots of an odd degree of E since $\det Q_\theta^n = (-1)^n$. Next, the matrices of the form $\begin{pmatrix} a & -b \\ b & a \end{pmatrix} = \sqrt{a^2+b^2} \cdot R_{\operatorname{Arc}\tan(b/a)}$: $a, b \in \mathbf{R}$ form a field (also, the map $\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \mapsto a + bi$ gives an isomorphism of this field onto the field of complex numbers) – see section H6.9 (the "Polar and Singular Value Decomposition Theorems" problem group above). A polynomial equation of degree n has at most n roots in a field; thus, there are at most n roots of degree n of E that have the form $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$.

E7.2

The (complexified) root subspaces of a root of E cannot have a (complex) dimension greater than 1 for topological reasons. Indeed, for a Jordan box of size $k>1 J = \begin{pmatrix} \lambda & 1 & & \\ & \lambda & 1 & \\ & & \ddots & \ddots \end{pmatrix}$ we will have $J^m = \begin{pmatrix} \lambda^m & m\lambda^{m-1} & \binom{m}{2}\lambda^{m-2} & \cdots \\ & \lambda^m & m\lambda^{m-1} & \binom{m}{2}\lambda^{m-2} & \cdots \\ & & \ddots & \ddots & \ddots \end{pmatrix}$, so that for some vector x (for example, which?) $J^m x$ grows as m^{k-1}, whereas $J^{qn}x = x$, $\forall q \in \mathbb{Z}$. **QED**.

E7.3

Let a linear operator be orthogonal with respect to two symmetric bilinear forms, one of which is positive definite (a scalar product). By the theorem about simultaneous diagonalization of two quadratic forms (in other words, the spectral theorem of symmetric linear operators), there is an orthogonal basis for both of them and one of them has the identity matrix, , with respect to this basis. Matrices of the operator

and second form in that basis, K and Δ respectively (Δ is diagonal), satisfy the equations

$$^t\mathrm{K} \circ \mathrm{K} = \mathrm{E}, \quad ^t\mathrm{K} \circ \Delta \circ \mathrm{K} = \Delta.$$

From the first equation $^t\mathrm{K} = \mathrm{K}^{-1}$, so that the second means that K commutes with Δ. K is a rotation (otherwise it would have the eigenvalues ±1; see section H7.1 above); therefore, the proof can be completed by reference to the fact that the rotation and diagonal matrices in $\mathbb{R}^2$ commute if and only if the first is ± E or the second is proportional to E (section E6.9, the "Polar and Singular Value Decomposition Theorems" problem group above).

E7.4 and E7.5

The minimal polynomials of roots of odd degrees n of E in a ring of 2 × 2 real-valued matrices are either $\mu(x) = x - 1$ or $\mu(x) = x^2 - 2\cos\theta_k^n x + 1$, $k = 1,\ldots, n-1$. For any even n we will have all of the above, as well as two extra possibilities: $\mu(x) = x + 1$, $\mu(x) = x^2 - 1$. For $\mu(x) = x \pm 1$ the corresponding roots are $\pm E$. If deg $\mu = 2$, then μ will coincide with the characteristic polynomial; therefore, in the terms of matrix elements

$$\begin{aligned}
&\mu(x) = x^2 - 2\cos\theta_k^n x + 1 \quad \Leftrightarrow \quad \alpha + \delta = 2\cos\theta_k^n, \quad \left(\alpha - \cos\theta_k^n\right)^2 + \beta\gamma = -\sin^2\theta_k^n;\\
&\mu(x) = x^2 - 1 \qquad\qquad \Leftrightarrow \quad \alpha + \delta = 0, \qquad\quad \alpha^2 + \beta\gamma = 1\,..
\end{aligned}$$

The first group of these equations corresponds to a hyperboloid of two sheets H_k^2 in a hyperplane $\left\{\alpha + \delta = 2\cos\theta_k^n\right\}$ of $\mathbf{R}^4_{\alpha,\beta,\gamma,\delta}$, and the second group corresponds to a hyperboloid of one sheet H^1 in a hyperplane $\{\alpha + \delta = 0\}$ of the same space. Verify that for the vertices of the sheets of H_k^2 we have $\alpha = \delta = \cos\theta_k^n$, $\beta = -\gamma \pm \sin\theta_k^n$, which corresponds to the matrices $R_{\theta_k^n}$, and on the neck (striction) circle of H^1 we have $\alpha = -\delta$, $\beta = \gamma$, $\alpha^2 + \beta^2 = 1$, which corresponds to the matrices Q_θ.[10] The points of H_k^2 and H^1 correspond to operators similar to $R_{\theta_k^n}$ and Q_θ (according to

[10] The correspondence of a point of a hyperboloid to an orthogonal operator and the location of the neck (striction) circle on this hyperboloid are connected to a scalar product. An observer using a different scalar product might refer to different linear operators as orthogonal and, similarly, to a different neck ellipse on the hyperboloid as the circle.

section P7.2**).[11] Next, orientation-preserving and –reversing components of $GL(n, \mathbb{R})$ give connected similitude orbits (according to section P6.8***); therefore, the sheets of H_k^2 should coincide with those orbits. (Why?) Lastly, H^1 is covered twice by the similitude action of $GL(n,\mathbb{R})$, just as it is covered by the orbit of each component, which is clear from the multiplications

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \circ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \circ \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} \sin 2\theta & \cos 2\theta \\ \cos 2\theta & -\sin 2\theta \end{pmatrix}$$
$$= \begin{pmatrix} \sin\theta & \cos\theta \\ \cos\theta & -\sin\theta \end{pmatrix} \circ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \circ \begin{pmatrix} \sin\theta & \cos\theta \\ \cos\theta & -\sin\theta \end{pmatrix}.$$

[Complete the proof using the polar decomposition (see the problem group "Polar and Singular Value Decomposition Theorems")].

E7.7

For an eigenvector x in the complexification of the vector space, corresponding to an eigenvalue λ, $|Ax| = |\lambda| \cdot |x|$, so $A^n x$ does not tend to zero if $|\lambda| \geq 1$. On the other hand, the Jordan box's powers $J^m = \begin{pmatrix} \lambda^m & m\lambda^{m-1} & \binom{m}{2}\lambda^{m-2} & \cdots \\ & \lambda^m & m\lambda^{m-1} & \binom{m}{2}\lambda^{m-2} & \cdots \\ & & \ddots & \ddots & \ddots \end{pmatrix}$ tend to zero for $|\lambda| < 1$. (Why?) Therefore, $A^n \to 0$ if all its eigenvalues are of moduli less than 1. Changing a Jordan basis in a root subspace $e_1, e_2, \ldots, e_k \mapsto e_1, \varepsilon e_2, \ldots, \varepsilon^{k-1} e_k$ substitutes the corresponding Jordan box as $\begin{pmatrix} \lambda & 1 & & \\ & \lambda & 1 & \\ & & \ddots & \ddots \end{pmatrix} \mapsto \begin{pmatrix} \lambda & \varepsilon & & \\ & \lambda & \varepsilon & \\ & & \ddots & \ddots \end{pmatrix}$. A norm of the last matrix (defined with respect to any fixed norm in the source vector space) equals $|\lambda| + O(\varepsilon)$; therefore, it is less than 1 for $|\lambda| < 1$ and small ε, which completes the solution of section P7.7***.

[11] Eigenvalues are similitude-invariant. Actually, eigenvalues lying on a unit circle in $\mathbb{C}$ are topologically invariant (Arnol'd 1975, 1978).

E7.8

A Jordan box of size k has the form $J = \lambda E + N$, where N is nilpotent of degree k, $N^k = 0$. By the binomial formula,

$$J^n = \lambda^n \cdot \left[E + \tfrac{nN}{\lambda} + \ldots + \binom{n}{k-1} \cdot \left(\tfrac{N}{\lambda}\right)^{k-1} \right] = \lambda^n \cdot O\left(n^{k-1}\right).$$

Because of this, for the norm $|x| := \max\limits_i |x_i|$, with the coordinates taken with respect to the Jordan basis, $|A^n x| \le \max\limits_{\lambda \in Spec(A)} |\lambda|^n \cdot O\left(n^{k-1}\right) \cdot |x|$. Hence, $\|A^n\| \le \max\limits_{\lambda \in Spec(A)} |\lambda|^n \cdot O\left(n^{k-1}\right)$, and similar inequalities, with appropriate functions $O(n^{k-1})$, will hold for all norms in the linear operators' space. (Why?) Extracting the root we come to the inequality $\sqrt[n]{\|A^n\|} \le \max\limits_{\lambda \in Spec(A)} |\lambda| \cdot (1 + o(1))$, which completes the solution of section P7.8***. (Find and fill the gap in our arguments).

E7.9

Verify that polynomials in λ with integral coefficients $p_k(\lambda) = \lambda^k + \ldots$, $q_k(\lambda)$, of greatest common divisor 1, which are the numerator and denominator of a fraction $\frac{p_k(\lambda)}{q_k(\lambda)} = \lambda - \frac{1}{\lambda - \frac{1}{\lambda - \ldots}}$ (of k stories on the right-hand side), can be taken as, respectively, the first and second components of a planar vector $\begin{pmatrix} p_k \\ q_k \end{pmatrix} = M^k \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, where a matrix is $M = \begin{pmatrix} \lambda & -1 \\ 1 & 0 \end{pmatrix}$. Next, it is easy to see that $p_n = \chi$. (Why?) Therefore, the characteristic equation is $p_n = 0$; in other words, the vector $M^n \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ lies on the ordinate axis. On the other hand, M maps this axis onto the abscissa axis, that is, the positive semiaxis of the ordinate onto the negative semiaxis of the abscissa. Derive from this that $M^{n+1} = \kappa E$. But since $\det M = 1$, we have $\kappa = \pm 1$; hence, M may have only eigenvalues $\mu_\pm = \cos\frac{\pi k}{n+1} \pm i \sin\frac{\pi k}{n+1}$, using one of $k = 1,\ldots, n$ ($k = 0, n+1$ are excluded as $M \ne \pm E$). Finally, M has $\chi_M(\mu) = \mu^2 - \lambda\mu + 1$ as its characteristic polynomial, so $\lambda = \mu_+ + \mu_-$, and these must be included for all $k = 1,\ldots, n$ since the matrix in section P7.9** cannot have multiple eigenvalues (verify), completing the proof.

Since $\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}=\begin{pmatrix}0\\1\end{pmatrix}$ and $\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix}=\begin{pmatrix}-1\\0\end{pmatrix}$, the determinant of the two-diagonal matrix from section P7.9** will be equal to zero for odd n and to $(-1)^{(n \bmod 4)/2}$ for even n, yielding the trigonometric identities $\prod_{k=1}^{n}\cos\frac{k\pi}{2n+1}=2^{-n}$, $\forall n=1,2,\ldots$. [Fill in all details; a different proof may be obtained using the Chebyshev polynomials, discussed in the "Least Squares and Chebyshev Systems" problem group below; also, see an elementary proof in the special case $n=7$ in Venkataraman (1971) or the issue of Trigg (1967) in Russian (1975): Тригг, Ч. Задачи с изюминкой. Под ред. В.М. Алексеева (Alexeev V.M., editor). "МИР" Press, Moscow.]

Next, we find that the eigenvalues of a three-diagonal $n\times n$ matrix $\Delta=\begin{pmatrix}-2 & 1 & & & \\ 1 & -2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & -2 & 1\\ & & & 1 & -2\end{pmatrix}$ are $\lambda_k=2\cos\frac{k\pi}{n+1}-2=-4\sin^2\frac{k\pi}{2n+2}$: $k=1,\ldots,n$. (This matrix results from the uniform grid digitalization of the operator d^2/dt^2). Also, an induction on n shows that $\begin{pmatrix}2 & -1\\ 1 & 0\end{pmatrix}^n\begin{pmatrix}1\\0\end{pmatrix}=\begin{pmatrix}n+1\\n\end{pmatrix}$ $(n=0,1,\ldots)$, so $\det\Delta=(-1)^n\cdot(n+1)$. Thus, identities $\prod_{k=1}^{n}\cos\frac{k\pi}{2n+2}=2^{-n}\sqrt{n+1}$, $\forall n=1,2,\ldots$, can be proved. (Furnish all details; a different proof may be obtained using the Chebyshev polynomials, discussed in the "Least Squares and Chebyshev Systems" problem group below).

Completing the Solution

S7.2

For $x=\begin{pmatrix}0\\ \vdots\\ 0\\ 1\end{pmatrix}$ we have $J^m x=\begin{pmatrix}\frac{m(m-1)\ldots(m-k+2)}{(k-1)!}\cdot\lambda^{m-k+1}\\ *\end{pmatrix}$, the magnitude of the first component of which grows as m^{k-1} [more precisely, as $m^{k-1}/(k-1)!$] for large m (as $|\lambda|=1$).

S7.3

The results of the problem sections P7.3** and P7.4** show that positive definite quadratic forms in $\mathbb{R}^2$ form an algebraic variety, which is a cylinder over a sheet of a hyperboloid of two sheets [a Cartesian product of this sheet and positive semiaxis $(0,\infty)$]. The same may be

verified directly because the entries of matrix $\begin{pmatrix} x & y \\ y & z \end{pmatrix}$ with a determinant equal to one are bound by the equation determining a hyperboloid of two sheets, $xz - y^2 = 1$, and, finally, the desired sheet is selected using the Sylvester criterion: $x,z > 0$. In turn, the problem section P7.4** can be solved using this result and the result of section P7.3**.

S7.4 and S7.5

Denote the similitude action of a linear operator T by A_T, $A_T X = T^{-1} \circ X \circ T$, and the set $\{A_T X: T \in S\}$ by $A_S X$. A proof of the theorem stating that the sheets of the hyperboloid H_k^2 coincide with the orientation-preserving and orientation-reversing similitude orbits $A_{GL^+(n,\mathbf{R})} R_{\theta_k^n}$ and $A_{GL^-(n,\mathbf{R})} R_{\theta_k^n}$, respectively, can be provided as follows. We have the representation of this hyperboloid as a union of connected subsets,

$$\mathrm{H}_k^2 = A_{GL(n,\mathbf{R})} R_{\theta_k^n} = A_{GL^+(n,\mathbf{R})} R_{\theta_k^n} \cup A_{GL^-(n,\mathbf{R})} R_{\theta_k^n},$$

and, on the other hand, its sheets are connected components; this proves the necessity because any connected subset is contained in some component. **QED**.

Perhaps readers will better understand this theorem when they consider the following statement:

- The commutativity identities $R_\theta \circ R_\varphi \circ R_\theta^{-1} = R_\varphi$, $Q_\theta \circ R_\varphi \circ Q_\theta^{-1} = {}^t R_\varphi = R_{-\varphi}$ (these show when a similitude transformation remains invariant or substitutes one sheet for another),
- The commutativity identity $\begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix} \circ \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \circ \begin{pmatrix} x^{-1} & 0 \\ 0 & y^{-1} \end{pmatrix} = \begin{pmatrix} a & -\frac{x}{y}b \\ \frac{y}{x}b & a \end{pmatrix}$,
- The arc $t \mapsto \begin{pmatrix} a & -\left(\frac{x}{y}\right)^t b \\ \left(\frac{y}{x}\right)^t b & a \end{pmatrix}$ $(t \in [0,1])$ connects the matrices $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ and $\begin{pmatrix} a & -\frac{x}{y}b \\ \frac{y}{x}b & a \end{pmatrix}$ $(x,y > 0)$,

and then we employ the polar decomposition. (The reader is encourage to complete the proof).

Turning to the hyperboloids H^1, the commutativity identities for $Q_0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ in section E7.4 and E7.5 show, taking into account that $H^1 = A_{GL(n,\mathbf{R})}Q_0$ (by virtue of section E7.2[12]) and using the polar decomposition, that $H^1 = A_{GL^+(n,\mathbf{R})}Q_0 = A_{GL^-(n,\mathbf{R})}Q_0$. If now $X \in H^1$ and $A_T X = Q_0$, then we have, depending on the sign of det T, $T \circ GL^{\pm}(n,\mathbb{R}) = GL^{\pm}(n,\mathbb{R})$ or $T \circ GL^{\pm}(n,\mathbb{R}) = GL^{\mp}(n,\mathbb{R})$,[13] which yields $A_{GL^{\pm}(n,\mathbf{R})}X = A_{GL^{\pm}(n,\mathbf{R})}Q_0$ or, respectively, $A_{GL^{\pm}(n,\mathbf{R})}X = A_{GL^{\mp}(n,\mathbf{R})}Q_0$, and, in any case, $H^1 = A_{GL^+(n,\mathbf{R})}X = A_{GL^-(n,\mathbf{R})}X$. **QED**.

S7.7

The binomial coefficients $\begin{pmatrix} m \\ k \end{pmatrix}$ of bounded k grow at polynomial rates on m, so that all entries of J^m for a Jordan box J corresponding to an eigenvalue λ of $|\lambda| < 1$ decrease exponentially as $m \to \infty$.

S7.8

The "gap" consists in the necessity of using, for nonreal eigenvalues, either complexified Jordan boxes in the complexified root subspaces or Jordan boxes over reals, which, however, does not affect the final result. (We leave it to the reader to fill in the details).

[12] Because all matrices corresponding to points of H^1 have the same Jordan canonical form. (Which exactly?)

[13] As an exercise, the reader can find explicitly all matrices T such that $A_T Q_\theta = Q_0$ to verify that these matrices are proportional to each other (for any fixed θ) and to determine the sign of det T (as a function of θ).

S7.9

The polynomials p_k, q_k defined by the formula $\begin{pmatrix} p_k \\ q_k \end{pmatrix} = M^k \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ satisfy a matrix equation $\begin{pmatrix} p_k \\ q_k \end{pmatrix} = M \begin{pmatrix} p_{k-1} \\ q_{k-1} \end{pmatrix}$, equivalent to the system of scalar recursion equations.[14]

$$p_{-1} = 0, \qquad p_0 = 1, \qquad p_{k+1} = \lambda p_k - p_{k-1}, \qquad q_k = p_{k-1}.$$

[Complete the details and show, by induction, that both polynomials $p_k(\lambda)$ and $q_k(\lambda)$ have leading coefficients equal to 1.] From those equations,

$$\frac{p_k}{q_k} = \lambda - \frac{q_{k-1}}{p_{k-1}} = \lambda - \frac{1}{p_{k-1}/q_{k-1}} = \lambda - \frac{1}{\lambda - \frac{1}{p_{k-2}/q_{k-2}}} = \ldots = \lambda - \cfrac{1}{\lambda - \cfrac{1}{\ddots \\ \lambda - \frac{1}{\lambda}}}.$$

The relative primality of $p_k(\lambda)$ and $q_k(\lambda)$ (in the ring $\mathbb{Z}[\lambda]$) may be verified as follows. The equalities $\begin{pmatrix} p_k \\ q_k \end{pmatrix} = M^k \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\begin{pmatrix} p_{k-1} \\ q_{k-1} \end{pmatrix} = M^{k-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ together give

$$\begin{pmatrix} p_k & p_{k-1} \\ q_k & q_{k-1} \end{pmatrix} = M^{k-1} \left(M \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \right) = M^{k-1} \circ \begin{pmatrix} \lambda & 1 \\ 1 & 0 \end{pmatrix},$$

[14] Similarly, for a linear recursive sequence [a sequence $p_0, p_1, \ldots$ satisfying a relation of the form $p_{k+m} = f(p_k, \ldots, p_{k+m-1})$, for any natural k, with a linear function of m variables f] the recursion relation may be written as a matrix system $\begin{pmatrix} p_{k+m} \\ \vdots \\ p_{k+1} \end{pmatrix} = \begin{pmatrix} \lambda_1 & \cdots & \cdots & \lambda_m \\ 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ & & 1 & 0 \end{pmatrix} \begin{pmatrix} p_{k+m-1} \\ \vdots \\ p_k \end{pmatrix}$. These are discrete analogs of linear autonomous ordinary differential equations (which are written as linear systems using the same, within a permutation of the rows, matrices). (Which matrix corresponds to the Fibonacci sequence?) The characteristic equation for the matrix of the system is $\mu^m = \lambda_1 \mu^{m-1} + \ldots + \lambda_m$, and a single proper direction corresponds to any eigenvalue. (Why?) Explicit formulas for the elements of recursive sequences may be obtained using the Jordan canonical forms of these matrices. Specifically, if all eigenvalues $\mu_1, \ldots, \mu_m$ are simple, then these elements are sums of geometric progressions: $p_k = \sum_{i=1}^{m} a_i \mu_i^k$. ($a_i$ are found using "initial" elements $p_0, \ldots, p_{m-1}$, or any other m consecutive elements of the sequence. Write out the corresponding system of equations on a_i and verify its unique solvability. Derive explicit formulas for the Fibonacci numbers). Readers will find further discussion of the subject in Arnol'd (1975). A similar approach provides a complete algebraic description of the continued fractions.

which yields $p_k q_{k-1} - p_{k-1} q_k = \det \begin{pmatrix} p_k & p_{k-1} \\ q_k & q_{k-1} \end{pmatrix} = -1$. (What follows from this equality? Fill in the details).

Lastly, $\chi(\lambda) = \frac{p_1}{q_1} \cdot \frac{p_2}{q_2} \cdot \ldots \cdot \frac{p_n}{q_n} = p_1 \cdot \frac{p_2}{p_1} \cdot \ldots \cdot \frac{p_n}{p_{n-1}} = p_n$. **QED**. (We leave it to the reader to fill in the details).

The matrix from section P7.9** is diagonalizable (since it is symmetric) and has all its eigenvalues simple (nonmultiple): indeed, for an eigenvector ${}^t(x_1, \ldots, x_n)$ corresponding to λ we have $x_2 = \lambda x_1$, $x_1 + x_3 = \lambda x_2$, $\ldots$, $x_{n-2} + x_n = \lambda x_{n-1}$, and a short computation shows by induction that all of these vectors are proportional to each other.

Furthermore, for a $k \times k$ matrix $A = \begin{pmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & \end{pmatrix}$ we have

$$(-1)^k \det A = \det(-A) = \chi(0) = p_k(0)$$
$$= \text{first component of vector} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^k \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

which is zero for $k = 2n + 1$; for $k = 2n$, the first trigonometric identity in section P7.9** is obtained equating the determinant to the product of the eigenvalues, taking into account some elementary trigonometric relations. (We leave it to the reader to fill in the details).

The second trigonometric identity in section P7.9** may be established by similar computations with $\chi(2) = \det(-\Delta)$. (Find it!)

A Property of Orthogonal Matrices

Problems

P8.0**

Preliminaries. Readers likely recall the following elementary geometric statement:

Theorem 1 *If the sides of an angle* α *in a Euclidean plane are orthogonal to the sides of an angle* β*, then* $\alpha = \beta$ *or* $\alpha + \beta = 180°$*; in other words,* $|\cos\alpha| = |\cos\beta|$*. The same rule ties the angles:* $\alpha = (\Pi, \Pi')$*between planes* Π*,* Π' *in a Euclidean space* $\mathbb{R}^3$ *defined as an angle between straight lines in* Π*,* Π' *orthogonal to* $\Pi \cap \Pi'$*, and* $\beta = S(\Lambda,\Lambda')$*, between the normals to* Π*,* Π'*, respectively.*

We may reformulate the foregoing theorem in a different way. Consider the orthogonal projection P: $\mathbb{R}^3 \to \Pi$ ($P \restriction_\Pi = E_\Pi$, $\ker P = \Lambda$). Define a function on the set of all polygons in Π' [or, if you wish, all **squarable** (also called **Jordan-measurable**) bounded subsets of Π'], which may be called an *area projection factor*

$$S \mapsto K_{\Pi,\Pi'}(S) := \frac{\text{area}(P(S))}{\text{area}(S)}.$$

Actually this factor is the same for all S, that is, $K_{\Pi,\Pi'} = |\cos\alpha|$ (because, in fact, $K_{\Pi,\Pi'}$ is equal to the determinant of a map $P \restriction_{\Pi'} : \Pi' \to \Pi$ calculated with respect to the orthogonal bases of Π, Π'). Considering an orthogonal projection $L = E_{\mathbf{R}^3} - P : \mathbf{R}^3 \to \Lambda$ yields for a *length projection factor* $K_{\Lambda,\Lambda'}(S) := \frac{\text{length}(L(S))}{\text{length}(S)}$ (defined for measurable bounded subsets $S \subset \Lambda'$) that $K_{\Lambda,\Lambda'} = |\cos\beta|$. Thus, Theorem 1 states that $K_{\Pi,\Pi'} = K_{\Lambda,\Lambda'}$ (Fig. 1).

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_8,

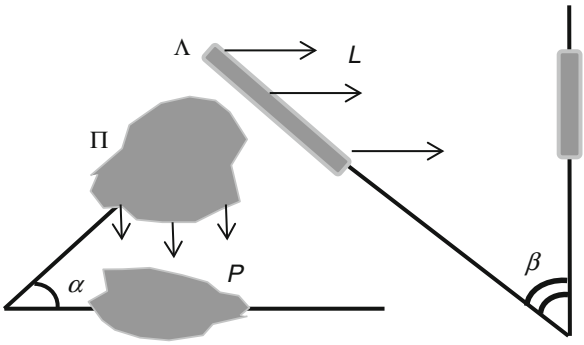


Fig. 1 Equality of projection factors

Let us formulate the following generalization:

Theorem 2 *Let* $\mathbb{R}^n = \Pi \bigoplus \Pi' = \Lambda \bigoplus \Lambda'$ *be orthogonal decompositions of a Euclidean space such that* Π, Π' *are hyperplanes (resp.* Λ, Λ' *are straight lines). Then* $K_{\Pi,\Pi'} = K_{\Lambda,\Lambda'}$ *(where the usual area is substituted by the* $n-1$*-dimensional area).*

Theorem 2 is evident when $\Pi = \Pi'$, and for $\Pi \neq \Pi'$ it can be reduced (how?) to Theorem 1 by considering a two-dimensional plane $\Lambda + \Lambda'$, which is the orthogonal complement to $\Pi \cap \Pi'$. But, alternatively, we can prove this geometric theorem if we understand its algebraic meaning. Assume for a moment that Π, Π' are not orthogonal [$\Leftrightarrow \Pi' \cap \ker P = 0 \Leftrightarrow \ker (P \mid_{\Pi'}) = 0 \Leftrightarrow \ker (L \mid_{\Lambda'}) = 0 \Leftrightarrow \Lambda$, Λ' are not orthogonal]. Consider an orthogonal linear operator A mapping an orthonormal basis $e_1,\ldots,e_n$ with $e_1,\ldots,e_{n-1} \in \Pi$ and $e_n \in \Lambda$ onto a basis with $Ae_1,\ldots, Ae_{n-1} \in \Pi'$ and $Ae_n \in \Lambda'$. Let $\varsigma = (a_{ij})_{i,j\,=\,1,\ldots,\,n}$ be the matrix of this operator with respect to the basis $e_1,\ldots,e_n$. Theorem 2 states that the element a_{nn} of this matrix by absolute value is equal to its cofactor,

$$\left(K_{\Lambda,\Lambda'} =\right)\ |a_{nn}| = |\det(a_{ij})|_{i,j=1,\ldots,n-1} \left(= K_{\Pi,\Pi'}\right),$$

which is true since the matrix is orthogonal (${}^t\varsigma = \varsigma^{-1}$). On the other hand, if Π, Π' are orthogonal, then Λ, Λ' are orthogonal also, hence, $K_{\Pi,\Pi'} = 0 = K_{\Lambda,\Lambda'}$, completing the proof of Theorem 2. We recommend that readers work out the details of this proof, although in what follows we will present a generalization that can require knowledge of more sophisticated techniques! The suggested generalization of Theorem 2 is as follows.

Theorem 3

- *In geometrical language*: Let $\mathbb{R}^n = \Pi \bigoplus \Pi' = \Lambda \bigoplus \Lambda'$ *be orthogonal decompositions of a Euclidean space such that* $\dim \Pi = \dim \Pi'$ *(and so* $\dim \Lambda = \dim \Lambda'$*). Then* $K_{\Pi,\Pi'} = K_{\Lambda,\Lambda'}$ *(for the k- and n-k-dimensional areas, respectively, where* $\dim \Pi = k$ *and* $\dim \Lambda = n - k$*).*
- *In algebraic language*: Minors of an orthogonal matrix by absolute value are equal to their cofactors.

This theorem relates to an interaction of skew symmetry and orthogonality features, i.e., properties of determinants and similar multilinear functions, and Euclidean spaces (real vector spaces equipped with scalar products). We presume readers are familiar with these features within the scope of a common university course (see the section "Using the Stars on Problems" for details), and on this basis we will develop the tools necessary to prove Theorem 3. In what follows, these tools are introduced as a series of problems (sections P8.1***, P8.2***, P8.3***, P8.4***, P8.5***, P8.6***, P8.7***, P8.8***, P8.9***, P8.10***, P8.11***, P8.12***, P8.13***, P8.14***, P8.15***, P8.16***, and P8.18***). Readers experienced in these techniques may start proving Theorem 3 (see section P8.19*** for the refined formulation) while skipping most of the other problems in this group. These tools have multiple applications. For example, they allow us to extend a definition of the angle between straight lines or between hyperplanes to k-dimensional planes of $\mathbb{R}^n$ for every $0 < k < n$ (section P8.17***) and to extend a definition of the vector product of two vectors in $\mathbb{R}^3$ – to any number of vectors and multivectors in $\mathbb{R}^n$ (sections P8.20*** and P8.21***). For more on the applications, consult Van der Waerden (1971, 1967), Lang (1965), Kostrikin and Manin (1980), Arnol'd (1989), Warner (1983), Godbillon (1969), Schutz (1982), and Dubrovin et al. (1986).

*P8.1****

Recall that a function or a map defined on a Cartesian product of vector spaces is **multilinear** (if there are k factors, it is called "***k*-linear**") if it is linear on each factor, while the arguments related to the remaining factors are fixed. Recall that a multilinear function or map is referred to as **skew-symmetric** if it equals zero any time when two elements of its multiargument input are equal. This definition is equivalent (except for vector spaces over *what* kind of fields?) to the following definition: the value of a function or map is multiplied by -1 when two elements of its multiargument input are transposed with each other. Also recall that the **skew-symmetric**, or **exterior**, ***k*-power**, $\wedge^k \mathbb{R}^n$, of a finite-dimensional space $\mathbb{R}^n$ [1] is

- $\mathbb{R}$ when $k = 0$;
- for $k > 0$, a vector space linearly generated by the elements of a Cartesian power $(\mathbf{R}^n)^k = \mathbf{R}^n \times \ldots \times \mathbf{R}^n{}_k$. The generators are denoted by $x_1 \wedge \ldots \wedge x_k$ or $\bigwedge_{i=1}^{k} x_i$. They are not linearly independentbecause they must obey linear relations of two kinds: (1) **multilinearity**, $[\ldots \wedge (\alpha x'_i + \beta x''_i) \wedge \ldots] = \alpha(\ldots \wedge x'_i \wedge \ldots) +$

[1] Readers experienced in abstract algebra know that exterior products and other multilinearity-related techniques are developed in algebra for free modules over associative–commutative unitary rings. However, here we do not need such a level of generality.

$\beta(\ldots \wedge x''_i \wedge \ldots)$, $\forall \alpha,\beta \in \mathbb{R}$, $\forall i = 1,\ldots, k$, and (2) **skew symmetry**, $\bigwedge_{i=1}^{k} x_i = 0$ if $x_i = x_j$ for some $i \neq j$ (or, equivalently, a transposition $x_i \rightleftarrows x_j$ multiplies the generator by -1; *prove* this equivalence).

In more technical terms, $\wedge^k \mathbb{R}^n$ is a quotient space of a continuum-dimensional vector space over $\mathbb{R}$ whose basis elements correspond to the points of $(\mathbb{R}^n)^k$ over its subspace linearly generated by all elements of the forms $\left(\ldots, \underset{i}{x}, \ldots, \underset{j}{x}, \ldots\right)$ and $(\ldots, \underset{m}{\alpha x + \beta y}, \ldots) - \alpha(\ldots, \underset{m}{x}, \ldots) - \beta(\ldots, \underset{m}{y}, \ldots)$ $(1 \le i \le k,\ 1 \le m \le k,\ \alpha,\ \beta \in \mathbb{R},\ x,\ y \in \mathbb{R}^n)$. *Establish* the **universality** property of the map $\begin{cases} (\mathbb{R}^n)^k & \overset{\wedge^k}{\to} & \wedge^k \mathbf{R}^n \\ (x_1, \ldots, x_k) & \mapsto & \bigwedge_{i=1}^{k} x_i \end{cases}$: it is multilinear, skew-symmetric, and such that for any multilinear skew-symmetric map $(\mathbb{R}^n)^k \to L$ there exists a unique linear map $\wedge^k \mathbb{R}^n \to L$ making the diagram $\begin{array}{ccc} (\mathbf{R}^n)^k & \overset{\wedge^k}{\to} & \wedge^k \mathbf{R}^n \\ & \searrow & \downarrow \\ & & L \end{array}$ commutative.

P8.2***

Prove the following key technical statement about exterior powers.

Lemma *For linear combinations* $y_1 = \alpha_{11}x_1 + \ldots + \alpha_{1k}x_k$, $\ldots$, $y_k = \alpha_{k1}x_1 + \ldots + \alpha_{kk}x_k$,

$$y_1 \wedge \ldots \wedge y_k = \det\left(\alpha_{ij}\right) \cdot x_1 \wedge \ldots \wedge x_k.$$

P8.3***

Recall that **skew-symmetric**, or **exterior, *k*-forms** in $\mathbb{R}^n$ are k-linear skew-symmetric functions $(\mathbb{R}^n)^k \to \mathbb{R}$. These forms are naturally identified with linear functionals on $\wedge^k \mathbb{R}^n$. To be exact, any such functional, f, determines a composition $(\mathbf{R}^n)^k \overset{\wedge^k}{\to} \wedge^k \mathbf{R}^n \overset{f}{\to} \mathbf{R}$, which is an exterior k-form. Thus, there is a map $f \mapsto f \circ \wedge^k$ defined on the vector space $(\wedge^k \mathbb{R}^n)^*$ [2] to the vector space of exterior k-forms in $\mathbb{R}^n$. (It is a linear map; *why*?) *Verify* the invertibility of this map (thus, it is a linear isomorphism).

[2] Recall that the dual space L^* for a vector space L is a vector space of all linear functionals in L.

At this point, readers can familiarize themselves with exterior forms, analyzing the following special cases.

- Exterior 0-forms are multiplications by scalars: $\alpha \mapsto \alpha^* : \begin{cases} \wedge^0 \mathbf{R}^n = \mathbf{R} \to \mathbf{R} \\ \beta \mapsto \alpha\beta \end{cases}$ ($\alpha \in \mathbb{R}$).
- Exterior 1-forms are linear functionals.
- Exterior n-forms in $\mathbb{R}^n$ are proportional to each other. (*Prove* using section P8.2***.) In linear algebra these forms are known as **determinants** and in geometry as **oriented volumes**. (*Explain* why the n-forms are identified with those objects.)

*P8.4****

Prove that $\dim \wedge^k \mathbf{R}^n = \binom{n}{k}$ and, specifically, for an ordered basis $e_1,\ldots,e_n$ in $\mathbb{R}^n$, $\binom{n}{k}$ elements $e_{i_1} \wedge \ldots \wedge e_{i_k}$ $(1 \le i_1 < \ldots < i_k \le n)$ form a basis of $\wedge^k \mathbb{R}^n$. Therefore, $\wedge^k \mathbb{R}^n = 0$ for $k > n$, $\wedge^n \mathbb{R}^n \cong \wedge^0 \mathbb{R}^n = \mathbb{R}$, $\wedge^{n-1} \mathbb{R}^n \cong \wedge^1 \mathbb{R}^n = \mathbb{R}^n$, and, generally, $\wedge^k \mathbb{R}^n \cong \wedge^{n-k} \mathbb{R}^n$, $\forall k = 0,\ldots,n$. [The isomorphism $\wedge^k \mathbb{R}^n \cong \wedge^{n-k} \mathbb{R}^n$, defined such that $e_{i_1} \wedge \ldots \wedge e_{i_k} \mapsto \pm e_{i_{k+1}} \wedge \ldots \wedge e_{i_n}$ ($\{i_1,\ldots,i_n\} = \{1, \ldots, n\}$, $i_1 < \ldots < i_k$, $i_{k+1} < \ldots < i_n$) for an ordered orthobasis $e_1,\ldots,e_n$ in a Euclidean space $\mathbb{R}^n$, is referred to as a "star" isomorphism. It is discussed in more detail below.]

*P8.5****

Show that $x_1 \wedge \ldots \wedge x_k$ equals zero if and only if x_i are linearly dependent. Also, *establish* the invertibility of the statement in section P8.2***: $y_1 \wedge \ldots \wedge y_k = \alpha \cdot x_1 \wedge \ldots \wedge x_k \neq 0$ only if all of y_i belong to the subspace of $\mathbb{R}^n$ generated by $x_1,\ldots,x_k$.

*P8.6****

Determine the situations where any linear combinations of the generators $x_1 \wedge \ldots \wedge x_k$ (in other words, all elements of the vector space $\wedge^k \mathbb{R}^n$) themselves are some generators $y_1 \wedge \ldots \wedge y_k$.

P8.7[***]

The **Grassmannian variety** of the k-dimensional vector subspaces of $\mathbb{R}^n$, in short, the ***k*-Grassmannian** $G(n,k)$, is the set of these subspaces.[3] Let $\Pi^k \in G(n,k)$. According to section P8.2[***], the map $\wedge^k$ restricted to a set of the bases of Π^k defines a straight line in $\wedge^k \mathbb{R}^n$. (*Work out* the details.) Therefore, we have defined a map from $G(n,k)$ to the space of straight lines in $\wedge^k \mathbb{R}^n$, which is denoted $P(\wedge^k \mathbb{R}^n)$.[4] According to section P8.5[***], it is always an injective map (that is, it does not paste points of $\wedge^k \mathbb{R}^n$ together). The result of section P8.6[***] determines the situations where this map is surjective [*surjectivity* is the quality whereby all points of $P(\wedge^k \mathbb{R}^n)$ have preimages].

An example: $G(4,2)$ is mapped onto a quadric in $\mathbb{R}P^5$ (Julius Plücker's theorem).

P8.8[***]

Recall that the **skew-symmetric**, or **exterior**, **algebra of** $\mathbb{R}^n$ is a formal direct sum $\wedge \mathbf{R}^n := \sum_{k=0}^{n} {}^{\oplus} \wedge^k \mathbf{R}^n$. (Usually, $\wedge^k \mathbb{R}^n$ are referred to as **homogeneous, of degree *k*, components** of $\wedge \mathbb{R}^n$.) It is a 2^n-dimensional vector space, but in addition, it is equipped with an extra bilinear operation $(\wedge \mathbf{R}^n) \times (\wedge \mathbf{R}^n) \xrightarrow{\wedge} \wedge \mathbf{R}^n$ which is called an **exterior multiplication**. "$\wedge$" may be set on a fixed basis of $\wedge \mathbb{R}^n$ and then extended to all of $\wedge \mathbb{R}^n$ by bilinearity. Fulfilling this program, for a fixed basis

$$\{1\} \cup \{e_{i_1} \wedge \ldots \wedge e_{i_k} : \quad k = 1, \ldots, n, \quad 1 \le i_1 < \ldots < i_k \le n\},$$

we will define

$$1 \wedge 1 := 1, \qquad 1 \wedge (e_{i_1} \wedge \ldots \wedge e_{i_k}) := (e_{i_1} \wedge \ldots \wedge e_{i_k}) \wedge 1 :$$
$$= e_{i_1} \wedge \ldots \wedge e_{i_k}, (e_{i_1} \wedge \ldots \wedge e_{i_k}) \wedge \left(e_{i_{k+1}} \wedge \ldots \wedge e_{i_{k+l}}\right) := e_{i_1} \wedge \ldots \wedge e_{i_{k+l}}.$$

Show the independence of this definition from the initial basis, that is, that

$$(x_1 \wedge \ldots \wedge x_k) \wedge (x_{k+1} \wedge \ldots \wedge x_{k+l}) = x_1 \wedge \ldots \wedge x_{k+l}, \quad \forall x_1, \ldots, x_{k+l} \in \mathbf{R}^n.$$

Establish the skew commutativity $x^k \wedge x^l = (-1)^{kl} x^l \wedge x^k$ ($x^k \in \wedge^k \mathbb{R}^n$, $x^l \in \wedge^l \mathbb{R}^n$).

Exterior multiplication is an associative operation, so it causes m-linear exterior multiplications $(\wedge \mathbf{R}^n)^m \xrightarrow{\wedge} \wedge \mathbf{R}^n$ for all $m = 0,1,2,\ldots$ such that

[3] Supplied with determinate topological and manifold structures, which we will not consider.

[4] Readers familiar with projective geometry know that $P(\wedge^k \mathbb{R}^n)$ is referred to as a projectivization of the vector space $\wedge^k \mathbb{R}^n$.

$$\left(\bigwedge_{i=1}^{k_1} x_i\right) \wedge \ldots \wedge \left(\bigwedge_{i=k_{m-1}+1}^{k_m} x_i\right) = \bigwedge_{i=1}^{k_1+\ldots+k_m} x_i.$$

Obviously, $\wedge\mathbb{R}^n$ as a ring is generated by its component of degree 1, $\wedge^1\mathbb{R}^n$, namely, the map $(\wedge^1\mathbf{R}^n)^k \xrightarrow{\wedge} \wedge^k\mathbf{R}^n$ coincides with $(\mathbf{R}^n)^k \xrightarrow{\wedge^k} \wedge^k\mathbf{R}^n$ defined in section P8.1***.

*P8.9****

The space of all exterior forms in $\mathbb{R}^n$ is defined as

$$(\wedge\mathbf{R}^n)^* = \left(\sum_{k=0}^{n} \oplus \wedge^k\mathbf{R}^n\right)^* \cong \sum_{k=0}^{n} \oplus \left(\wedge^k\mathbf{R}^n\right)^*.$$

Vector spaces $(\wedge\mathbb{R}^n)^*$ and $\wedge\mathbb{R}^{n*}$ have equal dimensions of 2^n. We want to define those spaces in such a way that will enable us to operate with the elements of $\wedge\mathbb{R}^{n*}$ as if they were exterior forms in $\mathbb{R}^n$. In other words, our aim is to have an exterior multiplication of the exterior forms. The aim is to preserve the homogeneity degree, so that the elements of $\wedge^k\mathbb{R}^{n*}$ are k-forms. To do this, let us fix dual ordered bases $e_1,\ldots,e_n$, $e_1^*,\ldots,e_n^*$ in $\mathbb{R}^n$ and $\mathbb{R}^{n*}$, respectively $[e_i^*(e_j) = \delta_{ij}]$. We stipulate by our definition that the pairs $\{1\}$, $\{1\}$ and $\{e_{i_1} \wedge \ldots \wedge e_{i_k} : \ 1 \le i_1 < \ldots < i_k \le n\}$, $\left\{e_{i_1}^* \wedge \ldots \wedge e_{i_k}^* : \ 1 \le i_1 < \ldots < i_k \le n\right\}$ form dual bases in, respectively, $\wedge^0\mathbb{R}^n$, $\wedge^0\mathbb{R}^{n*}$ and $\wedge^k\mathbb{R}^n$, $\wedge^k\mathbb{R}^{n*}$ $(k = 1,\ldots,n)$:

$$e_{i_1}^* \wedge \ldots \wedge e_{i_k}^*\left(e_{j_1} \wedge \ldots \wedge e_{j_k}\right) := \delta_{j_1,\ldots,j_k}^{i_1,\ldots,i_k} = \begin{cases} 1 & \text{when } i_v = j_v, \ \forall v = 1,\ldots,k, \\ 0 & \text{otherwise}. \end{cases}$$

After that, a linear action of the elements of $\wedge^k\mathbb{R}^{n*}$ on $\wedge^k\mathbb{R}^n$ is defined by bilinearity. *Prove* that this definition of $(\wedge\mathbb{R}^n)^* \cong \wedge\mathbb{R}^{n*}$ actually does not depend on a source basis $e_1,\ldots,e_n$. To be more exact, *describe* this construction in invariant terms *proving* that

1. The *elements of* $\wedge^1\mathbb{R}^{n*}$ *act on* $\wedge^1\mathbb{R}^n$ *in exactly the same way as they act on* $\mathbb{R}^n$ *when they are considered as elements of* $\mathbb{R}^{n*}$;
2. *The exterior product of* the *elements of* $\wedge^k\mathbb{R}^{n*}$ *and* $\wedge^l\mathbb{R}^{n*}$ *is calculated by the rule*

$$\omega^k \wedge \omega^l (x_1, \ldots, x_{k+l}) = \sum_\sigma \operatorname{sign}(\sigma) \cdot \omega^k \left(x_{\sigma(1)}, \ldots, x_{\sigma(k)}\right) \cdot \omega^l \left(x_{\sigma(k+1)}, \ldots, x_{\sigma(k+l)}\right),$$

where the sum is taken over all permutations of the set $1,\ldots, k + l$ *such that* $\sigma(1) < \ldots < \sigma(k)$ *and* $\sigma(k + 1) < \ldots < \sigma(k + l)$.

Write out a similar formula for an arbitrary number of exterior cofactors. Also, prove that for $x_1,\ldots, x_k \in \mathbb{R}^n$, $x_1^*, \ldots, x_k^* \in \mathbf{R}^{n*}$ $(k = 1,\ldots, n)$,

$$x_1^* \wedge \ldots \wedge x_k^*(x_1, \ldots, x_k) = \det \left(x_i^*(x_j)\right)_{i,j=1,\ldots,k}.$$

[In other words, for dual ordered bases $e_1,\ldots,e_n$, $e_1^*, \ldots, e_n^*$ in $\mathbb{R}^n$ and $\mathbb{R}^{n*}$, respectively, and $x_i = \sum x_{ij} e_j \in \mathbb{R}^n$ $(i = 1,\ldots, k \leq n)$, we have $e_{m_1}^* \wedge \ldots \wedge e_{m_k}^* \times (x_1, \ldots, x_k) = \det \left(x_{m_i, j}\right)_{i,j=1,\ldots,k}$.] Establish the skew commutativity $\omega^k \wedge \omega^l = (-1)^{kl} \omega^l \wedge \omega^k$ $(\omega^k \in \wedge^k \mathbb{R}^{n*}, \omega^l \in \wedge^l \mathbb{R}^{n*})$.

Readers experienced in abstract algebra have already noticed that we have defined a homogeneous (degree-preserving) isomorphism $\wedge \mathbb{R}^{n*} \cong (\wedge \mathbb{R}^n)^*$. [The preservation means that $\wedge^k \mathbb{R}^{n*}$ are mapped onto $(\wedge^k \mathbb{R}^n)^*$, $\forall k = 0,\ldots, n$.] Such an isomorphism is not unique, but the preceding one is commonly accepted and has important advantages. Another way of establishing such an isomorphism is by extracting it from functorial isomorphisms of tensor powers $\bigotimes^k \mathbb{R}^{n*} \cong (\bigotimes^k \mathbb{R}^n)^*$. This yields isomorphisms $\wedge^k \mathbb{R}^{n*} \cong (\wedge^k \mathbb{R}^n)^*$ differing from ours by scalar factors $(k!)^{-1}$, respectively. (Geometrically, these correspond to the volumes of simplexes, while ours correspond to the volumes of parallelepipeds.) A detailed discussion of this and related topics is beyond the scope of our problem book. Interested readers should consult advanced literature in abstract algebra.

P8.10[***]

An exterior n-form Ω with fixed first k arguments becomes an $n - k$-form. (Generally, an exterior $k + l$-form with fixed k arguments becomes an l-form for the rest of our argument.) That is, let $x_1,\ldots, x_k \in \mathbb{R}^n$. If they are linearly dependent, then this $n - k$-form is zero. If not, complement them to an ordered basis $x_1,\ldots, x_n$. Let $x_1^*, \ldots, x_n^*$ be the dual basis in $\mathbb{R}^{n*}$. We then have $\Omega = \alpha \cdot x_1^* \wedge \ldots \wedge x_n^*$ with $\alpha \in \mathbb{R}$. Fixing the arguments $x_1,\ldots, x_k$ yields, according to section P8.9[***], the $n - k$-form $\alpha \cdot x_{k+1}^* \wedge \ldots \wedge x_n^*$, which is, within proportionality, the Euclidean $(n - k$-dimensional) area of the orthogonal projection onto the orthogonal complement to a (k-dimensional) plane spanned on $x_1,\ldots, x_k$. Instead, we may, in accordance with section P8.3[***], fix the "k-dimensional argument" $x_1 \wedge \ldots \wedge x_k$, as is clearly shown in this diagram:

$$\begin{array}{ccccc} (x_1, \ldots, x_k) & \overset{\wedge^k}{\mapsto} & \bigwedge_{i=1}^k x_i \in \wedge^k \mathbf{R}^n & & \\ \times & & \times & \overset{\wedge}{\longrightarrow} \wedge^n \mathbf{R}^n \overset{\Omega}{\longrightarrow} \mathbf{R}. & \\ (\mathbf{R}^n)^{n-k} & \overset{\wedge^{n-k}}{\longrightarrow} & \wedge^{n-k} \mathbf{R}^n & & \end{array}$$

Using linearity, we will obtain an $n-k$-form from an n-form Ω by fixing any element of $\wedge^k\mathbb{R}^n$ as the "k-dimensional argument." Therefore, we have defined a so-called **insertion operator**

$$I_\Omega : \begin{cases} \wedge^k \mathbf{R}^n & \to & \wedge^{n-k}\mathbf{R}^{n*} \\ X & \mapsto & \imath_X\Omega := \Omega\left(\underbrace{X}_{k}, \underbrace{\ldots}_{n-k}\right) \end{cases}$$

By this definition, I_Ω is a linear operator. *When* is it an isomorphism?

*P8.11****

For a finite-dimensional space L, $\dim L = \dim L^*$. However, there is no "canonical" isomorphism between those spaces. You might, for instance, define an isomorphism $L \cong L^*$ using any pair of bases in L, L^*, respectively. But usually, it is defined in a more invariant way, namely, using a nondegenerate bilinear form $L \times L \xrightarrow{\Phi} field$. The following linear maps are connected to any fixed bilinear form Φ: $D'_\Phi : \begin{cases} L & \to & L^* \\ v & \mapsto & \Phi(v,.) \end{cases}$, $D''_\Phi : \begin{cases} L & \to & L^* \\ v & \mapsto & \Phi(.,v) \end{cases}$.[5] (*When* do they coincide?) *Prove* that if one of the connected linear maps is an isomorphism, then both of them are isomorphisms, and this is a necessary and sufficient condition in order for a bilinear form to be nondegenerate.

In what follows we will deal with isomorhisms (**dualities**) $L \xrightarrow{D} L^*$ connected to scalar products (bilinear symmetric positive definite forms) in L. *Check* that if an isomorphism $L \xrightarrow{D} L^*$ is connected to a scalar product in L, then for any orthobasis (with respect to this product) $e_1,\ldots$, the basis $De_1,\ldots$ in L^* is dual: $De_i(e_j) = \delta_{ij}$. *Show* that, conversely, for a pair of dual bases $e_1,\ldots,e_1^*,\ldots$ in L, L^*, respectively, an isomorphism $L \xrightarrow{D} L^*$, for which $De_1 = e_1^*,\ldots$, is connected to a scalar product for which $e_1,\ldots$ is an orthobasis. (Geometrically speaking, De, for a unit vector e, is a length of an orthogonal projection onto a straight line in L spanned on e and oriented by it; in other words, De is a linear functional whose 1-set is the affine hyperplane orthogonal to e at its end.) Also, *check* that for the isomorphism $L \xrightarrow{D} L^*$ connected to a scalar product Φ in L, a dual form Φ^* in L^* defined as $\Phi^*(x^*, y^*) :=$

[5] A linear map $L \xrightarrow{D} L^*$ defines the bilinear forms $\Phi'_D(x,y) := Dx(y)$, $\Phi''_D(x,y) := Dy(x)$. Readers experienced in abstract algebra know that the correspondences $D \mapsto \Phi'_D$, $\Phi \mapsto D'_\Phi$ and $D \mapsto \Phi''_D$, $\Phi \mapsto D''_\Phi$ define two pairs of functorial isomorphisms $\mathrm{Hom}(L,L^*) \rightleftarrows L^* \bigotimes L^*$ between vector spaces of linear operators $L \to L^*$ and bilinear forms in L.

$\Phi\,(D^{-1}x^*, D^{-1}y^*)$ is a scalar product, and the orthonormality of a basis $e_1, \ldots$ with respect to Φ is equivalent to the orthonormality of $De_1, \ldots$ with respect to Φ^*.

So that readers may feel comfortable with Φ^*, we suggest using it to solve the following simple problem: find the distance from the origin to an affine plane (a translation of a vector subspace) of codimension m in a Euclidean space. Let this plane (denote it M) be an intersection of m hyperplanes $\langle x, a_i\rangle = b_i$, $i = 1, \ldots, m$. The orthogonal complement L (spanned on $a_1, \ldots, a_m$) passing through the origin is a subspace of the Euclidean space and therefore is itself a Euclidean space. Consider the Gram matrix $G = \left(\langle a_i, a_j\rangle\right)_{i,j=1,\ldots,m}$ relating to the scalar product in L and its inverse $G^{-1} = (g^{ij})_{i,j\,=\,1,\ldots,n}$; thus, the squared distance solving this problem is ${}^tbG^{-1}b = \sum\limits_{i,j} g^{ij}b_ib_j$ (for $m = 1$, the distance is $|b|/|a|$). Indeed, define the duality $L \overset{D}{\to} L^*$ connected to the scalar product in L and the dual scalar product $\langle\rangle^{**}$ in L^*. The required distance is $|v| = |Dv|^*$, where v is the radius vector of the point of the intersection of M and L. Decomposing Dv with respect to the dual basis $a_1^*, \ldots$ $(a_i^*(a_j) = \delta_{ij})$ we evidently have $Dv = \langle v, .\rangle = \sum b_i a_i^*$. Now *prove* the following lemma.

Lemma *The Gram matrix relating to the scalar product* $\langle\rangle^*$ *in* L^* *is* G^{-1} *with respect to* the *basis* $a_1^*, \ldots$.

Hence, $|Dv|^{*2} = {}^tbG^{-1}b$. **QED**. [A clarification of this proof is attained by moving the tedious calculations (unavoidable for any proof of this claim) to the lemma!]

P8.12[***]

Let us extend the duality $\mathbf{R}^n \overset{D}{\to} \mathbf{R}^{n*}$ to a duality $\wedge\mathbf{R}^n \overset{\wedge D}{\to} \wedge\mathbf{R}^{n*}$ so just as for an orthobasis $e_1, \ldots, e_n$ and $x_1, \ldots, x_k \in \mathbb{R}^n$, $\left|\left[\wedge D\left(\underset{i=1}{\overset{k}{\wedge}} e_i\right)\right](x_1, \ldots, x_k)\right|$ would be the Euclidean (k-dimensional) area of the orthogonal projection of $x_1, \ldots, x_k$-sided (k-dimensional) parallelogram onto a (k-dimensional) plane spanned on $e_1, \ldots, e_k$. Thus, we are seeking $\wedge D\left(\underset{i=1}{\overset{k}{\wedge}} e_i\right) = \underset{i=1}{\overset{k}{\wedge}} e_i^*$, and according to section P8.11[***], $\wedge D\left(\underset{i=1}{\overset{k}{\wedge}} e_i\right) = \underset{i=1}{\overset{k}{\wedge}} De_i$. Show that the extension is unique, as there exists a unique isomorphism $\wedge\mathbf{R}^n \overset{\wedge D}{\to} \wedge\mathbf{R}^{n*}$ such that $\wedge D(\wedge^k \mathbb{R}^n) = \wedge^k\mathbb{R}^{n*}$ $(k = 0, \ldots, n)$, $\wedge D$ is the identity operator on $\wedge^0\mathbb{R}^n$, $\wedge D = D$ on $\wedge^1\mathbb{R}^n$, and $\wedge D\left(\underset{i}{\wedge} x^{k_i}\right) = \underset{i}{\wedge}\left[\wedge D\left(x^{k_i}\right)\right]$ $(x^{k_i} \in \wedge^{k_i}\mathbf{R})$. Using the identification of $\wedge\mathbb{R}^{n*}$ and $(\wedge\mathbb{R}^n)^*$ (as defined in section P8.9[***]), *show* that the isomorphism $\wedge\mathbf{R}^n \overset{\wedge D}{\to} (\wedge\mathbf{R}^n)^*$ will be a duality

(connected to a scalar product) in $\wedge\mathbb{R}^n$ if D is a duality in $\mathbb{R}^n$; also *show* that $\{1\} \cup \{e_{i_1} \wedge \ldots \wedge e_{i_k} : k = 1, \ldots, n,\ \ 1 \le i_1 < \ldots < i_k \le n\}$ will be an orthobasis in $\wedge\mathbb{R}^n$ when $e_1, \ldots$ is an orthobasis in $\mathbb{R}^n$. Therefore, $| x_1 \wedge \ldots \wedge x_k |$ will equal the Euclidean (k- dimensional) area of the $x_1, \ldots, x_k$-sided (k-dimensional) parallelogram. (*Prove*.)

Hence, by the (multidimensional) Pythagorean theorem, the squared (k-dimensional) area of a (k-dimensional) parallelogram in a Euclidean space of dimension $n \ge k$ equals the sum of squared (k-dimensional) areas of orthogonal projections of this parallelogram onto all k-dimensional coordinate planes – for any fixed orthogonal coordinate system. For (k-dimensional) squares (the parallelograms whose sides are formed by orthogonal unit vectors) this result can be reformulated as follows: $\forall k = 1, \ldots, n$, the sum of all squared $k \times k$ minors contained in any fixed set of k rows (or columns) of an orthogonal $n \times n$ matrix equals 1.

We will now prove the well-known **Lagrange identity** $(\sum x_i^2) \cdot (\sum y_i^2) - (\sum x_i y_i)^2 = \sum_{i<j} (x_i y_j - x_j y_i)^2$ using this result. The **Cauchy-Schwarz-Bunyakovskii (CSB) inequality** enables one to define angles between vectors in a Euclidean space so that $\cos \sphericalangle(x, y) = \langle x, y \rangle / (|x||y|)$ ($\langle ., . \rangle$ denotes the scalar product). Thus, for the vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$ equipped with a scalar product $\langle x, y \rangle = \sum x_i y_i$ we have

$$\begin{aligned}
&\left(\sum x_i^2\right) \cdot \left(\sum y_i^2\right) - \left(\sum x_i y_i\right)^2 = |x|^2 |y|^2 - \langle x, y \rangle^2 = |x|^2 |y|^2 \cdot \left(1 - \cos^2 \sphericalangle(x, y)\right) \\
&= |x|^2 |y|^2 \cdot \sin^2 \sphericalangle(x, y) \\
&= \text{the squared area of a parallelogram with the sides } x,\ y \\
&= \text{the sum of squared areas of its orthogonal projections onto the coordinate planes} \\
&= \sum_i \left(x_i y_j - x_j y_i\right)^2.
\end{aligned}$$

(In section P9.11**, the "Convexity and Related Classical Inequalities" problem group, we discuss a proof of Lagrange's identity using properties of determinants and not referring to the CSB inequality.)

Recall that the (k-dimensional) area of a union of a finite number of (k-dimensional) parallelograms intersecting with each other at most by their boundaries is defined as the sum of their areas, and the area of any other **squarable** (or **Jordan-measurable**) set S as the common limit for areas of S_q' and S_q'', which are unions of the preceding form and such that $S_q' \subseteq S \subseteq S_q''$, $\forall q$ and $area(S_q'') - area(S_q') \to 0$ as $q \to \infty$. *Prove* the foregoing claim about the sum of squared (k-dimensional) areas of orthogonal projections for arbitrary squarable *planar* (i.e., not extending beyond some k-dimensional plane) set.

*P8.13****

The next step consists in considering **oriented** Euclidean spaces. As readers know, we can orient the plane $\mathbb{R}^2$ by specifying the first and second of given basis vectors. (Obviously, the orientation may be established by choosing the clockwise or

counterclockwise direction for the shorter of the two rotations between these vectors.) Two ordered bases define the same or opposite orientations of the plane, depending on the direction of the rotation from the first to the second basis vector of those bases. Hence, the bases defining the same (opposite) orientations on $\mathbb{R}^2$ are bound by linear transformations of positive (resp. negative) determinants. (*Work out* the details.) Similarly, we will orient $\mathbb{R}^n$ by ordering any fixed basis. The orientation is in fact an equivalence class of bases bound by linear transformations having positive determinants. Therefore, a vector space may have two orientations opposite to one another.[6]

From section P8.3*** above we know that the totality of all exterior n-forms in $\mathbb{R}^n$ is a one-dimensional space, that is, geometrically, a straight line. Specifying the direction, formally, by choosing one of the two connected components after deleting the origin creates an orientation on a straight line. We can establish a correspondence between two parts of the punctured straight line $(\wedge^n \mathbb{R}^n)^* \setminus \{0\}$ and the equivalence classes of bases in $\mathbb{R}^n$. Therefore, a nonzero exterior n-form in $\mathbb{R}^n$ defines the orientation of this space. The **oriented** (n-dimensional) Euclidean volume is an exterior n-form Ω^n (we usually drop the upper index n) in a Euclidean space $\mathbb{R}^n$ equal to 1 on an orienting orthobasis (i.e., the basis that has been selected to define the orientation). (Put less formally, Ω is the volume of an n-dimensional unit cube with properly numbered edge vectors.) *Show* that for an orienting orthobasis $e_1,\ldots$ in $\mathbb{R}^n$, $\Omega = e_1^* \wedge \ldots \wedge e_n^*$, where $e_1^*, \ldots$ is the dual basis in $\mathbb{R}^{n*}$, so (by virtue of section P8.9***), for $x_1,\ldots, x_n \in \mathbb{R}^n$, $\Omega(x_1,\ldots, x_n) = \det (\alpha_{ij})_{i,j\,=\,1,\ldots,n}$, where α_{ij} are coefficients with respect to this basis, $x_i = \sum_{j=1}^{n} \alpha_{ij} e_j$ $(i = 1,\ldots, n)$.

*P8.14****

Consider an oriented Euclidean space $\mathbb{R}^n$. Let the isomorphism $\mathbf{R}^n \xrightarrow{D} \mathbf{R}^{n*}$ be connected to the scalar product, as in section P8.11***, and $\wedge \mathbf{R}^n \xrightarrow{\wedge D} \wedge \mathbf{R}^{n*}$ its extension, as in section P8.12***. Furthermore, let Ω be the oriented Euclidean volume and $\wedge \mathbf{R}^n \xrightarrow{I_\Omega} \wedge \mathbf{R}^{n*}$ the corresponding isomorphism, as in section P8.10***. For purposes of clarity, in what follows we will keep to the notations D, I instead of $\wedge D$, I_Ω, respectively. The **Hodge star operator** (or just "**star**" for short) is a common name of the two linear automorphisms

$$* := I^{-1} \circ D : \wedge \mathbf{R}^n \to \wedge \mathbf{R}^n, \qquad *: = D \circ I^{-1} : \wedge \mathbf{R}^{n*} \to \wedge \mathbf{R}^{n*}.$$

[6] Readers preferring more developed terminology may say that orienting $\mathbb{R}^n$ is choosing one of two connected components of the linear group $GL(n,\mathbb{R})$ (see section P6.8*** in the "Polar and Singular Value Decomposition Theorems" problem group above).

By this definition, D and I conjugate the star operator: $* = D \circ * \circ D^{-1} = I \circ * \circ I^{-1}$. The star operator maps homogeneous elements to homogeneous elements of complementary degree: $*(\wedge^k \mathbb{R}^n) = \wedge^{n-k} \mathbb{R}^n$, $*(\wedge^k \mathbb{R}^{n*}) = \wedge^{n-k} \mathbb{R}^{n*}$ ($k = 0,\ldots, n$). *Prove* this by *translating* the defining formulas into the following form: $DX = \imath_{*X}\,\Omega = *\imath_X\,\Omega$ ($X \in \wedge\mathbb{R}^n$), where $\imath_X\Omega = \mathrm{I}_\Omega(X)$ (see section P8.10***).

*P8.15****

The claim of section P8.14*** can be refined as follows: for a k-dimensional vector subspace Π, the star operator $\wedge \mathbf{R}^n \xrightarrow{*} \wedge\mathbf{R}^n$ maps its exterior k-power $\wedge^k\Pi$ onto the exterior $n-k$-power of the orthogonal complement to Π, $*(\wedge^k\Pi) = \wedge^{n-k}\Pi^\perp$. Adding even more details, for an orienting orthobasis $e_1,\ldots$ in $\mathbb{R}^n$, an ordered subset $\mathrm{S} \subseteq \{1,\ldots,n\}$ and its ordered complement $S' = \{1,\ldots, n\}\backslash S$,

$$* \left(\bigwedge_{i\in S} e_i \right) = \mathrm{sign}(\sigma) \cdot \bigwedge_{i\in S'} e_i,$$

where σ is a permutation of $\{1,\ldots,n\}$ such that the elements of S appear after the elements of S'. Prove it**.** In particular, $* \bigwedge_{i=1}^{n} e_i = 1$, $* 1 = \bigwedge_{i=1}^{n} e_i$. *Show* that similar dual claims for $\wedge \mathbf{R}^{n*}$ also hold. (Therefore, $* \Omega = 1, *1 = \Omega$). Hence, the iterated star operator is $*^2 = (-1)^{k(n-k)} E_{\wedge^k}$ on $\wedge^k$. *Prove* that $\Omega(x_1,\ldots, x_n) = *\,(x_1 \wedge \ldots \wedge x_n)$ ($x_1,\ldots, x_n \in \mathbb{R}^n$).

*P8.16****

Consider the isomorphism $\wedge \mathbf{R}^n \xrightarrow{D} \wedge\mathbf{R}^{n*}$ for an oriented Euclidean space $\mathbb{R}^n$, as in section P8.14***, and the connected scalar product in $\wedge\mathbb{R}^n$, as in section P8.12*** (extending the scalar product in $\mathbb{R}^n$). Denoting it by $\langle\rangle$, *establish* the formula

$$* \langle X, Y\rangle = (*X) \wedge Y = (*Y) \wedge X \qquad \big(X, Y \in \wedge^k\mathbf{R}^n,\ \ k = 0,\ldots,n\big).$$

Also, *prove* the orthogonality of the star operator with respect to $\langle\rangle$:

$$\langle *X,\ *Y\rangle = \langle X, Y\rangle \qquad \big(X, Y \in \wedge^k\mathbf{R}^n,\ \ k = 0,\ldots,n\big).$$

P8.17[***]

The methods of elementary geometry make it possible to define angles between straight lines and angles between hyperplanes in Euclidean spaces. Let us now define angles between k-dimensional planes of a Euclidean space $\mathbb{R}^n$. For this, consider the Euclidean space $\wedge^k\mathbb{R}^n$, with the scalar product defined in section P8.12[***]. Also, consider the embedding $G(n,k) \to P(\wedge^k \mathbb{R}^n)$ defined in section P8.7[***]. Thus, the angle between elements of $G(n,k)$ may be defined as an angle between the straight lines that are images of these elements in $P(\wedge^k \mathbb{R}^n)$. Obviously, this definition is equivalent to the elementary-geometrical definition for $k = 1$ (when k-dimensional planes are straight lines). *Establish* a similar equivalence for $k = n - 1$ (the elementary-geometrical definition was discussed in section P8.0[**]).

Verify that the definition of angle $\alpha = (\Pi,\Pi')$ between k-dimensional planes Π, Π' means, geometrically speaking, that for an orthobasis $e_1,\ldots,e_k$ of Π and the orthogonal projection P of the space onto Π', $|\cos\alpha|$ is the k-dimensional "nonoriented" area of the k-dimensional $Pe_1,\ldots,Pe_k$-sided parallelogram in Π' (that is, is the k-dimensional area projection factor).

Carry out two or three numerical exercises as follows: given two k-dimensional planes in $\mathbb{R}^n$ (each defined by a system of $n - k$ linear equations; say, take $k = 2$, $n = 4$), calculate the angle. For example, find the angle between the planes $\{(x,y,z,t)\in\mathbb{R}^4: x + y + z + t = 0,\ x - z = 0\}$ and $\{(x,y,z,t)\in\mathbb{R}^4: x - y + z - t = 0,\ y - t = 0\}$ (90° is an answer).

P8.18[***]

Let $\mathbf{R}^n = \Pi^k \oplus \Lambda^{n-k}$ be an orthogonal decomposition of an oriented Euclidean space, with the summands oriented such that the concatenation of orienting bases of Π^k first and of Λ^{n-k} second produces an orienting basis of $\mathbb{R}^n$. *Prove* that if the restriction of a k-form $\omega^k\in(\wedge^k \mathbb{R}^n)^*$ to Π^k (more precisely, to the subspace $\wedge^k \Pi^k \subseteq \wedge^k \mathbb{R}^n$) coincides with the oriented Euclidean (k-dimensional) area in Π^k, then the restriction of an $n - k$-form $*\omega^k\in(\wedge^{n-k} \mathbb{R}^n)^*$ to Λ^{n-k} (that is, to the subspace $\wedge^{n-k} \Lambda^{n-k} \subseteq \wedge^{n-k} \mathbb{R}^n$) coincides with the oriented Euclidean ($n - k$-dimensional) area in Λ^{n-k}, multiplied by $(-1)^{k(n-k)}$.

P8.19[***]

Now readers have all the necessary tools for proving Theorem 3 from section P8.0[**]. Thus, *show* that the complementary minors M, M' of an orthogonal matrix A are equal by absolute values and $\frac{\text{sign } M}{\text{sign } M'} = \det A \cdot \text{sign}\begin{pmatrix} i_1 & \cdots & i_n \\ j_1 & \cdots & j_n \end{pmatrix}$, where $i_1 < \ldots < i_k$, $j_1 < \ldots < j_k$ ($i_{k+1} < \ldots < i_n$, $j_{k+1} < \ldots < j_n$) are indices of rows and columns of M (resp. M').

P8.20[***]

Let us define a **vector product in an oriented Euclidean space** $\mathbb{R}^n$ as a multilinear skew-symmetric map $\prod\left(\wedge^{k_i}\mathbf{R}^n\right) \xrightarrow{[\,]} \wedge^{n-\sum k_i}\mathbf{R}^n$ by the formula

$$\left[x^{k_1},\ldots,x^{k_m}\right]:=*\bigwedge_{i=1}^{m}x^{k_i}.$$

Show that in a special case of $n=3$, $m=2$, $k_1=k_2=1$, this formula is equivalent to the usual definition of a vector product in $\mathbb{R}^3$. To do so, prove for $x_2,\ldots,x_n\in\mathbb{R}^n$, and $x_1:=[x_2,\ldots,x_n]$, that:

(i) If $x_{i1},\ldots,x_{in}$ are coordinates of x_i $(i=2,\ldots,n)$ with respect to an orienting orthobasis $e_1,\ldots,e_n$, then the coordinates of x_1 are given by decomposing the symbolic determinant $x_1=\det\begin{pmatrix}e_1 & \cdots & e_n\\ x_{21} & \cdots & x_{2n}\\ \vdots & & \vdots\\ x_{n1} & \cdots & x_{nn}\end{pmatrix}$ with respect to its upper row.

(ii) x_1 is orthogonal to $x_2,\ldots,x_n$, and $|x_1|$ is the (Euclidean $n-1$-dimensional) area of an $x_2,\ldots,x_n$-sided ($n-1$-dimensional) parallelogram. If $x_2,\ldots,x_n$ are linearly independent (otherwise $x_1=0$), the direction of x_1 is such that $x_1,\ldots,x_n$ is an orienting basis of $\mathbb{R}^n$.

Prove for $x_1,\ldots,x_n\in\mathbb{R}^n$ that the oriented Euclidean volume of the $x_1,\ldots,x_n$-edged (n-dimensional) parallelepiped is the **mixed product** $\Omega(x_1,\ldots,x_n)=\langle x_1,[x_2,\ldots,x_n]\rangle$.

P8.21[***]

Prove for the **double vector product** in $\mathbb{R}^n$ equalities similar to those known for $\mathbb{R}^3$ (the second of which is known as the **Jacobi identity for Lie algebras**):

$$[X,[Y,Z]]=\langle X,Z\rangle Y-\langle X,Y\rangle Z,$$
$$[X,[Y,Z]]+[Z,[X,Y]]+[Y,[Z,X]]=0\quad(X,Y,Z\in\mathbf{R}^n)\ .$$

(***Warning***: for $n\neq 0,1,3$, $\mathbb{R}^n$ is not a Lie algebra with respect to the vector product as $X, Y\in\mathbb{R}^n \Rightarrow [X,Y]\in\mathbb{R}^n$ only when $n=0,1,3$.)

Hint

H8.1

For the sake of simplicity, let $k = 2$. If $x \wedge y + y \wedge x = 0, \forall x,y$, then setting $x = y$ yields $2x \wedge x = 0$. Therefore, $x \wedge x = 0$. (Over what fields does the implication not apply to vector spaces?) Conversely, if $x \wedge x = 0, \forall x$, then

$$0 = (x + y) \wedge (x + y) = x \wedge x + y \wedge y + x \wedge y + y \wedge x \Rightarrow x \wedge y + y \wedge x = 0.$$

Similar arguments are used for $k > 2$. Turning to the basic property of the map $\wedge^k$, denote, respectively, $\mathcal{P}$, $\mathcal{Q}$, as an infinite-dimensional vector space spanned on a basis indexed by the points of $(\mathbb{R}^n)^k$, and its subspace generated by $\left(\ldots, \underset{i}{x}, \ldots, \underset{j}{x}, \ldots\right)$ and $(\ldots, \underset{m}{\alpha x + \beta y}, \ldots) - \alpha(\ldots, \underset{m}{x}, \ldots) - \beta(\ldots, \underset{m}{y}, \ldots)$ $(x,y \in \mathbb{R}^n$, $\alpha,\beta \in \mathbb{R}$, $1 \leq m \leq k$, $1 \leq i < j \leq k)$. Let "$\subset$" be the inclusion map $(\mathbf{R}^n)^k \xrightarrow{\subset} \mathcal{P}$. Consider a multilinear skew-symmetric map F: $(\mathbb{R}^n)^k \to L$. A map F is uniquely extended to a linear map $\mathcal{F}: \mathcal{P} \to L$. (Why?) Obviously, $F = \mathcal{F} \circ \subset$. For a multilinear skew-symmetric F we have an inclusion $\mathcal{Q} \subseteq \ker \mathcal{F}$. (Why?) Hence, $\mathcal{F}$ factors through the quotient map $\pi : \mathcal{P} \to \mathcal{P}/\mathcal{Q} = \wedge^k \mathbb{R}^n$: $\mathcal{F} = \varphi \circ \pi$. (Why?) Thus, $\pi \circ \subset = \wedge^k$, and φ is a desired linear map $\wedge^k \mathbb{R}^n \to L$ as $F = \varphi \circ \wedge^k$. The uniqueness of φ is evident. (Why?)

H8.2

Decomposing $(\sum a_{1j}x_j) \wedge \ldots \wedge (\sum a_{kj}x_j)$ by multilinearity and skew symmetry and taking into account that $\det A = \det {}^tA$ yields the required solution.

H8.3

The invertibility of $f \mapsto f \circ \wedge^k$, which is a map on $(\wedge^k \mathbb{R}^n)^*$ to the space of k-forms in $\mathbb{R}^n$, follows from the basic property of $\wedge^k$ discussed in section P8.1***. Indeed, by this property, an exterior k-form is represented by a composition $f \circ \wedge^k$ with a unique $f \in (\wedge^k \mathbb{R}^n)^*$. (Why?)

H8.4

Applying section P8.2*** to the decomposition of the generators $x_1 \wedge \ldots \wedge x_k$ confirms that the elements $e_{i_1} \wedge \ldots \wedge e_{i_k}$ $(1 \leq i_1 < \ldots < i_k \leq n)$ jointly form a sufficient generating set, so we must verify their linear independence. First, establish this independence for $k = n$, that is, prove that $\wedge^n \mathbb{R}^n \neq \{0\}$; to do this, use section P8.3*** and the existence of nonzero determinants. Next, given a basis $e_1,\ldots,e_n$ of $\mathbb{R}^n$, arrange, using the basic property of $\wedge^k$ from section P8.1***, a linear map $\wedge^k \mathbb{R}^n \to \wedge^n \mathbb{R}^n$ such that $e_1 \wedge \ldots \wedge e_k$ is mapped onto $e_1 \wedge \ldots \wedge e_n$ and the rest of $e_{i_1} \wedge \ldots \wedge e_{i_k}$ are mapped to zero. This proves the linear independence of all of them. (Why?)

H8.5

If x_i are linearly dependent, then, without loss of generality, $x_k = \sum_{i=1}^{k-1} \alpha_i x_i$, so

$$x_1 \wedge \ldots \wedge x_k = x_1 \wedge \ldots \wedge x_{k-1} \wedge \sum_{i=1}^{k-1} \alpha_i x_i = \sum_{i=1}^{k-1} \alpha_i \cdot x_1 \wedge \ldots \wedge x_{k-1} \wedge x_i = \sum 0 = 0.$$

The remaining claims in section P8.5*** follow directly from section P8.4***.

H8.6

The claims in section P8.5*** show that applying the Gram orthogonalization to the set of vectors $x_1,\ldots,x_k$ leaves the linear generator $x_1 \wedge \ldots \wedge x_k$ unchanged, to within proportionality. Hence, the dimension of the variety consisting of these generators exceeds by 1 the dimension of the variety of the sets of k orthonormal vectors considered up to linear equivalence. Using this find the dimension of the generators' variety and compare your result to $\dim \wedge^k \mathbf{R}^n = \binom{n}{k}$.

H8.7

The embedding $G(n,k) \to P(\wedge^k \mathbb{R}^n)$ identifying k-dimensional subspaces of $\mathbb{R}^n$ with straight lines in $\wedge^k \mathbb{R}^n$ is surjective only for $k = 0, 1, n - 1, n$.

H8.8

To extend the equality $(x_1 \wedge \ldots \wedge x_k) \wedge (x_{k+1} \wedge \ldots \wedge x_{k+l}) = x_1 \wedge \ldots \wedge x_{k+l}$ from given basis elements $e_1,\ldots$ to all $x_1,\ldots$, set $x_i = \sum a_{ij} e_j$ and decompose both sides by the multilinearity of $\wedge^k$, $\wedge^l$, and $\wedge$ using the **Laplace** formula for the **decomposition** of a determinant with respect to a set of its rows (or columns). Next, the skew-commutativity formula holds for basis elements and is multilinear, so it holds throughout $\wedge\mathbb{R}^n$.

H8.9

The desired formula for the exterior product of an arbitrary number of cofactors is obtained by induction from the case of two cofactors and looks as follows:

$$\wedge\,\omega^{k_i}\left(x_1,\ldots,x_{\sum k_i}\right) = \sum_{\sigma} \operatorname{sign}(\sigma) \cdot \prod \omega^{k_i}\left(x_{\sigma\left(\sum_j k_j+1\right)},\ldots,x_{\sigma\left(\sum_{j\le i} k_j\right)}\right),$$

where the sum is taken over all permutations of the set $1,\ldots,\sum k_i$ such that $\sigma\left(\sum_j k_j + 1\right) < \ldots < \sigma\left(\sum_{j\le i} k_j\right)$, $\forall i$. To prove this formula for two cofactors, first establish the right-hand side as an exterior $k + l$-form, that is, a $k + l$-linear skew-symmetric function on $(\mathbb{R}^n)^{k+l}$. For this, apply similar arguments as used to prove the **Laplace decomposition** formula for determinants (see section E8.8). Next, verify that the right-hand side is bilinear with respect to the forms ω^k, ω^l and coincides with the source definition for basis forms $\omega^k = e^*_{i_1} \wedge \ldots \wedge e^*_{i_k}$, $\omega^l = e^*_{j_1} \wedge \ldots \wedge e^*_{j_l}$, which will complete the proof. Next, we can establish the determinant formula for $x_1^* \wedge \ldots \wedge x_k^*(x_1,\ldots,x_k)$, verifying that this formula is k-linear and skew-symmetric with respect to both $x_1,\ldots,x_k$ and $x_1^*,\ldots,x_k^*$, and that it is valid for the basis elements $e_{i_1},\ldots,e_{i_k}$, $e^*_{j_1},\ldots,e^*_{j_k}$ $(i_1 < \ldots < i_k, j_1 < \ldots < j_k)$. (What follows from that?) Finally, the skew-commutativity formula holds for basis elements and is multilinear, so it holds throughout $(\wedge\mathbb{R}^n)^*$; alternatively, it can be derived from the skew commutativity of the exterior product in $\wedge\mathbb{R}^n$. (Complete the proof using this guidance.)

H8.10

The linear operator I_Ω: $\wedge\mathbb{R}^n \to (\wedge\mathbb{R}^n)^*$ is an isomorphism if $\Omega \neq 0$ and otherwise is the zero operator.

H8.11

Consider bilinear forms on a pair of vector spaces $X \times Y \xrightarrow{\Phi} field$ and the connected linear maps $D'_\Phi : \begin{cases} X & \to & Y^* \\ x & \mapsto & \Phi(x,.) \end{cases}$, $D''_\Phi : \begin{cases} Y & \to & X^* \\ y & \mapsto & \Phi(.,y) \end{cases}$. A subspace $Y^\perp := \{x \in X: \Phi(x,y) = 0, \forall y \in Y\}$ ($X^\perp := \{y \in Y: \Phi(x,y) = 0, \forall x \in X\}$) is referred to as the **left** (resp. **right**) **kernel** of the form Φ. The form **degenerates from the left** (resp. **right**) when $Y^\perp \neq 0$ (resp. $X^\perp \neq 0$). If the form does not degenerate from any side, then it is referred to as **nondegenerate**. Verify that $\ker D'_\Phi = Y^\perp$, $\ker D''_\Phi = X^\perp$. We consider only the finite-dimensional spaces. Prove the following claims:

1. $\dim X = \dim Y$ *if* Φ *is nondegenerate*.
2. *(**A generalization**:)* $\operatorname{codim} Y^\perp (= \dim X - \dim Y^\perp) = \operatorname{codim} X^\perp (= \dim Y - \dim X^\perp)$.

From the second claim, $\dim X = \dim Y \Rightarrow \dim Y^\perp = \dim X^\perp$, in particular, Φ is nondegenerate if one of D'_Φ, D''_Φ is, and only if both of them are isomorhisms. **QED**.

The same technique allows us to prove other statements that readers are familiar with! For an **orthogonal complement** to a subspace $U \subseteq X$, $U^\perp := \{y \in Y: \Phi(x,y) = 0, \forall x \in U\}$ [resp. to a subspace $V \subseteq Y$, $V^\perp := \{x \in X: \Phi(x,y) = 0, \forall y \in V\}$] prove that

3. $U \cap Y^\perp = 0 \Rightarrow \dim U^\perp + \dim U = \dim Y$, $V \cap X^\perp = 0 \Rightarrow \dim V^\perp + \dim V = \dim X$*; in particular, an orthogonal complement always has a complementary dimension when* Φ *is nondegenerate*. (***Warning***: a subspace may have a nonzero intersection with its orthogonal complement if Φ is not a positive definite form!)

Also, considering a common identification of $m \times n$ matrices with bilinear forms $X^m \times Y^n \to$ *field*, establish the equality of the matrices' row and column ranks.

Next, let G be a Gram matrix with respect to the basis $a_1, \ldots$, $G = \left(\langle a_i, a_j \rangle\right)_{i,j=1,\ldots,m}$. A scalar product $\langle\rangle^*$ will have the same Gram matrix with respect to the basis $Da_1, \ldots$ ($Da_i = \langle a_i, . \rangle$). For the coefficients of $a_1^*, \ldots$ in the basis $Da_1, \ldots$, $a_i^* = \sum_j c_{ij} Da_j$,, we will have equations $\delta_{ik} = \sum_j c_{ij} \langle a_j, a_k \rangle$, or, in matrix form, $E = C \circ G$, so $C = G^{-1}$. Therefore, a Gram matrix with respect to the basis $a_1^*, \ldots$ is ${}^tG^{-1} \circ G \circ G^{-1} = G^{-1}$. (Why?)

The rest of section P8.11*** is more or less straightforward.

H8.12

Define $\wedge D(1) := 1$ and, for a fixed (not necessarily orthonormal) basis $e_1,\ldots$ in $\mathbb{R}^n$, $\wedge D\left(\bigwedge_{j=1}^{k} e_{i_j}\right) := \bigwedge_{j=1}^{k} De_{i_j}$ $(k = 1,\ldots, n,\ 1 \le i_1 < \ldots < i_k \le n)$, and extend it by multilinearity to all of $\wedge\mathbb{R}^n$. The permutability of $\wedge D$ and $\wedge$ can be deduced using section P8.2*** from the linearity of the first and multilinearity of the second. (Provide the details!) A bilinear form on $\wedge\mathbb{R}^n$, to which $\wedge D$ is connected, is a scalar product by virtue of section P8.11***. Considering an orthobasis $e_1,\ldots$ such that $e_1,\ldots,e_k$ belong to the (k-dimensional) plane spanned on $x_1,\ldots,x_k$, so that $x_i = \sum_{j=1}^{k} \alpha_{ij} e_j$ $(i = 1,\ldots, k)$, and using section P8.2*** yields

$$|x_1 \wedge \ldots \wedge x_k| = \left|\det\left(\alpha_{ij}\right)_{i,j=1,\ldots,k}\right| \cdot |e_1 \wedge \ldots \wedge e_k| = \left|\det\left(\alpha_{ij}\right)_{i,j=1,\ldots,k}\right|$$
$$= \left|e_1^* \wedge \ldots \wedge e_k^*(x_1, \ldots, x_k)\right|,$$

which is the area of the $x_1,\ldots,x_k$ -sided parallelogram.

Proof of the claim about the sum of squared areas of orthogonal projections for arbitrary squarable planar sets. Consider a planar set that is a union of N k-dimensional parallelograms intersecting with each other at most by their boundaries; let a_v, $v = 1,\ldots, N$ be their (k-dimensional) areas and b_{vi} the (k-dimensional) areas of their orthogonal projections onto the k-dimensional coordinate planes using a fixed orthogonal coordinate system $\left(i = 1, \ldots, \binom{n}{k}\right)$; we will prove that $\left(\sum_v a_v\right)^2 = \sum_i \left(\sum_v b_{vi}\right)^2$. Since $a_v^2 = \sum_i b_{vi}^2$, $\forall v$, it is sufficient to verify that $a_\mu a_\nu = \sum_i b_{\mu i} b_{\nu i}$, or $\sqrt{\sum_i b_{\mu i}^2}\sqrt{\sum_i b_{\nu i}^2} = \sum_i b_{\mu i} b_{\nu i}$, $\forall \mu,\nu$. But since the *area projection factor* (section P8.0**) is the same for all k-dimensional parallelograms in the same k-dimensional plane, b_μ and b_ν are proportional vectors, $b_\mu/a_\mu = b_\nu/a_\nu$, and so these equations are satisfied (this is simply the case of equality in the CSB inequality). We leave it to the reader to work out the details and finalize the proof.

H8.13

Follows directly from section P8.9***.

H8.15

The first formula follows from the permutability of the duality D ($= \wedge D$), with the exterior product (section P8.12***), and the formula $DX = \imath_{*X}\Omega$ from section P8.14***. The second formula (concerning $*^2$) is multilinear and holds for an orthobasis in $\wedge^k \mathbb{R}^n$ (by virtue of the first formula), so it holds for all of $\wedge^k \mathbb{R}^n$. (Why?) (Because of the permutability of D with * and $\wedge$, similar statements hold for the star operator on $\wedge\mathbb{R}^{n^*}$.) The third formula follows from the first, using the claims of sections P8.13*** and P8.2***, as consideration of decompositions $x_i = \sum_{j=1}^{n} \alpha_{ij} e_j (i = 1,\ldots, n)$ gives

$$\begin{aligned}\Omega(x_1,\ldots,x_n) &= \det\left(\alpha_{ij}\right)_{i,j=1,\ldots,n} = \det\left(\alpha_{ij}\right)_{i,j=1,\ldots,n} \cdot 1 = \det\left(\alpha_{ij}\right)_{i,j=1,\ldots,n} \cdot * \bigwedge_{i=1}^{n} e_i \\ &= *\left[\det\left(\alpha_{ij}\right)_{i,j=1,\ldots,n} \cdot \bigwedge_{i=1}^{n} e_i\right] = * \bigwedge_{i=1}^{n} x_i \ .\end{aligned}$$

We can attain the same solution while avoiding computations by using determinants:

$$\begin{aligned}* (x_1 \wedge \ldots \wedge x_n) &= D * (x_1 \wedge \ldots \wedge x_n) = *D(x_1 \wedge \ldots \wedge x_n) = ** \ \imath_{x_1 \wedge \ldots \wedge x_n}\Omega \\ &= \Omega(x_1,\ldots,x_n).\end{aligned}$$

H8.16

We have

$$* \langle X, Y\rangle = *[DX(Y)] = *(\imath_Y DX) = *(\imath_Y \imath_{*X}\Omega) = *\left[\imath_{(*X)\wedge Y}\Omega\right] = (*X) \wedge Y$$

[by symmetry, this also equals $(*Y) \wedge X$]. The orthogonality of * follows from the equalities

$$\begin{aligned}*\langle X, Y\rangle &= * \langle Y, X\rangle = (*Y) \wedge X, \\ *\langle *X, *Y\rangle &= (** X) \wedge Y = (-1)^{k(n-k)} X \wedge *Y = (*Y) \wedge X.\end{aligned}$$

(Because of the permutability of D with *, $\wedge$ and $\langle\rangle$, similar statements hold for the star operator on $\wedge\mathbb{R}^{n^*}$.)

H8.17

The equivalence, for $k = n - 1$, of the definitions of angles between hyperplanes, namely, the elementary-geometrical definition, and the definition from section P8.17*** follows from (1) a similar equivalence, for $k = 1$, (2) the orthogonality of the star operator (section P8.16***), and (3) elementary-geometric Theorem 1 from section P8.0**. Provide the details.

H8.18

Let $e_1,\ldots$ be an orienting orthobasis such that $e_1,\ldots,e_k \in \Pi^k$. We have $\omega^k = e_1^* \wedge \ldots \wedge e_k^* + \ldots$ (the terms proportional to $e_{i_1}^* \wedge \ldots \wedge e_{i_k}^*$ with $\{i_1,\ldots,i_k\} \neq \{1,\ldots,k\}$ are not written). (Why?) According to section P8.15***, $* \omega^k = (-1)^{k(n-k)} e_{k+1}^* \wedge \ldots \wedge e_n^* + \ldots$ (the terms proportional to $e_{i_{k+1}}^* \wedge \ldots \wedge e_{i_n}^*$ with $\{i_{k+1},\ldots,i_n\} \neq \{k+1,\ldots,n\}$ are not written). This proves the required statement. (Why?)

H8.19

Let $e_1,\ldots$ be an orienting orthobasis in $\mathbb{R}^n$. An orthogonal operator A has, with respect to it, an orthogonal matrix $\varsigma = (\alpha_{ij})_{i,j=0,\ldots,n}$. $Ae_1,\ldots$ is also an orthobasis (orienting or not, depending on the sign $\det A$). Considering the dual basis $(Ae_1)^*,\ldots$, the matrix elements are $\alpha_{ij} = \langle Ae_j,\ e_i\rangle = \left(Ae_j\right)^*(e_i)$. Apply P8.18*** to an exterior k-form $\omega = \left(Ae_{j_1}\right)^* \wedge \ldots \wedge \left(Ae_{j_k}\right)^*$ and a k-dimensional plane Π spanned on $e_{i_1},\ldots,e_{i_k}$.

H8.20

Complementary minors for the first row are the coefficients of decomposition of $x_2 \wedge \ldots \wedge x_n$ with respect to the orthobasis $\{e_{i_1} \wedge \ldots \wedge e_{i_{n-1}} :\ 1 \le i_1 < \ldots < i_{n-1} \le n\}$ in $\wedge^{n-1}\mathbb{R}^n$. (Why?) Hence, the coefficients of $x_1 = * \bigwedge_{i=2}^{n} x_i$ with respect to the basis $e_1,\ldots,e_n$ are cofactors of those elements (why?), which completes the proof of claim (i). The orthogonality of $[x_2,\ldots,x_n]$ to the hyperplane Π^{n-1} spanned on $x_2,\ldots,x_n$ follows from section P8.15***. The vector product's modulus and the mixed product can be found by inconvenient elementary computations with determinants. On the other hand, we have aimed at equipping the reader with a powerful technique of skew symmetry and making it accessible. A theory is a good one if it requires

tiresome computations when established, but never while being applied! Therefore, readers are invited to use the tools provided with the claims in sections P8.14***, P8.15, and P8.16***.

H8.21

Since both sides of the first equality being proved are bilinear and skew-symmetric with respect to Y, Z, we may restrict ourselves to the orthogonal unit vectors Y, Z. Consider an orienting orthobasis $e_1,\ldots$ such that $e_{n-1} = Y$, $e_n = Z$. Because of linearity, with respect to X as well, we may consider X an element of the same basis, $X = e_i$. Now evaluate $[X, [Y,Z]]$. The second formula (the Jacobi identity) is a formal implication of the first one. (Provide the details.)

Explanation

E8.1

Answering the first question "why?" from section H8.1, prove, for a vector space X (of any, perhaps infinite, dimension) and a subset $T = \{x_1,\ldots\} \subseteq X$, that an arbitrary map to a vector space $T \xrightarrow{f} Y$ is uniquely extended to a linear map $X \to Y$ if and only if T is a basis (a linearly independent, linearly generating X set). The reader may argue as follows: As is known, a subset $\Gamma \subseteq X \times Y$ containing pairs (x,y) for $\forall x \in X$ is a graph of a map $X \to Y$ if and only if $(x,y), (x,y') \in \Gamma \Rightarrow y = y'$, $\forall x \in X$, $\forall y,y' \in Y$. Let T be a basis of X. Consider the subset Γ of the Cartesian product $X \times Y$ consisting of the elements $(\sum \alpha_i x_i, \sum \alpha_i f(x_i))$ for all linear combinations $x = \sum \alpha_i x_i \in X$. Show that Γ is the graph of a map $X \to Y$ if and only if $(0, y) \in \Gamma \Rightarrow y = 0$, $\forall y$, and in this case, the map is linear. But since $y = \sum \alpha_i f(x_i) \neq 0$, we have $\sum \alpha_i x_i \neq 0$ only if some coefficients $\alpha_i \neq 0$. Thus, f possesses a unique (why?) linear extension $X \to Y$. Now let T be linearly dependent, and let $\sum_{i=1}^{N} \alpha_i x_i = 0$ be a nontrivial linear relation among its elements. Show that for a vector space Y of dimension N, with a basis $\{y_1,\ldots y_N\}$, a map $f(x_i) := \begin{cases} y_i, & 1 \le i \le N \\ 0 & \text{otherwise} \end{cases}$ is not extended to a linear map $X \to Y$. If T is linearly independent but has a span $X' \neq X$, then T is extended to a basis of X, but not uniquely, so a map $T \xrightarrow{f} Y$ is extended to a linear map $X \to Y$ (by the first part of this proof), but not uniquely. This completes the proof.

E8.4

We can prove that $\wedge^n \mathbb{R}^n \neq \{0\}$ as follows. $\wedge^n \mathbb{R}^n = \{0\} \Leftrightarrow (\wedge^n \mathbb{R}^n)^* = \{0\}$. According to section P8.3***, all exterior n-forms in $\mathbb{R}^n$ would be zero, which does not hold as there exist nonzero determinants. (Which one, for example?) Next, a required linear map $\wedge^k \mathbb{R}^n \to \wedge^n \mathbb{R}^n$ is arranged, given a basis $e_1,\ldots,e_n$ in $\mathbb{R}^n$, in the following manner. Consider embedding $\begin{cases} (\mathbf{R}^n)^k & \stackrel{\iota}{\to} & (\mathbf{R}^n)^n \\ (x_1,\ldots,x_k) & \mapsto & (x_1,\ldots,x_k,e_{k+1},\ldots,e_n) \end{cases}$. . Taking a composition $(\mathbf{R}^n)^k \stackrel{\wedge^n\circ\iota}{\to} \wedge^n\mathbf{R}^n$ yields a k-linear skew-symmetric map, so, by the basic property of $\wedge^k$ from section P8.1***, there is a unique linear map $\wedge^k\mathbf{R}^n \stackrel{\Phi}{\to} \wedge^n\mathbf{R}^n$ such that $\wedge^n \circ \iota = \Phi \circ \wedge^k$. We have

$$\begin{aligned}\Phi(e_{i_1} \wedge \ldots \wedge e_{i_k}) &= e_{i_1} \wedge \ldots \wedge e_{i_k} \wedge e_{k+1} \wedge \ldots \wedge e_n \\ &= \begin{cases} e_1 \wedge \ldots \wedge e_n & \text{if } \{i_1,\ldots,i_k\} = \{1,\ldots,k\}, \\ 0 & \text{otherwise,} \end{cases}\end{aligned}$$

which completes the proof.

E8.6

We do not expect the reader to look for the dimensions of orthogonal groups using the theory of Lie groups and algebras. We suggest calculating the dimension of the variety of the sets of k orthonormal vectors considered up to linear equivalence using a much easier technique. A unit vector $e_1 \in \mathbb{R}^n$ has $n-1$ linearly independent infinitesimal variations. (Why?) In turn, a unit vector e_2, orthogonal to e_1, has $n-2$ linearly independent infinitesimal variations, and so on. Therefore, a set of k orthonormal vectors has $(n-1)+\ldots+(n-k)$ linearly independent infinitesimal variations, which equals the dimension of the variety of these sets. Considering them up to linear equivalence decreases the dimension by the dimension of the variety of orthonormal bases of a k-dimensional plane, which, by the same argument, is equal to $(k-1)+\ldots+1$. Finally, the dimension of the generators' variety is

$$[(n-1)+\ldots+(n-k)] - [(k-1)+\ldots+1] + 1 = k(n-k)+1.$$

Comparing this to dim $\wedge^n \mathbb{R}^n$, verify that $\binom{n}{k} = k(n-k)+1$ for $k = 0,\ 1,\ n-1,\ n$, and make a rough estimate as $\binom{n}{k} \gg k(n-k)$ when $1 \ll k \ll n$.

The complete answer is that $\binom{n}{k} > (n-k)+1$, except for the equalities for $k = 0$, 1, $n-1$, n. Prove this elementary number-theoretic statement, and provide a geometric explanation of the equalities for $k = 1$, $n-1$. (Once you encounter the difficulties in this explanation for $k = n-1$, return to it after solving Problem P8.15*** below.)

E8.8

The **Laplace decomposition** formula for determinants states that for any $k = 1,\ldots,$ $n-1$, selected row indices $1 \le i_1 < \ldots < i_k \le n$,

$$\det(a_{ij})_{i,j=1,\ldots,n} = \sum_{1 \le j_1 < \ldots < j_k \le n} \operatorname{sign}\begin{pmatrix} i_1 & \cdots & i_n \\ j_1 & \cdots & j_n \end{pmatrix} \det\begin{pmatrix} a_{i_1 j_1} & \cdots & a_{i_1 j_k} \\ \vdots & & \vdots \\ a_{i_k j_1} & \cdots & a_{i_k j_k} \end{pmatrix} \det\begin{pmatrix} a_{i_{k+1} j_{k+1}} & \cdots & a_{i_{k+1} j_n} \\ \vdots & & \vdots \\ a_{i_n j_{k+1}} & \cdots & a_{i_n j_n} \end{pmatrix}.$$

In this formula, $i_{k+1} < \ldots < i_n$ and $(i_1,\ldots, i_n) = \{1,\ldots, n\}$, and the same convention is retained with respect to $j_1,\ldots,j_n$. A similar formula holds for selected column indices. Readers are likely familiar with the Laplace formula in the special case where $k = 1$ (or, equivalently, where $k = n-1$). For arbitrary $1 \le k \le n-1$ it can be proved as follows. The right-hand side is n-linear with respect to the columns $\begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}$ $(j = 1,\ldots,n)$. It is also skew-symmetric. Indeed, let $a_{ij'} = a_{ij''}$, $\forall i = 1,\ldots,n$ $(j' < j'')$. The terms on the right-hand side are equal to zero if both j' and j'' appear in one of the two lists $j_1,\ldots,j_k$, $j_{k+1},\ldots,j_n$, and they are mutually annihilable if j', j'' appear in different lists. (Why?) Thus, the right-hand side is an exterior n-form, so it is a determinant. Which determinant? Readers can specify it using a fixed set of n columns (why?), say, the columns of the identity matrix. On this set, the right-hand side turns into 1, which completes the proof.

E8.11

Proofs of the claims from section H8.11.

Claim 1. Embeddings $X \overset{D'_\Phi}{\to} Y^*$, $Y \overset{D''_\Phi}{\to} X^*$ show that $\dim X = \dim Y$ as

$$\dim X \leq \dim Y^* = \dim Y \leq \dim X^* = \dim X.$$

Claim 2. A function $(X/Y^\perp) \times (Y/X^\perp) \xrightarrow{\bar{\Phi}} field$ given by $\bar{\Phi}(x + Y^\perp,\ y + X^\perp) := \Phi(x, y)$ is a well-defined nondegenerate bilinear form (why?), so claim 2 follows from claim 1.

Claim 3. Consider bilinear forms $U \times (Y/U^\perp) \xrightarrow{\Phi'} field$, $(X/V^\perp) \times V \xrightarrow{\Phi''} field$ given by, respectively, $\Phi'(u, y + U^\perp) := \Phi(u,y)$, $\Phi''(x + V^\perp, v) := \Phi(x,v)$ and apply claim 1.

The equality of the matrix row and column ranks follow from claim 2, as they are, respectively, codim $Y^\perp$ and codim $X^\perp$. (Why?)

E8.12

The permutability of $\wedge D$ and $\wedge$ is derived as follows. Inserting the decompositions $x_i = \sum_{j=1}^{n} \alpha_{ij} e_j$ $(i = 1,\ldots, k)$ into $\bigwedge_i x_i$ we have

$$\begin{aligned}
\wedge D\left(\bigwedge_i x_i\right) &= \wedge D\left(\bigwedge_i \sum_{j=1}^{n} \alpha_{ij} e_j\right) = \wedge D\left(\sum_{1 \leq j_1 < \ldots < j_k \leq n} \det\left(\alpha_{ij}\right)_{\substack{i=1,\ldots,k \\ j=j_1,\ldots,j_k}} \bigwedge_{m=1}^{k} e_{j_m}\right) \\
&= \sum_{1 \leq j_1 < \ldots < j_k \leq n} \det\left(\alpha_{ij}\right)_{\substack{i=1,\ldots,k \\ j=j_1,\ldots,j_k}} \wedge D\left(\bigwedge_{m=1}^{k} e_{j_m}\right) \\
&= \sum_{1 \leq j_1 < \ldots < j_k \leq n} \det\left(\alpha_{ij}\right)_{\substack{i=1,\ldots,k \\ j=j_1,\ldots,j_k}} \bigwedge_{m=1}^{k} De_{j_m} \\
&= \bigwedge_i \left(\sum_{j=1}^{n} \alpha_{ij} De_j\right) = \bigwedge_i D\left(\sum_{j=1}^{n} \alpha_{ij} e_j\right) = \bigwedge_i Dx_i\,.
\end{aligned}$$

E8.19

Keeping to the notations of section H8.19, we have $\omega = M \cdot e^*_{i_1} \wedge \ldots \wedge e^*_{i_k}$ on $\wedge^k \Pi$. (Why?) According to section P8.18, $*\,\omega = (-1)^{k(n-k)} \text{sign}\ \sigma_1 \cdot M \cdot e^*_{i_{k+1}} \wedge \ldots \wedge e^*_{i_n}$ on $\wedge^{k-n} \Pi^\perp$, where $\sigma_1 = \begin{pmatrix} 1 & \cdots & n \\ i_1 & \cdots & i_n \end{pmatrix}$. On the other hand, we have, in

accordance with section P8.15, $*\,\omega = (-1)^{k(n-k)} \det A \cdot \operatorname{sign}\sigma_2 \cdot (Ae)^*_{j_{k+1}} \wedge \ldots \wedge (Ae)^*_{j_n}$, where $\sigma_2 = \begin{pmatrix} 1 & \cdots & n \\ j_1 & \cdots & j_n \end{pmatrix}$. Hence,

$$\begin{aligned} M' &= (Ae)^*_{j_{k+1}} \wedge \ldots \wedge (Ae)^*_{j_n}\left(e_{i_{k+1}}, \ldots, e_{i_n}\right) \\ &= (-1)^{k(n-k)} \det A \cdot \operatorname{sign}\sigma_2 \cdot *\,\omega\left(e_{i_{k+1}}, \ldots, e_{i_n}\right), \end{aligned}$$

so $M' = \det A \cdot \operatorname{sign}\sigma_1 \cdot \operatorname{sign}\sigma_2 \cdot M$, which completes the proof, as

$$\operatorname{sign}\sigma_1 \cdot \operatorname{sign}\sigma_2 = \operatorname{sign}\sigma_1^{-1} \cdot \operatorname{sign}\sigma_2 = \operatorname{sign}\left(\sigma_1^{-1}\sigma_2\right) = \operatorname{sign}\begin{pmatrix} i_1 & \cdots & i_n \\ j_1 & \cdots & j_n \end{pmatrix}.$$

E8.20

The mixed product of any $x_1, \ldots, x_n \in \mathbb{R}^n$ can be calculated as follows:

$$\begin{aligned} \langle x_1, [x_2, \ldots, x_n]\rangle &= *\left((*[x_2, \ldots, x_n]) \wedge x_1\right) = (-1)^{n-1} * (x_2 \wedge \ldots \wedge x_n \wedge x_1) \\ &= *(x_1 \wedge \ldots \wedge x_n) = \Omega(x_1, \ldots, x_n). \end{aligned}$$

In particular, for $x_1 := [x_2, \ldots, x_n]$ we have

$$|x_1|^2 = \langle x_1, x_1\rangle = \langle x_1, [x_2, \ldots, x_n]\rangle = \Omega(x_1, \ldots, x_n).$$

On the other hand, because of the orthogonality $x_1 := [x_2, \ldots, x_n] \perp x_2, \ldots, x_n$, we have, expressing the ($n-1$-dimensional) Euclidean area on Π^{n-1} by the formula from section P8.12,

$$|x_1| \cdot |x_2 \wedge \ldots \wedge x_n| \quad = \quad \Omega(x_1, \ldots, x_n),$$

so $|x_1| = |x_2 \wedge \ldots \wedge x_n| = \sqrt{\Omega(x_1, \ldots, x_n)}$, **QED**. Also, the equality $|x_1| = |x_2 \wedge \ldots \wedge x_n|$ can be obtained avoiding geometrical arguments, using section P8.12 and the orthogonality of *:

$$|x_2 \wedge \ldots \wedge x_n| = |*\,(x_2 \wedge \ldots \wedge x_n)| = |x_1|.$$

Readers who still prefer a more conventional way of calculating may proceed as follows:

$$\begin{aligned}\Omega(x_1,\ldots,x_n) &= \det\begin{pmatrix} x_{11} & \ldots & x_{1n} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nn}\end{pmatrix} \\ &= \sum_{i=1}^{n}(-1)^{i-1}x_{1i}\det\begin{pmatrix} x_{21} & \ldots & \widehat{x_{2i}} & \cdots & x_{2n} \\ \vdots & & & & \\ x_{n1} & \ldots & \widehat{x_{ni}} & \ldots & x_{nn}\end{pmatrix} \\ &= \sum_{i=1}^{n}\left[\det\begin{pmatrix} x_{21} & \ldots & \widehat{x_{2i}} & \cdots & x_{2n} \\ \vdots & & & & \\ x_{n1} & \ldots & \widehat{x_{ni}} & \ldots & x_{nn}\end{pmatrix}\right]^2 = \sum_{i=1}^{n}x_{1i}^2 = |x_1|^2.\end{aligned}$$

Also, those readers may provide similar elementary computation of the mixed product.

E8.21

Restricting ourselves to X, Y, Z as prescribed in section H8.21 brings

$$[X,[Y,Z]] = *[e_i \wedge *(e_{n-1}\wedge e_n)] = *(e_i \wedge e_1 \wedge \ldots \wedge e_{n-2}),$$

which equals zero if $i \notin \{n-1, n\}$, that is, when $X \neq Y$ and $X \neq Z$. In this case, the right-hand side of the formula being proved also equals zero, as $\langle X,Y\rangle = \langle X,Z\rangle = 0$. If X equals Y or Z, say, $X = Y$, then

$$*\,(e_i \wedge e_1 \wedge \ldots \wedge e_{n-2}) = *(-1)^{n-2}e_1 \wedge \ldots \wedge e_{n-1} = -e_n = -1\cdot Z = -\langle X,Y\rangle Z$$

(and similarly for $X = Z$). **QED**.

Completing the Solution

S8.1

Keeping to the notations of section H8.1, a map φ: $\mathcal{P}/\mathcal{Q} \to L$ may be defined in a unique manner, namely, as $\varphi(p+\mathcal{Q}) = \mathcal{F} \circ \pi^{-1}(p+\mathcal{Q})$. The correctness of the

definition (insensitivity to the selection of $p \in p + \mathcal{Q}$) is provided by the inclusion $\mathcal{Q} \subseteq \ker \mathcal{F}$. The linearity of φ is evident. (We leave it to the reader to work out the details.)

The uniqueness of φ provides the uniqueness, within a unique isomorphism, of the definition of $\wedge^k \mathbb{R}^n$ using its basic property described in section P8.1***. (Indeed, consider $L = \wedge^k \mathbb{R}^n$.)

Turning to the central auxiliary statement in section E8.1, readers will easily verify the following statement:

A relation on a pair of vector spaces $\Gamma \subseteq X \times Y$ *closed with respect to* the *summations* $((x,y),(x',y') \in \Gamma \Rightarrow (x + x',\ y + y') \in \Gamma$: $\forall x,\ x' \in X,\ \forall y,\ y' \in Y)$ *is a graph of a map* $X \to Y$ *if and only if* $(0, y) \in \Gamma \Rightarrow y = 0$, $\forall y \in Y$. [*With this, the map will be linear if, in addition,* Γ *is closed with respect to multiplications by scalars:* $(x,y) \in \Gamma \Rightarrow (\alpha x,\ \alpha\ y) \in \Gamma$, $\forall x \in X$, $\forall y \in Y$, $\forall \alpha$.[7]]

S8.6

Proof of the elementary number-theoretic claim from section E6.8. Consider a polynomial of degree k in n:

$$p_k(n) := \binom{n}{k} - k(n-k) - 1$$
$$= \begin{cases} 0, & k = 0, \\ \dfrac{(n-k+1)\ \ldots\ n}{k!} - k(n-k) - 1, & k = 1, 2, \ldots\ . \end{cases}$$

For $k = 0$ and $k = 1$, $p_k(n)$ equals zero identically, and, since $p_k(n) = p_{n-k}(n)$ for natural n, $n = k + 1$ and $n = k$ are roots of p_k for any k. Therefore, we must establish that $p_k(n) > 0$ for $1 < k < n - 1$. Do this by induction on k by deriving from the inductive hypothesis an inequality $p_k'(n) > 0$ for $1 < k \leq n - 1$, which will make it possible to complete the proof. (Why?)

[7] The closure, with respect to multiplications by rational scalars, follows from the closure with respect to summations (why?), but for real scalars it does not. (Provide examples.)

Following this plan, estimate the derivative from below:

$$\begin{aligned} p'_k(n) &= \tfrac{d}{dn}\left[\frac{(n-k+1)\ \dots\ n}{k!}\right] - k \\ &= (k!)^{-1}\left[\sum_{i=1}^{k}(n-k+1)\dots\widehat{(n-k+i)}\dots n\right] - k \\ &> \frac{(n-k+1)\ \dots\ (n-1)}{(k-1)!} - k = \binom{n-1}{k-1} - k \end{aligned}$$

(the hat means that the corresponding term is missing). Now note that $(k-1)(n-k)+1 \geq k$ for $1 \leq k \leq n-1$, and so

$$p'_k(n) > \binom{n-1}{k-1} - (k-1)(n-k) - 1 = p_{k-1}(n-1).$$

For $k-1=1$, $p_{k-1}(n-1)$ equals zero identically (for all n). For $k-1>1$, $p_{k-1}(n-1)=0$ under $n=k+1$ and becomes positive for $n>k+1$ (by the inductive hypothesis). In all cases, $p_{k-1}(n-1) \geq 0$ for $1<k\leq n-1$, so $p_k'(n) > 0$. **QED**.

S8.11

A different proof of the equality of the row and column ranks was discussed in the "Polar and Singular Value Decomposition Theorems" problem group (section S6.3).

Also, readers must answer the following question, which arises in connection with claim 3 from section H8.11. As we have shown, the orthogonal (with respect to a nondegenerate bilinear form) complement to a subspace must have a complementary dimension: $\dim U + \dim U^{\perp} = \dim X$. Given this, does the equality $U + U^{\perp} = X$ also hold?

S8.17

The proof of equivalence of the two definitions for angles between hyperplanes, as discussed in section H8.17, can be done as follows. The angle between hyperplanes Π_1, Π_2, defined via the isomorphism $G(n,n-1) \cong P(\wedge^{n-1}\mathbb{R}^n)$, equals, by virtue of section P8.16***, the angle between properly oriented normals to Π_1, Π_2, respectively, which is defined via the isomorphism $G(n,1) \cong P(\wedge^1\mathbb{R}^n) = P(\mathbb{R}^n) = \mathbb{R}P^n$. The angle between the normals, in turn, equals the angle between

them defined in the usual way (using, for example, the tools of elementary geometry). This last angle equals the "elementary-geometrical" angle between Π_1 and Π_2, which may be immediately verified applying Theorem 1 from section P8.0** for a two-dimensional plane orthogonal to $\Pi_1 \cap \Pi_2$. (Work out the details, paying attention to the orientations of the normals. Also, we leave it to the reader to complete a case of parallel Π_1, Π_2.)

Also, readers may prove the foregoing equivalence using the following arguments. For the element of oriented Euclidean volume Ω, isomorphism I_Ω (defined in section P8.10***) maps $\wedge^{n-1}\mathbb{R}^n$ onto $\mathbb{R}^{n^*}$. But since $\mathrm{I}_\Omega = D \circ *^{-1}$ (section P8.14***) and since the star operator preserves the scalar product on the exterior algebra (section P8.16***), I_Ω maps the scalar product in $\wedge^{n-1}\mathbb{R}^n$ to the scalar product in $\mathbb{R}^{n^*}$, dual with the scalar product in $\mathbb{R}^n$. Therefore, we need only verify that the "elementary-geometrical" angle between Π_1 and Π_2 is the same, as the angle between the straight lines determined by Π_1 and Π_2 in the space of linear functionals $\mathbb{R}^{n^*}$; precisely this is done by applying Theorem 1 from section P8.0** to a two-dimensional plane orthogonal to $\Pi_1 \cap \Pi_2$. (Complete the proof.)

Convexity and Related Classic Inequalities

Problems

P9.0

Preliminaries. Concepts of convexity and related techniques are widely used in various branches of mathematics and applications, such as functional analysis (Dunford and Schwartz 1957; Hille and Phillips 1957; Yosida 1965; Edwards 1965; Riesz and Nagy 1972; Rudin 1973; Nirenberg 1974; Kolmogorov and Fomin 1976; Reed and Simon 1972; Bourbaki 1981; Krein et al. 1972; Kirillov and Gvishiani 1988; Lubich 1988), calculus of variations and mathematical physics (Ekeland and Temam 1976; Arnol'd 1989; Polya and Szegö 1951), geometry (Hilbert and Cohn-Vossen 1932; Burago and Zalgaller 1980), number theory (Cassels 1959, 1978; Schmidt 1980), game theory (von Neumann and Morgenstern 1953; Luce and Raiffa 1957; Moulin 1981), optimization and other computational methods and convex programming (Collatz 1964; Krasnoselskii et al. 1969; Céa 1971; Kantorovich and Akilov 1972; Balakrishnan 1976; Aubin and Ekeland 1984), integral geometry (tomography) (Gelfand et al. 1959–1962; Helgason 1980, 1984; Herman et al. 1987; Gardner 1995), and many others. In addition, there is a vast literature on the fundamental topics in convexity and related inequalities; interested readers may refer to Beckenbach and Bellman (1961), Hardy et al. (1934), Marcus and Minc (1964), Leichtweiss (1980), Rockafellar (1970), Kutateladze and Rubinov (1976), and the multiple references therein.

In this problem group, we will focus mainly on the basic elements of the theory of convex functions, which can be taught to readers with limited experience (sections P9.1*, P9.2*, P9.3*, P9.4*, P9.6*, P9.7**, P9.8*, P9.9*, P9.21*, P9.22*, P9.23, P9.24***, P9.27***, and P9.28***). The remaining problems in this group are directed toward specific applications such as convex and linear programming (section P9.5*), the hierarchy of power means (sections P9.10**, P9.12*, and P9.14**), geometric inequalities (sections P9.13**, P9.16**, and P9.17), the Hölder and Minkowski inequalities (sections P9.15** and P9.18**), Young's inequality and

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_9,

the Legendre transform (sections P9.19** and P9.20**), and functions on metric and normed vector spaces (sections P9.25***, P9.26***, P9.29***, and P9.30***). The necessary introductory material is provided in these sections. (The reader will find additional references, definitions, and tools related to problems in these sections in the "Hint" section of the book.) A specially added section (P9.11**) discusses the widely known Cauchy-Schwarz-Bunyakovskii inequality and Lagrange-type identities in connection with the hierarchy of the power means.

*P9.1**

Recall that a real-valued function of one or many real variables, $f(x)$, is referred to as **convex** if it satisfies the so-called **Jensen inequality**

$$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha)f(y), \qquad \forall \alpha \in (0,1),\ \forall x, y.$$

Natural domains for convex functions are **convex sets**, which are the subsets of (real) vector spaces containing, together with points x and y, the line segment that connects them, $\{\alpha x + (1-\alpha)y : 0 \le \alpha \le 1\}$. Throughout this section we consider convex functions defined on convex domains. *Translate* the definition of a convex function into the language of geometry (Fig. 1). *Verify* that **affine transforms**

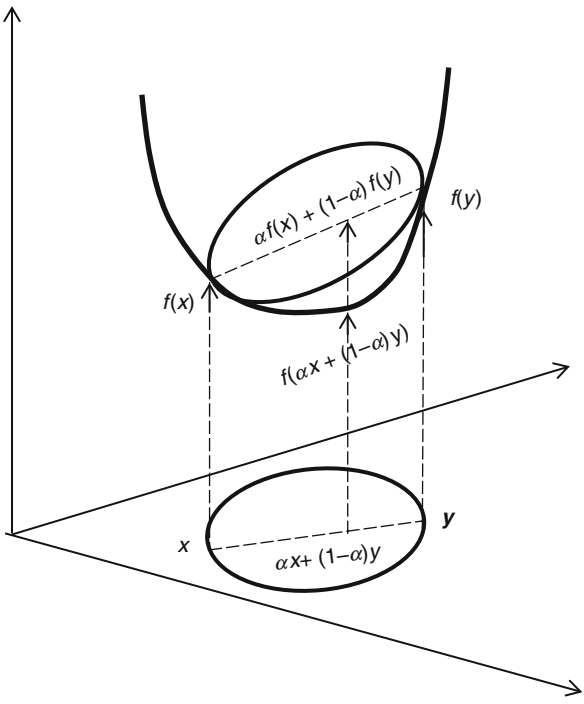


Fig. 1 The graph of a convex function

(combinations of linear transforms and translations) of domains affect neither their convexity nor the convexity of the functions on them. Also, *verify* that affine transformations of the range of values that do not change its orientation do not affect the convexity of the functions.

A **concave** function is defined by the Jensen inequality using an inverted sign. Hence, $f(x)$ is convex if and only if $-f(x)$ is concave, and, obviously, linear functions are both convex and concave (*prove* the inverse statement). *Provide* more examples of convex and concave functions. *Show* that linear combinations with positive coefficients of convex (concave) functions on a common domain are convex (resp. concave). Therefore, adding a linear function does not affect convexity (concavity).

Prove that pointwise limits of convex (concave) functions on a common domain, if they exist, are convex (resp. concave). Similarly, *prove* that pointwise supremums (infimums) of convex (resp. concave) functions on a common domain, uniformly bounded above (resp. below) on any element, are convex (resp. concave).

For convex functions f, g (say, of one variable), *can* any statements be made about the convexity or concavity of functions (1) fg, (2) $1/g$ (out of $\{g = 0\}$), (3) f', (4)** $\int f$, (5) f^{-1} (for continuous monotonic f), or (6) $f \circ g$, provided that f does not decrease as well?

Since $f(x)$ is convex if and only if $-f(x)$ is concave, those functions have similar properties. (The definitions of convex and concave functions in some literature sources are transposed.) *Therefore, in what follows we will discuss only convex functions.*

*P9.2**

Verify that a whole vector space and all its vector and affine subspaces are convex sets. *Show* that the closure and the interior (if nonempty) of a convex set itself is convex. Convex sets may be bounded or unbounded; they may be open, closed, or not open and not closed. (*Provide* examples.) Can a convex set be disconnected? *Verify* that intersections of convex sets themselves are convex sets. *Prove* that a convex set in a finite-dimensional space lies in a proper subspace or has a nonempty interior. *Enumerate* all types of convex sets on a straight line (real axis).

*P9.3**

Following sources such as Kutateladze and Rubinov (1976), we will define an **epigraph** of a function f: $D \to \mathbb{R}$ as a set of the points in $D \times \mathbb{R}$ lying above or on the graph of f, $epi(f) := \{(x,y): x \in D, y \geq f(x)\}$. *Which* property of the set $epi(f)$ is equivalent to the convexity of a function f defined on a convex set?

P9.4[*]

Prove, for a convex function, a more general form of the Jensen inequality

$$f\left(\sum_{i=1}^{n}\alpha_i x_i\right)\le\sum_{i=1}^{n}\alpha_i f(x_i)\quad\left(n=1,2,\ldots;\ \ 0\le\alpha_i\le 1,\ \ \forall i;\quad\sum\alpha_i=1\right).$$

[The usual related terminology is as follows: α_i are referred to as **relative weights**, and the sums $\sum\alpha_i x_i$ are called **weighted** (**arithmetic**) **means**, **weighted centers of gravity,** or **convex linear combinations**.]

P9.5[*]

Obviously, any subset M of a vector space is contained in a convex subset (the space itself), but is there a minimal convex subset containing M, a so-called **convex hull** of M? *Show* that the convex hull exists, is unique, and is an intersection of the convex sets containing M. In particular, the convex hull of M is always contained in the same affine subspace (or another convex subset) as M. *Determine* whether the convex hull of a closed subset is closed. *Prove* that the convex hull of M coincides with the set of weighted means of all points of M:

$$conv(M)=\left\{\sum_{x\in M}\alpha_x x:0\le\alpha_x\le 1,\ \ \alpha_x>0\text{ for a finite number of }x,\quad\sum\alpha_x=1\right\}.$$

We will use the definition of a **polyhedron** in $\mathbb{R}^n$ as a nonempty bounded set that is a closure of its interior, with the piecewise hyperplanar boundary being a union of polyhedrons of dimension $n-1$ (faces) that intersect along polyhedrons of dimensions $\le n-2$ (edges) that intersect along polyhedrons of dimensions $\le n-3$ (edges), and so on (up to the level of vertices).[1] *Show* that a bounded set with a nonempty interior, which is a finite intersection of closed half-spaces, is a polyhedron; the aforementioned set is referred to as a **convex polyhedron**.

Show that a convex polyhedron is a convex hull of its vertices and that the convex hull of any polyhedron coincides with the convex hull of the set of its vertices.

Show that a k-dimensional face of a polyhedron contains $k+1$ vertices that are not contained in a plane of a smaller dimension.

[1] This definition allows disconnected polyhedrons and unions of polyhedrons intersecting along edges of codimensions greater than one. Examples are polyhedrons with a common vertex and polyhedrons obtained by partitioning faces (with the corresponding partition and addition of edges of all dimensions) into smaller ones, which would count as different.

Prove that the convex hull of a finite set is a convex polyhedron with vertices that are some points of this set. (*Which* points exactly?) Therefore, the convex hull of a polyhedron M is a convex polyhedron with vertices that are some vertices of M.[2]

Show that if the maximum of a convex function is achieved at some internal point of its domain, then this function has a constant value.

Prove the *"Fundamental theorem of convex programming"*: The maximum of a convex function on any polyhedron M within this function's domain is achieved at a vertex of conv(M).

Derive from it the *"Fundamental theorem of linear programming"*: The maximum and minimum of a linear function on a *polyhedron M are achieved at some vertices of conv*(M).

(See the corresponding Hint for additional definitions or tools related to certain problems in this section.)

*P9.6**

We say that a function f: $D \to \mathbb{R}$ on a convex domain D allows a weight $\alpha_0 \in (0,1)$ if f obeys the Jensen inequality with relative weights α_0, $1 - \alpha_0$. *Prove* that the set of all allowed weights is always either empty or everywhere dense in [0,1] and contains all rational numbers of this segment so that f will satisfy the Jensen inequality with arbitrary rational weights:

$$f\left(\sum_{i=1}^{n}\frac{p_i}{q_i}x_i\right) \le \sum_{i=1}^{n}\frac{p_i}{q_i}f(x_i)$$
$$\left(n = 1,2,\ldots; \quad x_i \in D,\ p_i, q_i \in \not\subset,\ 0 \le \frac{p_i}{q_i} \le 1,\ \forall i; \quad \sum\frac{p_i}{q_i} = 1\right).$$

Show that a function that is continuous on a convex domain and that allows any weights is convex; in other words, the set of the allowed weights is all of (0,1).

Looking ahead, the continuity of a function on an open domain in a finite-dimensional space is necessary for convexity. *Can* the continuity of this function be derived from the fact that it allows weights (and omitted in the formulation of the last claim), or can you give examples of nonconvex functions that allow weights?

[2] Proving these claims, obvious for planar polygons, requires more sophisticated methods for dimensions greater than two. Working on this problem will help the reader in learning those methods because most real-life problems are of high dimensions.

P9.7[**]

Establish for a continuous convex function f: $D \to \mathbb{R}$ on a convex domain D the Jensen inequality in the following general form:

$$f\left(\int_0^1 X(t)dt\right) \le \int_0^1 f(X(t))dt,$$

for an integrable function X: $[0,1] \to D$; specifically, for a countable set of relative weights $\{\alpha_i\}$ $(\alpha_i \ge 0, \sum_{i=1}^{\infty} \alpha_i = 1)$ and a bounded sequence $\{x_i\} \subset D$, this will look like

$$f\left(\sum_{i=1}^{\infty} \alpha_i x_i\right) \le \sum_{i=1}^{\infty} \alpha_i f(x_i).$$

(Try to find a proof that is valid in both univariate and multivariate cases. Readers with limited experience in multivariate analysismay restrict themselves to a function of one variable.)

Readers familiar with elements of probability theory may establish the Jensen inequality in an even more general form:

$$f\left(\int_\Omega X\, d\mu\right) \le \int_\Omega f(X)\, d\mu$$

for an integrable function X: $\Omega \to D$ on a space Ω with a probability measure $d\mu$.

P9.8[*]

A hyperplane P in a vector space is referred to as a **hyperplane of support** for a subset M of this space at a point $a \in M$ if $a \in P$ and M is located in a closed half-space bounded by P. *Prove* that *epi*(f) for a convex function of one variable has straight lines of support at the points of the graph of f. [At the points $(x, f(x))$, corresponding to internal x, the lines of support are nonvertical (nonparallel to the ordinate axes); *why*?]

An extension of this result to the multivariate case is far from trivial (this problem is formulated in section P9.27[***] below). However, if the function f is differentiable, then the existence of hyperplanes of support can be proved quite simply for any number of variables (in the next section readers are invited to do so).

Conversely, *prove* that a function f defined on a convex set will be convex if *epi*(f) has hyperplanes of support at the points $(x, f(x))$ corresponding to internal x. (A proof may be provided that handles the multivariate case by a slight modification of the univariate one. Therefore, readers with limited experience may restrict themselves to a function of one variable, but more advanced readers should try to find a proof that is valid for a multivariate case.)

Will f be convex if the support condition is weakened to a local support, which means that there is a neighborhood $U(x)$ for any x, such that $epi(f \mid_{U(x)})$ has a hyperplane of support at $(x, f(x))$?

P9.9*

Prove that a differentiable function f: $D \to \mathbb{R}$ on an open convex domain is convex if and only if the tangent hyperplanes to the graph are hyperplanes of support for *epi*(f). Analytically, $f(x) \geq f(x_0) + f'_{x_0}(x - x_0), \ \forall x, x_0$, where f'_{x_0} is the differential at x_0 [a linear functional on the tangent hyperplane; in univariate cases, $f'_{x_0}(x - x_0)$ is equal to a product of the derivative at x_0 and the increment $x - x_0$].

Based on the foregoing discussion, *derive* the claim that a convex function achieves its global minimum at every critical point (the internal points x_0 such that $f'_{x_0} = 0$).

Also, *prove* that a twice-differentiable function on the same domain is convex if and only if it has its second differentials (symmetric bilinear forms on the corresponding tangent hyperplanes) positive semidefinite at all points, $f''_{x_0}(x - x_0, x - x_0) \geq 0, \ \forall x, x_0$. [In the univariate case, $f''_{x_0}(x - x_0, x - x_0)$ is equal to a product of the second derivative at x_0 and the squared increment $x - x_0$.]

(Proofs may be provided that handle the multivariate case by a slight modification of the univariate one.)

P9.10**

Prove that a positive function X: [0,1] $\to$ (0,∞), integrable with power q, $\int_0^1 [X(t)]^q dt < \infty$, is integrable with any power between q and 0; also,

$$\left(\int_0^1 [X(t)]^p dt \right)^q \leq \left(\int_0^1 [X(t)]^q dt \right)^p$$

when it is integrable with powers $p < q$. *Derive* from this the hierarchy of the weighted power means

$$\left(\sum \alpha_i x_i^p\right)^{1/p} \le \left(\sum \alpha_i x_i^q\right)^{1/q} \quad \left(p,q \ne 0,\ p<q;\quad x_i>0,\ 0 \le \alpha_i \le 1,\ \forall i;\quad \sum \alpha_i = 1\right).$$

(For $p,q = -1,1,2$, these are referred to as harmonic, arithmetic, and quadratic means, respectively.) *Determine* the cases of equalities in these inequalities.

P9.11[**]

We discuss here several important identities and inequalities associated with the inequality between arithmetic and quadratic means from section P9.10[**]. Being written as $n\sum x_i^2 \ge \left(\sum x_i\right)^2$ (using equal weights of x_i), the inequality between arithmetic and quadratic means is a special case of the CSB inequality $\left(\sum x_i^2\right)\left(\sum y_i^2\right) \ge \left(\sum x_i y_i\right)^2$. . [In turn, this famous CSB inequality is a special case of the Hölder inequality (section P9.15[**]) that has important applications in functional analysis.] The reader should know a common proof of this inequality in its geometric form: for two vectors x, y in a Euclidean space, $|x|^2|y|^2 \ge \langle x,y\rangle^2$ because the difference between the right- and left-hand sides is proportional to the discriminant of a nonnegative quadratic trinomial in t, $\langle x+ty, x+ty\rangle$.

Following Beckenbach and Bellman (1961), generalize this solution. The positive definiteness of a scalar product yields for any $t_1,\ldots,t_n$ the following inequalities:

$$\begin{aligned} 0 &\le \langle t_1e_1 + \ldots + t_ne_n,\ \ t_1e_1 + \ldots + t_ne_n\rangle \\ &= \begin{pmatrix} t_1 & \ldots & t_n \end{pmatrix} \begin{pmatrix} \langle e_1,e_1\rangle & \ldots & \langle e_1,e_n\rangle \\ \vdots & & \vdots \\ \langle e_n,e_1\rangle & \cdots & \langle e_n,e_n\rangle \end{pmatrix} \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}, \end{aligned}$$

meaning that the Gram matrix $(\langle e_i,e_j\rangle)_{i,j=1,\ldots,n}$ is positive semidefinite. (For linearly independent $e_1,\ldots,e_n$, this is positive definiteness). The positiveness of the corner minors including the CSB inequality (for 2×2 minors) follows from **Sylvester's criterion**. (*Complete* the details.)

Also, we may obtain this inequality by the following short calculation:

$$0 \le |x|y| - y|x||^2 = 2|x|^2|y|^2 - 2|x||y|\langle x,y\rangle = 2|x||y|(|x||y| - \langle x,y\rangle).$$

Readers who have worked with the book of Arnol'd (1975) know the following nice proof of the CSB inequality: we may consider only a two-dimensional Euclidean plane (*why*?); in turn, for dimension two this inequality becomes $\left(x_1^2+x_2^2\right)\left(y_1^2+y_2^2\right) \ge \left(x_1y_1+x_2y_2\right)^2$, which is verified directly. (*Complete* the details.) Another proof utilizes a minimization of the quadratic form in x, y using differential calculus. Lastly, the CSB inequality may be proved using Lagrange's method, which consists of a representation of the difference between the right- and left-hand sides by a linear combination of monomials of even powers with positive

coefficients, which obviously is a nonnegative function. In our special case, this representation is $n\sum x_i^2 - \left(\sum x_i\right)^2 = \sum_i \left(x_i - x_j\right)^2$, which is easy to prove:

$$n\sum x_i^2 - \left(\sum x_i\right)^2 = \det\begin{pmatrix} \sum x_i^2 & \sum x_j \\ \sum x_i & n \end{pmatrix}$$
$$= \sum_{i,j} \det\begin{pmatrix} x_i^2 & x_j \\ x_i & 1 \end{pmatrix} = \sum_{i,j} \left(x_i^2 - x_i x_j\right) = \sum_i \left(x_i - x_j\right)^2.$$

Prove the complete Lagrange identity $\left(\sum x_i^2\right)\left(\sum y_i^2\right) - \left(\sum x_i y_i\right)^2 = \sum_i \left(x_i y_j - x_j y_i\right)^2$. In fact, there is a Lagrange-type identity for any nonnegative homogeneous power form. Interested readers will find this theorem and some other generalizations of the above Lagrange identity in Beckenbach and Bellman (1961), Hardy et al. (1934), and Polya and Szegö (1964). One of the generalizations was discussed in section P3.2** ("A Combinatorial Algorithm in Multiexponential Analysis" problem group above):

$$\det\begin{pmatrix} \sum x_i^{2k} & \cdots & \sum x_i^k \\ \vdots & & \vdots \\ \sum x_i^k & \cdots & n \end{pmatrix} = \begin{cases} 0 & \text{when } n \le k, \\ \sum\limits_{1 \le i_0 < \ldots < i_k \le n} \prod\limits_{0 \le \mu < \nu \le k} \left(x_{i_\nu} - x_{i_\mu}\right)^2 & \text{otherwise,} \end{cases}$$

or, completely,

$$\det\begin{pmatrix} \sum x_i^{2k} & \cdots & \sum x_i^k y_i^k \\ \vdots & & \vdots \\ \sum x_i^k y_i^k & \cdots & \sum y_i^{2k} \end{pmatrix} = \begin{cases} 0 & \text{when } n \le k, \\ \sum\limits_{1 \le i_0 < \ldots < i_k \le n} \prod\limits_{0 \le \mu < \nu \le k} \left(x_{i_\nu} y_{i_\mu} - x_{i_\mu} y_{i_\nu}\right)^2 & \text{otherwise.} \end{cases}$$

Advanced readers may have encountered this determinant used for a different purpose in interpolation theory while performing a least-squares fitting of data $\{(x_i, \varphi(x_i))\}_{i = 1,\ldots,n}$ ($x_i \ne x_j$ for $i \ne j$) with a polynomial of degree k [$k = 1$ corresponds to the rectilinear fitting; for $k = n - 1$ the Lagrange interpolating polynomial $\varphi(x)$ is obtained]. The nonnegativity of this determinant, including the positiveness for $n > k$, may be established using other techniques discussed in the "Least Squares and Chebyshev Systems" problem group.

*P9.12**

Include the weighted geometric mean $\prod x_i^{\alpha_i}$ ($x_i > 0$, $0 \le \alpha_i \le 1$, $\forall i$; $\sum \alpha_i = 1$) as a mean of power zero in the hierarchy from section P9.10** by *establishing* the limit relation $\lim\limits_{p \to 0,\, p \ne 0} \left(\sum \alpha_i x_i^p\right)^{1/p} = \prod x_i^{\alpha_i}$.

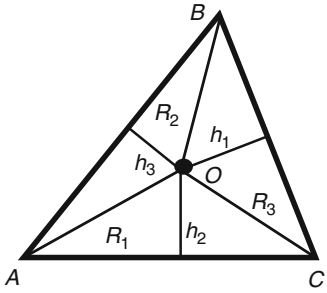


Fig. 2 To the formulation of the Erdös-Mordell inequality

And thus we have $\left(\sum \alpha_i x_i^p\right)^{1/p} \geq \prod x_i^{\alpha_i}$, $\forall p > 0$ and $\left(\sum \alpha_i x_i^p\right)^{1/p} \leq \prod x_i^{\alpha_i}$, $\forall p < 0$. For $p = 1$, this is the widely known **Cauchy inequality** between arithmetic and geometric means.

P9.13[**]

The famous **Erdös-Mordell geometric inequality** states that given a point O inside a triangle ABC and denoting R_1, R_2, R_3 as the lengths of segments OA, OB, and OC, respectively, and h_1, h_2, h_3 as the heights at vertex O of triangles OBC, OAC, and OAB, respectively (Fig. 2), $2\sum h_i \leq \sum R_i$, with equalities for and only for the equilateral triangle and the point at its center. The history, several nice geometric proofs of this inequality by various mathematicians (L.J. Mordell, D.F. Barrow, D.K. Kazarinoff, L. Bankoff, G.R. Veldkamp, H.H. Eggleston, H. Bralant, Z.A. Skopetz), and a discussion of related topics can be found in Kazarinoff (1961), Chentzov et al. (1970), and references therein. Some of these mathematicians actually proved a stronger inequality $2\sum l_i \leq \sum R_i$ (with the same cases of equalities), where l_1, l_2, l_3 are bisectors at vertex O of triangles OBC, OAC, and OAB, respectively. *Verify* that, denoting by $\gamma_1, \gamma_2, \gamma_3$ the halves of angles $\sphericalangle AOB$, $\sphericalangle BOC$, $\sphericalangle COA$, respectively, this stronger inequality may be written in algebraic form as

$$\sum_{i=1}^{3} \cos\gamma_i \cdot H\left(R_i, R_{(i+1)\bmod 3}\right) \leq \cos\frac{\pi}{3} \cdot \sum_{i=1}^{3} R_i, \quad R_i > 0, \quad 0 \leq \gamma_i \leq \frac{\pi}{2}, \quad \sum \gamma_i = \pi, \tag{*}$$

where $H(x,y)$ is the harmonic mean, $H(x,y) = 2/(x^{-1} + y^{-1})$.

F. Tóth suggested a natural generalization of the Erdös-Mordell inequality for a convex n-angle: $\sum_{i=1}^{n} h_i \leq \cos\frac{\pi}{n} \cdot \sum_{i=1}^{n} R_i$ ($n \geq 3$, with equalities for and only for the perfect polygon and the point at its center) (Tóth 1948); this generalization holds for the stronger inequality as well (Florian 1958; Lenhard 1961; Leuenberger 1962; Vogler 1966). Let us attempt to sharpen of the foregoing Erdös-Mordell-Tóth-

Florian-Lenhard-Leuenberger-Vogler inequality (briefly, the Erdös-Mordell inequality) for arbitrary $n \geq 3$ by replacing harmonic means with the geometric ones (in accordance with section P9.12*):

$$\sum_{i=1}^{n} \cos\gamma_i \cdot \sqrt{R_i R_{(i+1)\bmod n}} \leq \cos\tfrac{\pi}{n} \cdot \sum_{i=1}^{n} R_i, \quad R_i>0,\ 0 \leq \gamma_i \leq \tfrac{\pi}{2}, \sum \gamma_i = \pi, \quad (**)$$

or, with a substitution of variables $x_i = R_i^{1/2}$,

$$\sum_{i=1}^{n} \cos\gamma_i \cdot x_i x_{(i+1)\bmod n} \leq \cos\tfrac{\pi}{n} \cdot \sum_{i=1}^{n} x_i^2, \quad 0 \leq \gamma_i \leq \tfrac{\pi}{2}, \sum \gamma_i = \pi. \quad (***)$$

Prove the inequality (***) for $n = 3$ and *investigate* the cases of equality. [In doing this readers will also establish the stronger Erdös-Mordell inequality (*).] Could the inequality (**) itself be strengthened by substituting power means of positive powers for the geometric mean?

Try to answer whether (***) holds true for any $n > 3$. (If yes, the reader will also establish Tóth's generalization of the Erdös-Mordell inequality.)

*P9.14***

We define for nonnegative-element vectors $X = (x_1,\ldots)$ the **weighted power means**

$$M_p(X) = M_p(X, \alpha_1, \ldots): = \begin{cases} \left(\sum \alpha_i x_i^p\right)^{1/p} & \text{when } p \neq 0, \\ \prod x_i^{\alpha_i} & \text{otherwise} \end{cases} \quad \left(0 \leq \alpha_i \leq 1,\ \sum \alpha_i = 1\right).$$

In the continuous version, the **integral power means** are defined for functions X: $[0,1] \to [0,\infty]$ integrable with any powers (readers preferring not to deal with functions with infinite values may consider them bounded from zero and infinity):

$$M_p(X): = \begin{cases} \left(\int_0^1 [X(t)]^p dt\right)^{1/p} & \text{when } p \neq 0, \\ \exp\left(\int_0^1 \log X(t)\, dt\right) & \text{otherwise.} \end{cases}$$

From sections P9.10** and P9.12* we have the monotonicity of $M_p(X)$ on p. Find the limits $\lim_{p\to\infty} M_p(X)$ and $\lim_{p\to-\infty} M_p(X)$ for the vectors $X = (x_1,\ldots)$ and for the functions X: $[0,1] \to [0,\infty]$.

Readers familiar with elements of probability theory may do the same for the integral power means

$$M_p(X, d\mu) := \begin{cases} \left(\int\limits_{\Omega} X^p d\mu \right)^{1/p} & \text{when } p \neq 0, \\ \exp\left(\int\limits_{\Omega} \log X \, d\mu \right) & \text{otherwise,} \end{cases}$$

defined for the continuous function, integrable with any powers, X: $\Omega \to [0,\infty]$ on a topological space Ω with a probability measure $d\mu$.

*P9.15***

Derive from the Cauchy inequality the **Hölder inequality**: for a (rectangular) nonnegative-element matrix (x_{ij}) and relative weights α_i ($0 \le \alpha_i \le 1$, $\sum \alpha_i = 1$),

$$\sum_j \prod_i x_{ij}^{\alpha_i} \le \prod_i \left(\sum_j x_{ij} \right)^{\alpha_i},$$

with equalities if and only if the matrix (x_{ij}) has some of its rows equal to zero or has all its rows proportional. Also, we may have an integral form of this inequality. For this, multiply both sides by Δt and distribute $(\Delta t)^{\alpha_i}$ among the factors on the right-hand side. Then, considering the sums as integral ones, pass on to the limit. This will result, for integrable nonnegative functions X_i: $[0,1] \to [0,\infty)$, in the integrability of a product of their α_i-powers, and the desired inequality

$$\int_a^b \prod [X_i(t)]^{\alpha_i} dt \le \prod \left(\int_a^b X_i(t)\, dt \right)^{\alpha_i},$$

with equalities if and only if some of the functions X_i are zero almost everywhere (up to a set of zero volume) or all of them are proportional almost everywhere. (Advanced readers may substitute a measure space for a rectilinear segment $[a,b]$.) Work out the details of this construction. The case that is most widely used is the case of two functions; it is commonly written as

$$\int_a^b |X(t)Y(t)|\, dt \le \left(\int_a^b |X(t)|^p\, dt \right)^{1/p} \left(\int_a^b |Y(t)|^q\, dt \right)^{1/q} \quad \left(p>1,\ \frac{1}{p}+\frac{1}{q}=1 \right),$$

or, in discrete form, $\sum x_i y_i \le \left(\sum x_i^p\right)^{1/p} \left(\sum y_i^p\right)^{1/q}$ $(x_i, y_i \ge 0)$. Specifically, $p = q = 2$ corresponds to the CSB inequality.

P9.16[**]

Inequalities may be used to solve problems instead of more awkward analytical techniques. Let us consider as an example a simple problem of determining the distance from the origin of a Euclidean space to an affine hyperplane P (a subspace of codimension one shifted from the origin). The usual solution to this problem uses linear algebra or differential calculus (we may consider it as a Lagrange minimization problem – minimization of the values of a positive definite quadratic form on a hyperplane). At the same time, the CSB inequality provides a solution in a much easier way. *Verify* that!

The effectiveness of the CSB inequality-based technique in this and related problems stems from a deep connection of that inequality to the Euclidean metric. *Verify* the equivalence for the Euclidean metric of the triangular inequality $|x + y| \leq |x| + |y|$ to the CSB inequality.

P9.17

Let us pose the following problem. Given a planar polygon P (a closed set in $\mathbb{R}^2$ bounded by a closed finite-sided polygonal line, not necessarily connected, free of self-intersections), determine a point $x_m(P) \in \mathbb{R}^2$ with a minimal value $m(P)$ of summed squared distances from the sides (or their extensions), and determine $m(P)$ itself. A similar problem arises for polyhedrons in $\mathbb{R}^n$. *Provide* a complete solution where you perform the following steps:

- [*] *Establish* the unique existence of $x_m(P)$ for the polygons (polyhedrons). [***Warning***: generally, $x_m(P)$ lacks ties with the center of gravity, though it is invariant with respect to orthogonal linear transforms and parallel translations T: $\mathbb{R}^n \to \mathbb{R}^n$: $T(x_m(P)) = x_m(T(P))$.] *Does* the same hold for any polyhedral-shaped bodies obtained by combining finite unions and intersections of closed half-spaces?[3]
- [**] *Provide* a numerically realizable algorithm for the determination of $m(P)$ and $x_m(P)$, given a polygon (polyhedron) P.
- [**] *Prove* the inclusion $x_m(P) \in \text{int } P$ for the triangles (simplexes). *Does* it hold for any convex polygons (polyhedrons)?
- [***] *Find* explicitly a function $f(P)$ of a polyhedron of dimension n, which is determined with ($n - 1$-dimensional) areas of all its faces $A_1, \ldots, A_k$ (k is a number of the faces) and its (n-dimensional) volume V, and is such that $m(P) \geq f(P)$ for any polyhedrons P and $m(P) = f(P)$ for the simplexes (when $k = n + 1$). *Provide* examples, not simplexes, of $m(P) = f(P)$ and examples of $m(P) > f(P)$.

Find an appropriate method that requires no modifications for extensions to $n > 2$ and $k > n + 1$!

[3] For advanced readers, the proper name for these shapes is n-dimensional submanifolds with piecewise-planar boundaries.

P9.18[**]

Keeping the notations of section P9.14[**], the Hölder inequality (P9.15[**]) is

$$\sum_j M_0\big(X_j,\ \alpha_1,\dots\big) \le M_0\left(\sum_j X_j,\ \alpha_1,\dots\right) \quad \big(X_j = \big(x_{1j}, x_{2j},\dots\big)\big).$$

The extension to any weighted power mean is referred to as the **Minkowski inequality**. For nonnegative (positive, for $p < 0$) vectors of equal dimensions X_j this inequality states that

$$\sum_j M_p\big(X_j,\ \alpha_1,\dots\big) \le (\ge) M_p\left(\sum_j X_j,\ \alpha_1,\dots\right) \quad \text{when } p<1 (\text{resp. } p>1),$$

or (for $p \ne 0$)

$$\sum_j \left(\sum_i \alpha_i x_{ij}^p\right)^{1/p} \le (\ge) \left(\sum_i \alpha_i \left(\sum_j x_{ij}\right)^p\right)^{1/p} \quad \text{when } p<1 (\text{resp. } p>1),$$

with equalities achieved if and only if matrix (x_{ij}) has all its rows proportional. Actually, we may simplify this inequality by eliminating variables that are not essential for it. Note that α_i are not weights in the usual sense, as they may obviously be multiplied all together by any positive factor without violating the inequality. Therefore, if the inequality holds for any $\alpha_i \ge 0$ with $\sum \alpha_i > 0$, one may get rid of them by substitution of variables. Making the substitution $x_{ij} \mapsto \alpha_i^{1/p} x_{ij}$ reduces the preceding inequality to

$$\sum_j \left(\sum_i x_{ij}^p\right)^{1/p} \le (\ge) \left(\sum_i \left(\sum_j x_{ij}\right)^p\right)^{1/p} \quad \text{when } p<1 \ (\text{respectively, } p>1).$$

Explain the geometrical meaning of this inequality for a two-element set of indices j and $p = 2$. *Derive* the Minkowski inequality including the cases of equalities from the Hölder inequality. Also, *establish* three integral forms of the Minkowski inequality by inserting Δu, or Δv, or both Δu and Δv simultaneously in proper places on both sides of the inequality and passing on to the limits in the obtained integral sums:

$$\sum_j \left(\int_a^b |X_j(u)|^p du\right)^{1/p} \le (\ge) \left(\int_a^b \sum_j |X_j(u)|^p\, du\right)^{1/p},$$

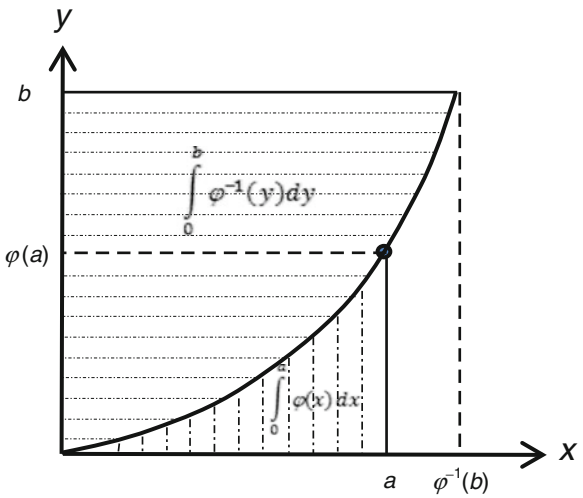


Fig. 3 To the formulation of the young's inequality

$$\int_c^d \left(\sum_i |X_i(v)|^p \right)^{1/p} dv \le (\ge) \left(\sum_i \left(\int_c^d |X_i(v)|\, dv \right)^p \right)^{1/p},$$

$$\int_c^d \left(\int_a^b |X(u,v)|^p du \right)^{1/p} dv \le (\ge) \left(\int_a^b \left(\int_c^d |X(u,v)|\, dv \right)^p du \right)^{1/p},$$

when $p < 1$ (respectively, $p > 1$). (Advanced readers may substitute measure spaces for line segments $[a,b]$, $[c,d]$.)

*P9.19***

Establish **Young's inequality**: for a function $\varphi(x)$ continuous and strictly growing on $\{x \ge 0\}$, with $\varphi(0) = 0$,

$$ab \le \int_0^a \varphi(x)\, dx + \int_0^b \varphi^{-1}(y) dy \quad \left(\varphi^{-1} \text{ is inverse to } \varphi;\ \ a, b \ge 0\right),$$

with the equality achieved if and only if $b = \varphi(a)$ (Fig. 3).

A famous special case of Young's inequality corresponds to $\varphi = x^{p-1}$:

$$ab \le \frac{a^p}{p} + \frac{b^q}{q} \quad \left(a, b \ge 0;\ \ p>1;\ \ \frac{1}{p} + \frac{1}{q} = 1\right),$$

with equality when $a^p = b^q$. [*Verify* that also, this is a special case of the Cauchy inequality (section P9.12*).] *Establish* this inequality in a different way that makes it possible to *establish* the accompanying inequality

$$ab \geq \frac{a^p}{p} + \frac{b^q}{q} \qquad \left(a, b>0;\ p<1,\ p \neq 0;\ \frac{1}{p} + \frac{1}{q} = 1\right),$$

with the same cases of equality. *Derive* from these two inequalities a special case of the Hölder inequality (section P9.15**) and the accompanying one:

$$\sum x_i y_i \leq (\geq) \left(\sum x_i^p\right)^{1/p} \left(\sum y_i^q\right)^{1/q},\ \ x_i, y_i \geq 0,\ \ p>1\ \ (x_i, y_i>0,\ p<1,\ p \neq 0);$$
$$\frac{1}{p} + \frac{1}{q} = 1,$$

with equality when vectors x, y are proportional (in particular, if one of them is zero).

Establish another known special case of Young's inequality,

$$ab \leq a \log a - a + e^b \qquad (a \geq 1, \quad b \geq 0),$$

and *determine* the equality conditions.

*P9.20***

The source of Young's inequality is the **Legendre transform**. A convex function is called strictly convex when the Jensen inequality for it is strict (equality is achieved only for equal x_i). (*Provide* geometric formulations.) The Legendre image of a strictly convex function of one variable $f(x)$ is a function of a new variable $g(p)$ defined as follows: given a number p, let us consider a straight line $y = px$ in $\mathbb{R}^2$ and take a point $x = x(p)$ of the maximal distance, in the vertical direction, from this line to the graph of f (Fig. 4). *Show* that $x(p)$ is unique if it exists.

Analytically, $x(p)$ is a maximum point of a function $F(x, p) = px - f(x)$ with respect to x, keeping p fixed. *Derive* from this the generalized Young inequality

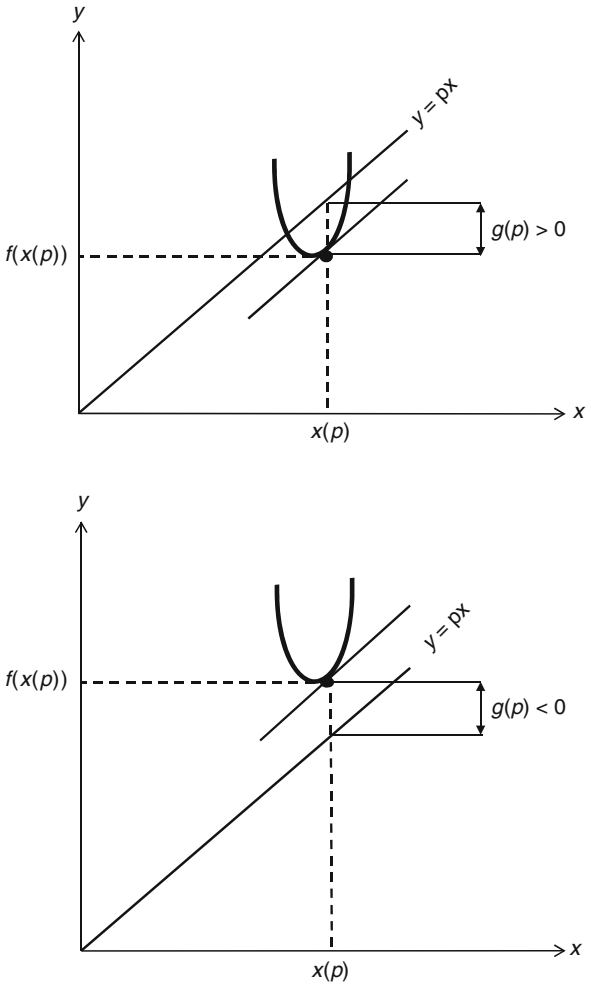
$$px \leq f(x) + g(p), \qquad \forall p, x \qquad (g(p) \text{ is the Legendre image of } f(x)).$$

Prove the convexity of the function $g(p)$.

Verify that for a differentiable f, $x(p)$ is defined by an equation $f'(x) = p$, that is, geometrically, a tangent to the graph of f at a point $(x(p), f(x(p)))$ is parallel to the line $y = px$.

Calculate the Legendre image of $f = kx^2$. *Show* that in this case $f(x(p)) = g(p)$. For *which* k is f equal to its Legendre image in the sense that substituting p for x in $f(x)$ yields $g(p)$?

Fig. 4 Legendre image



Prove that for a function $f(x) = \int_0^x \varphi \, dx$ with $\varphi(x)$ defined as in section P9.19**, the Legendre image is $g(p) = \int_0^p \varphi^{-1} \, dp$. Therefore, Young's inequality (section P9.19**) is a special case of the generalized Young inequality.

The Legendre transform in these examples[4] is an involution: if $g(p)$ is the Legendre image of $f(x)$, then the Legendre image of $g(p)$ once again is $f(x)$. *Prove* the same for any strictly convex differentiable function.

Legendre images of convex, but not strictly convex, functions are defined in the same way as above since the nonuniqueness of $x(p)$ in this case still allows us to define a valid $g(p)$. (*Why*?) For instance, the Legendre image of a linear function is a function on a one-point domain. (*Prove*.) What is the Legendre image of a continuous piecewise linear function (the graph of this function is a broken line)? *Verify* the involution property. Legendre images are defined for concave functions by substituting the minimum of $F(p,x)$ for its maximum. Moreover, Legendre images are defined for inflected functions, though this leads to cusped multifunctions: *calculate* the Legendre image of $y = x^3$. The involution property remains even in this case! (Advanced readers might have guessed that this property's source is the projective duality, as one can prove that the functions' duality by Legendre is equivalent to the projective duality of their graphs. Interested readers will find a discussion of the singularities, the multivariate version, and multiple applications of the Legendre transform and far-reaching developments in Courant and Hilbert (1953–1962), Arnol'd (1978, 1989, 2002), Arnol'd et al. (1982, 1985), and references therein.)

P9.21[*]

Carrying out a further study of convex functions, *prove* the following technical lemma.

Lemma *Let* $f(0) = 0$, *f be bounded above in a neighborhood D of* $x = 0$ *and have a property* $f(\alpha x) \geq \alpha f(x)$, $\forall \alpha \geq 1$, $\forall x$ *with* $\alpha x \in D$. *Then* $x \to 0$, $f(x) \geq 0 \Rightarrow f(x) \to 0$.

[Proving this lemma is achievable even for readers with very limited experience, but since it holds for multivariate functions and for functions on infinite-dimensional domains (in normed spaces) as well, more experienced readers should be sure to provide a dimension-independent proof.]

P9.22[*]

Establish continuity of convex functions of one variable at the internal points of their domains. *What* can be stated about their behavior at the endpoints? (See Fig. 5.)

[4] The first example is a special case of the second one.

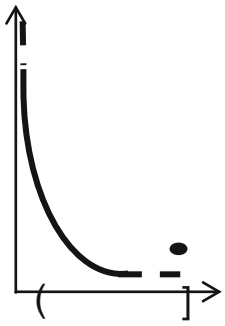


Fig. 5 Possible boundary behaviour of a convex function

P9.23[*]

A function of one variable defined to the right (left) side of a point x_0 (including x_0 itself) is referred to as a right (resp. left) side differentiable at that point if there exists a right-side derivative $f'_+(x_0) := \lim_{h \to +0} h^{-1}[f(x_0 + h) - f(x_0)]$ (resp. left-side derivative $f'_-(x_0) := \lim_{h \to -0} h^{-1}[f(x_0 + h) - f(x_0)]$). A function that has both one-sided derivatives at all internal points is called **one-sided differentiable**. *Establish* the one-sided differentiability of convex functions. (*What* can be stated about their behavior at endpoints?) *Deduce* from this a differentiability of a convex function, except at most at a countable number of points.

P9.24[***]

Show that *epi*(f) is a closed set for a continuous function f. Actually, closure of *epi* (f) is equivalent to a weaker condition than continuity. A real-valued function (on some topological space D) is referred to as **lower** (**upper**) **semicontinuous** at $x_0 \in D$ if for any $\varepsilon > 0, f(x) - f(x_0) \geq -\varepsilon$ [resp. $f(x) - f(x_0) \leq \varepsilon$] for neighboring points x. *Prove* that *epi*(f) is closed (in the topology of $D \times \mathbb{R}$) if and only if f is lower semicontinuous.

Obviously, lower plus upper semicontinuities equal continuity. *Provide* simple examples of discontinuous lower (or upper) continuous functions.

P9.25[***]

The famous **Hahn-Banach theorem**, usually taught to second- to third-year students, states, in one of its geometrical versions, that a closed convex set has

hyperplanes of support at all its face points.[5] Using this theorem, *show* that the epigraph of a continuous convex function has hyperplanes of support at the points of its graph. (This claim holds, including the convex functions on infinite-dimensional domains, as long as the Hahn-Banach theorem holds.)

The Hahn-Banach theorem in its complete geometric version states that a closed convex set is separated from a point outside it by a closed hyperplane. One of the implications is a definition of closed convex sets as intersections of closed half-spaces (determined by the hyperplanes of support). For a nonconvex set this intersection, taken over all hyperplanes of support, coincides with the set's convex hull, as discussed in section P9.5*. (The same result may also be obtained using the **Minkowski duality** technique discussed in sections H9.5 and E9.5.) This theorem finds practical application for image reconstruction in computerized tomography (see section P5.0 in the problem group "A Dynamical System with a Strange Attractor" and references therein). Consider a planar compactly supported binary distribution $f(x,y)$ (this means that the function f is defined on a plane, has only values 0 and 1, and has value 1 only on some compact set); we obtain such a distribution by extending the density of a homogeneous compact planar body by zero values at the external points. The **Radon image** of f consists of the set of integrals of this distribution calculated along a "radiation beam" – all possible affine straight lines. (Obviously, the value of each integral equals the length of the chord in the body cut out by the corresponding line.) The task of tomography is a reconstruction of the distribution f based on its Radon image. In real life a finite distribution $\{f_i\}$ as relating to the cells of a fixed finite partition of a bounded planar region is reconstructed, and the "beam" is quantized by a finite number of rays, but even after that, the numbers of both unknowns and of equations remain huge. However, these numbers can be significantly reduced – to a set of the pixels related to the convex hull of the f's support (f's nonzero set) for unknowns and a set of the rays intersecting with that hull for equations. These smaller subsets can be determined directly from the Radon image by the Hahn-Banach theorem; that is, we can determine the lines of support for $\{f \neq 0\}$ as straight lines l such that the preceding integrals are zero for all lines l' lying on one side of l and nonzero for l' lines lying on the opposite side of l. The described reduction in the number of unknowns and equations using the convex hull significantly increases the stability of the reconstruction algorithms. The same is true also for the reconstruction of nonbinary distributions (corresponding to inhomogeneous bodies), the three-dimensional reconstruction, etc. (Herman et al. 1987; Robinson and Roytvarf 2001).

*P9.26****

A function is referred to as **locally bounded below** (**above**) when it is bounded below (resp. above) in a neighborhood of any point. *Prove* that boundedness above of a convex function in a neighborhood of some internal point of its domain implies

(i) Upper semicontinuity at this point;
(ii) Lower semicontinuity at this point.

[5] In an infinite-dimensional case, the hyperplanes of support are required to be closed.

Therefore, a convex function locally bounded above in the interior of its domain is continuous there. Readers should *try* to provide a dimension-independent proof, as claims (i) and (ii) actually hold at least for the convex functions on any normed vector spaces.

Will similar results hold when the point is not internal?

The inversion of claim (i) is trivial for any convex and nonconvex functions (and the points do not have to be internal). The inversion of claim (ii) is not trivial, but for convex functions it is a true statement as stated in

Gelfand's lemma. *For a convex function, lower semicontinuity in the interior of its domain implies local boundedness above in the interior.*

Prove this statement. *Prove* that the gelfand's lemma holds also for any complete metric vector space [a vector space supplied with a **distance function** (**metric**) so that with respect to it the summation of vectors and multiplication of vectors by scalars become continuous operations, and the space is complete in the Cauchy sense].

*P9.27****

Show that a convex function f on a finite-dimensional (convex) domain is continuous in the interior of the domain. *Derive* from it that $epi(f)$ has hyperplanes of support at the points $(x, f(x))$ corresponding to internal points x.

*P9.28****

Prove that a univariate function f: $(a,b) \to \mathbb{R}$ allowing weights (as defined in section P9.6*) and at some point discontinuous is unbounded above in any interval $(c,d) \subseteq (a,b)$.

*P9.29****

A real-valued function on a vector space, that is subadditive, $f(x + y) \leq f(x) + f(y)$, and strongly positive homogeneous, $f(\alpha x) = |\alpha|\ f(x)$ $(\alpha \in \mathbb{R})$, is referred to as a **convex functional**, or **seminorm**. Examples: absolute values of the linear functionals, the norms. *Show* the following claims:

- Convex functionals are nonnegative-valued functions and $f(0) = 0$;
- Zero sets of convex functionals are vector subspaces (in general, not hyperplanes);
- Convex functionals are convex functions.

Establish that convex functionals (specifically, the norms) on finite-dimensional spaces are continuous. (This claim does not hold for infinite dimensions: as advanced readers probably know, there are discontinuous linear functionals on any infinite-dimensional metric vector spaces,[6] so taking absolute values of those functionals creates discontinuous convex functionals with everywhere dense zero sets.)

P9.30***

Verify that a pointwise supremum of convex functionals uniformly bounded on any vector is a convex functional. *Prove* that for continuous convex functionals this supremum itself is continuous.[7] *Provide* arguments actually proving the continuity of the supremum of continuous convex functions. *Be sure* to use arguments that are valid for a **Banach** (complete normed vector) **space**.

Hint

H9.1

Geometrically, a function f: $D \to \mathbb{R}$ on a convex domain D is convex when for $x,y \in D$, the segment $\{(z,f(z)): z = \alpha x + (1 - \alpha) y, 0 \le \alpha \le 1\}$ of its graph is not found above the line segment in $D \times \mathbb{R}$ connecting the points $(x, f(x))$, $(y, f(y))$.

[6] Actually, there are discontinuous linear functionals on any topological vector space (a vector space over either the real or the complex field of scalars supplied with a topology so that with respect to it the summation of vectors and multiplication of them by scalars become continuous operations) possessing a base of neighborhoods of zero of a cardinality not exceeding the algebraic dimension of the space. Indeed, following sources such as Kirillov and Gvishiani (1988), let $U \mapsto e_U$ be an injective map on this base to any fixed basis of our space and k_U be positive numbers such that $k_U e_U \in U$, $\forall U$ (k_U cannot be chosen without the famous **axiom of choice,** but of course we accept it). A linear functional f taking values ${k_U}^{-1}$ on the corresponding e_U and arbitrary values on the rest of the basis is discontinuous since the set $\{f < 1\}$, which is the preimage of (an open) neighborhood of zero in $\mathbb{R}$, includes no U and so cannot be a neighborhood of zero in our space. [We leave it to readers familiar with the elements of general topology to work out the details, including the fact that any neighborhood of zero in a topological vector space contains vectors of all directions (in fact, contains some half-interval $[0, \alpha_v v)$ of the rectilinear ray spanned on any vector v ($\alpha_v > 0$)), which follows from the continuity of the multiplication by scalars. We must warn readers, however, that the definition of a topological vector space uses a linguistic trick: a vector space of positive dimension with discrete topology (which means that every subset is open) is a topological Abelian group (with respect to the summation of vectors) but is not a topological vector space.]

[7] This theorem, due to I.M. Gelfand, is analogous to the **Banach-Steinhaus theorem** for linear operators.

Obviously, this property remains unchanged under the affine transformations of the domain and range (without changing its orientation). Also, use the geometric definition to show that the function that is simultaneously convex and concave is linear.

Neither convexity nor concavity is guaranteed for (1)–(4) (provide counterexamples); (5) for continuous monotonic convex f, f^{-1} is concave (convex) if and only if f is monotonically increasing (resp. decreasing); (6) $f \circ g$ is convex, provided that f does not decrease as well.

H9.2

The convex sets are arcwise connected (as discussed in section H8.6, "Polar and Singular Value Decomposition Theorems" problem group).

If points $x_1, \ldots$ of a convex set D in a finite-dimensional space do not lie in a proper subspace, a polyhedron of those vertices contained in D has a nonempty interior.

The convex subsets of a line are the line itself, the open and closed half-lines (rays), the open, half-open, and closed intervals, and the single-point sets.

H9.3

f is a convex function if and only if $epi(f)$ is a convex set.

H9.4

The Jensen inequality is proved by induction on n. (Provide the details.)

H9.5

The convex hull of a closed set is not necessarily closed. For example, a planar subset $\{(x,y)\colon x > 0, y = \pm x^{-1}\}$ is closed but has an open half-plane $\{x > 0\}$ as its convex hull. Prove that the convex hull of a compact (bounded and closed) set is compact.[8]

[8] Advanced readers probably know that in infinite-dimensional analysis, compactness does not mean the same as boundedness + closure. A topological space is compact if and only if every infinite subset has a limit point. Readers may prove this using the common definition of compactness that they are familiar with or find it in a textbook on general topology, for example, Hausdorff (1914) or Bourbaki (1960). Use this equivalent definition to prove the statement about the compactness of the convex hull of a compact set.

conv(M) is a convex set, so it contains the weighted means of the points of M. To establish the inverse inclusion, show that the set of the weighted means is convex. Use the fact that a weighted mean with relative weights γ, $1 - \gamma$ of two weighted means $\sum\alpha_x x$, $\sum\beta_x x$ is a weighted mean, because

$$\sum[\gamma\alpha_x + (1-\gamma)\beta_x] = \gamma\sum\alpha_x + (1-\gamma)\sum\beta_x = \gamma + 1 - \gamma = 1.$$

The sets that are finite intersections of closed half-spaces are obviously closures of their interiors, so a proof that the convex polyhedrons are polyhedrons may be completed by induction on dimension.

For a polyhedron M, *conv*(M) = *conv*(*vert*(M)). Indeed, any line segment between two points of M will meet the boundary of M when being rectilinearly extended across its ends, so the proof may be completed by induction on dimension. (Provide the details.)

By virtue of the inclusion *conv*(*vert*(G)) = *conv*(G) $\supseteq G$, a face G is contained in a plane if its vertices are also contained in that plane; hence, vertices cannot be contained in a plane of smaller dimension.

A determination of convex hulls and many related problems of Convex analysis can be solved using the **Minkowski duality** technique. Let us define for a nonempty subset of a Euclidean space $X \subseteq \mathbb{R}^n$ a dual set $X^* = \{p \in \mathbf{R}^n : \langle x, p\rangle \leq 1, \forall x \in X\}$. Obviously, $X^* \neq \varnothing$ (as $0 \in X^*$), $X \subseteq Y \Rightarrow Y^* \subseteq X^*$, $(X \cup Y)^* = X^* \cap Y^*$, $(X \cap Y)^* \supseteq X^* \cup Y^*$ (when $X \cap Y \neq \varnothing$) and $X \subseteq X^{**}$ (where $X^{**} := (X^*)^*$). What about the inverse inclusions? Which set is dual to a vector subspace and which to an affine subspace (a translation of a vector subspace)?

A. Verify the convexity (and closure) of X^*. Show that X^* has a nonempty interior if X is bounded (specifically, finite). Now turn to a finite X. We assume that X is not contained in a proper affine subspace (otherwise we will have a similar problem for a space of smaller dimension). We will call this finite set *good* if the spatial origin is contained in the interior of a simplex with vertices in X. (A proper translation of the origin makes a finite set good if this set is not contained in proper subspaces.) Show that X^* is a convex polyhedron if X is good.
B. Let S be a simplex with vertices in a set X and S° be S with vertices removed; prove that the points of S° play no part in defining X^*, that is, $X^* = (X \backslash S^\circ)^*$. We will call a point on the boundary of X a point of nonconvexity if it belongs to S° for a simplex S with vertices in X. We see that these points play no part in defining X^*. Verify that for a good set X the remaining points (the points of convexity) do play a part in defining X^*.
C. Show that for a convex polyhedron Y containing the origin in its interior Y^* is also a convex polyhedron and $Y^{**} = Y$. [*Famous classic examples*: a hexahedron (cube) is dual with an octahedron, a dodecahedron is dual with an icosahedron, and a tetrahedron's shape is autodual (Hilbert and Cohn-Vossen 1932).]
D. Lastly, prove, using the results of **A**–**C**, that for a good set X, $X^{**} = conv(X)$ and the vertices of X^{**} are the points of convexity in X.

The "fundamental theorems" follow directly from the geometric definition of convexity (of sets and functions).

H9.6

We suggest proving that the allowed weights are everywhere dense in a few steps as follows.

A. Verify that if f satisfies the Jensen inequality with a set of related weights $\{\alpha_i\}_{i\in I}$, then the sums $\sum_{i\in J}\alpha_i$ for any index subsets $\varnothing \neq J \underset{\neq}{\subset} I$ are allowed weights.

B. Verify that if f satisfies the Jensen inequalities with two sets of related weights $\{\alpha_i\}_{i\in I}$ and $\{\beta_i\}_{i\in J}$, then it satisfies this inequality with a set of related weights $\alpha_i\beta_j$. Therefore, the allowed weights form a multiplicative semigroup: α, β are allowed $\Rightarrow \alpha\beta$ is allowed.

C. Prove that if f satisfies the Jensen inequality with a set of related weights $\{\alpha_i\}_{i\in I}$, then it satisfies, for any nonempty subset $J \subseteq I$, a similar inequality with a set of related weights $\{\beta_i\}_{i\in J}$, where $\beta_j = \alpha_j \Big/ \sum_{k\in J}\alpha_k$. [This step is not completely trivial; readers can overcome the difficulties using arguments close to the **Cauchy downward induction technique** as discussed in Hardy et al. (1934) and Beckenbach and Bellman (1961).]

D. Show that if α_0 is an allowed weight, then f satisfies the Jensen inequality with a set of related weights $\alpha_{n,i,j}(\alpha_0) = \alpha_0^i(1-\alpha_0)^{n-i}$: $i = 0,\ldots, n$, $j = 1,\ldots,\binom{n}{i}$.

E. For $0 < i < n$, there are equal weights $\alpha_{n,i,j_1} = \alpha_{n,i,j_2}$. Applying paragraphs **C** and **A** find that f satisfies the Jensen inequality with relative weights ½, ½, so (by paragraph **B**), with relative weights $\underbrace{1/2^n,\ldots,1/2^n}_{2^n}$, and (by paragraph **C**), with relative weights $\underbrace{1/n,\ldots,1/n}_{n}$;; thus (by paragraph **A**), any rational weights are allowed. **QED**.

Allowing weights does not imply continuity. To provide counterexamples, consider additive functions f: $\mathbb{R} \to \mathbb{R}$, that is, functions with a property $f(x + y) = f(x) + f(y)$ (x,$y\in\mathbb{R}$). Verify that the additivity is the same as linearity over the rationals. Therefore, additive functions allow rational weights. The real field $\mathbb{R}$ is a vector space over the rational field $\mathbb{Q}$, so that the readers possess a variety of additive functions; find discontinuous functions among them. (***Warning***: the examples are constructed with the famous **axiom of choice**.) Also, verify that continuous additive functions are linear over the reals, whereas graphs of discontinuous additive functions are everywhere dense in the plane $\mathbf{R}^2_{x,y}$ (Gelbaum and Olmsted 1964).

Therefore, a discontinuous additive function is "very discontinuous" because it is unbounded in any interval. Actually, all univariate discontinuous functions allowing weights have the same property; see Problem P9.28*** below.

H9.7

The Jensen inequality holds for the integral sums, for $X(t)$ on the left-hand side and for $f(X(t))$ on the right, so passing on to the limit and taking the continuity of f into account yields the required solution.

H9.8

Show that any limit position of a secant of a graph of f passing through the points $(x_{\pm}, f(x_{\pm}))$ $(x_{-} \le x \le x_{+}, x_{-} < x_{+})$, as $x_{\pm} \to x \pm 0$, is a straight line of support for $epi(f)$ at point x, and establish the existence of this position. Vertical straight lines at internal points have a graph of f on both their sides, so they cannot be of support.

In a univariate case, if a point $(x, f(x))$ is placed above a straight line passing through the points $(x_{\pm}, f(x_{\pm}))$ $(x_{-} < x < x_{+})$, then $epi(f)$ cannot have a straight line of support at $(x, f(x))$ (provide a drawing), so f is convex if $epi(f)$ has straight lines of support at the points $(x, f(x))$ corresponding to internal x. A multivariate case can be reduced to a univariate one by considering the two-dimensional vertical cross sections of the space $D_x \times \mathbb{R}_f$, taking into account that the convexity of a function equals the convexity of all its restrictions to line segments of the domain, and the hyperplanes of support are formed by straight lines of support in those cross sections.

The local support condition does not provide convexity. (Readers can easily give counterexamples.)

H9.9

If each tangent hyperplane to a graph of f is of support for $epi(f)$, then f is convex by virtue of section P9.8*. Conversely, if a convex function $f: D \to \mathbb{R}$ defined on a convex domain is differentiable at an internal point x_0, then the tangent hyperplane to the graph at the point $(x_0, f(x_0))$ is the unique hyperplane of support for $epi(f)$ at that point. Indeed, in the univariate case, the tangent is the limit position of a secant, so it is the unique straight line of support, as follows from the arguments in section H9.8. The multivariate case can be reduced to the univariate one using the facts that the convexity of a function equals the convexity of all its restrictions to line segments of the domain and that the differential of the restriction of a

function to any subvariety in its domain equals the restriction of the differential to the vectors tangent to that subvariety [thus, the inequality $f(x) \geq f(x_0) + f'_{x_0}(x - x_0)$, $\forall x$ can be established by considering the restriction of f to the straight line in D, passing via x_0 and x]. (We leave it to the reader to work out all the details.)

The claim about critical points follows from the former claim by using the horizontality of the tangent plane to a graph at a critical point.

The claim about twice-differentiable functions also follows from the earlier claim using Taylor's formula.

H9.10

Verify that a substitution of variable $x \mapsto x^p$ $(x > 0)$ reduces the inequality to its special case $p = 1$. The desired inequality follows from the claim in section P9.7**, applied to the function $f(x) = x^q$ $(x > 0)$. This function is continuous, **strictly concave** for $0 < q < 1$, linear for $q = 1$, and **strictly convex** for the remaining q. Strict convexity (concavity) means that the Jensen inequality is strict, with the exception of equal x_i. The convexity (concavity) of $f(x)$ may be established by analyzing the second derivative (section P9.9**). Simple additional arguments show that $f'' > 0$ $(f'' < 0)$ is a sufficient but not necessary[9] condition for the strict convexity (resp. concavity) of a function f; $f = x^q$ satisfies this extra condition. Therefore, inequalities in section P9.10** turn into equalities only when all x_i are equal.

The inequalities for the weighted power means can be established with different approaches; some of them are quite elementary, usually requiring many computations. Those computations, however, can be simplified by the following observation: one may restrict oneself to a special case of equally weighted means (because this is actually equivalent to the general case; why?).

H9.12

An equality $\lim_{p \to 0, p \neq 0} \left(\sum \alpha_i x_i^p \right)^{1/p} = \prod x_i^{\alpha_i}$ is easily proved with l'Hôpital's rule applied to the logarithm of the left-hand side.

The Cauchy inequality between arithmetic and geometric means may also be obtained by directly applying the Jensen inequality (section P9.4*) to the function $f = \log x$. Interested readers will find in the book by Beckenbach and Bellman

[9] Readers may provide examples.

(1961) 12 proofs belonging to different mathematicians. Lev Radzivilovsky taught the following interesting geometric proof of the Cauchy inequality. (This proof was not published in the aforementioned book.) Consider the positive orthant in $\mathbb{R}^n$, defined by the inequalities $0 \le x_i$, $\forall$i, and inside it the pyramids $K_i(a)$, defined by two conditions: $x_i = \max(x_1,\ldots,x_n)$ and $x_i \le a$ (i = 1,..., n). Obviously, n equal pyramids of that kind form a cube of volume a^n; therefore, the volume of such a pyramid is a^n/n. Clearly, n pyramids $K_i(a_i)$ cover the parallelepiped $\{0 \le x_i \le a, \forall i\}$ (with positive a_i). (Provide figures for n = 2,3.) The volume of the parallelepiped is $\prod a_i$; the volume of $K_i(a_i)$, as was already found, is a_i^n/n. Thus we obtain the Cauchy inequality $\sum a_i^n/n \ge \prod a_i$. The same approach shows that the inequality becomes an equality if and only if all a_i are equal.

H9.13

Show that the modulus of a vector $\overrightarrow{OX}$, which is a convex linear combination of $\overrightarrow{OA}$ and $\overrightarrow{OB}$, equals $|OX| = \frac{|OA|\cdot|OB|\cdot\sin AOB}{|OA|\cdot\sin AOX+|OB|\cdot\sin XOB}$ (provide a figure), so for the bisector (if $AOX = \frac{1}{2}AOB$), $|OX| = H(|OA|,|OB|)\cdot\cos\frac{1}{2}AOB$. (Using more traditional argument of elementary geometry yields the same result.) This yields the desired form (*) of the inequality with the bisectors.

According to section P9.12*, (*) will hold if (**) holds; show, assuming the validity of (**), that equality in (*) takes place if and only if all R_i and all γ_i are equal.

To prove the inequality (***) for $n = 3$ without handling the general case, we suggest proceeding by any of the following two methods.

(1) Maximization of the left-hand side with respect to cos γ_i under fixed x_i using differential calculus. Calculate the values of a function $f = f_{a_1,a_2,a_n}(\gamma_1,\gamma_2,\gamma_3) := \sum a_i\cos\gamma_i$ at the critical points of the Lagrangian $L = L_{a_1,a_2,a_n}(\gamma_1,\gamma_2,\gamma_3,\lambda) := f + \lambda\sum\gamma_i$ that satisfy the condition $\sum\gamma_i = \pi$, and calculate a maximal value of f on the boundary of the triangle $\Delta = \{0 \le \gamma_i \le \pi/2,\ \sum\gamma_i = \pi\} \subset \mathbf{R}^3_{\gamma_1,\gamma_2,\gamma_3}$ (depict it!). (What does passing on the boundary correspond to, in the source geometric problem?) Find $f_{\max}$ among these values.

(2) Finding the maximal eigenvalue of the quadratic form in x_i on the left-hand side while the coefficients cos γ_i, are fixed. Show that $\sum\cos\gamma_i\cdot x_i x_{(i+1)\bmod n} - \frac{1}{2}\sum x_i^2$ is, for any $(\gamma_1,\gamma_2,\gamma_3)\in\Delta$, a negative semidefinite quadratic form with a one-dimensional kernel. Readers will address this by applying the famous Sylvester criterion to a matrix corresponding to this form and establishing the following trigonometric identity.

Lemma *A function* $\varphi = \sum\cos^2\gamma_i + 2\prod\cos\gamma_i$ *is constant (equal to one) on all planes* $\{\sum\gamma_i = (2k+1)\pi\}$ *(*$k\in\mathbb{Z}$*).*

Readers will learn, from method (2), that equality in (**) takes place for any $(\gamma_1,\gamma_2,\gamma_3)\in\Delta$ (with appropriate R_i). Because of this, (**) could not be sharpened by substituting positive power means for geometric means. (Why?)

The inequality (***) holds for any $n \geq 3$. However, it is difficult to attain a complete proof with method (1) or (2), including an enumeration of the cases of equality, for $n > 3$. Additional difficulties arise because of a routine growing computational complexity with n and other factors. We recommend that readers complete the proof for $n = 4$ using method (2). They will find the quadratic form $\sum \cos\gamma_i \cdot x_i x_{(i+1)\bmod n} - \frac{1}{2}\sum x_i^2$ to be negative semidefinite throughout the octahedron $\Delta = \{0 \leq \gamma_i \leq \pi/2,\ \sum\gamma_i = \pi\} \subset \mathbf{R}^4_{\gamma_1,\gamma_2,\gamma_3,\gamma_4}$ but for some $(\gamma_1,\gamma_2,\gamma_3,\gamma_4)\in\Delta$, negative definite (of zero kernel). For example, for four of six vertices of Δ, the maximal eigenvalue of the form $\sum \cos\gamma_i \cdot x_i x_{(i+1)\bmod n}$ is $\frac{1}{\sqrt{2}} = \cos\frac{\pi}{4}$, whereas for the other two vertices it is $\frac{1}{2} = \cos\frac{\pi}{3}$. (Determine those vertices.)

For $n > 3$, method (2) is fraught with tedious computations with trigonometric functions. Perhaps readers could suggest conceptual arguments replacing those computations. Meanwhile, we can suggest a different method to prove (***) (and, with complementary efforts, enumerating the cases of equality) for arbitrary $n \geq 3$. Evidently, we may assume $x_i \geq 0$, $\forall i$. The dependence on angles γ_i on the left-hand side can be removed with the use of a nice trick. Consider a set of planar vectors $\overrightarrow{r}_i = (u_i, v_i)$, $i = 1,\ldots, n$, of lengths $\left|\overrightarrow{r}_i\right| = x_i$, complemented by vector $\overrightarrow{r}_{n+1} = -\overrightarrow{r}_1$, such that $\left(\overrightarrow{r}_i, \overrightarrow{r}_{i+1}\right) = \gamma_i\,(i = 1,\ldots, n)$. (Provide a figure.) With the inequality from section P7.10** we find

$$\begin{aligned}\sum_{i=1}^{n} \cos\gamma_i \cdot x_i x_{(i+1)\bmod n} &= \left\langle \overrightarrow{r}_1, \overrightarrow{r}_2\right\rangle + \ldots + \left\langle \overrightarrow{r}_{n-1}, \overrightarrow{r}_n\right\rangle - \left\langle \overrightarrow{r}_n, \overrightarrow{r}_1\right\rangle \\ &= (u_1u_2 + \ldots + u_{n-1}u_n - u_nu_1) + (v_1v_2 + \ldots + v_{n-1}v_n - v_nv_1) \\ &\leq \cos\frac{\pi}{n}\sum_{i=1}^{n}\left(u_i^2 + v_i^2\right) = \cos\frac{\pi}{n}\sum_{i=1}^{n} x_i^2\,.\end{aligned}$$

H9.14

Find the limit when $p \to \infty$; then the second of the limits is obtained with help of the evident equality $M_{-p}(X) = (M_p(X))^{-1}$. Find $\lim\limits_{p\to\infty} M_p(X)$ first for the vectors and then for the functions using integral sums. The answer, both for the vectors and the functions, is $\lim\limits_{p\to\infty} M_p(X) = \max(X)$, $\lim\limits_{p\to-\infty} M_p(X) = \min(X)$ [if, of course, we retain a natural – for vectors – definition of max (X) as $\max\limits_{\alpha_i>0} x_i$].

H9.15

(Beckenbach and Bellman 1961; Hardy et al. 1934). If matrix (x_{ij}) contains zero rows, then the equality being proved is obvious; otherwise, we will derive

$$\frac{\sum\limits_j \prod\limits_i x_{ij}^{\alpha_i}}{\prod\limits_i \left(\sum\limits_j x_{ij}\right)^{\alpha_i}} = \sum_j \prod_i \left(\frac{x_{ij}}{\sum\limits_k x_{ik}}\right)^{\alpha_i} \le \sum_j \sum_i \alpha_i \frac{x_{ij}}{\sum\limits_k x_{ik}}$$

$$= \sum_i \alpha_i \sum_j \frac{x_{ij}}{\sum\limits_k x_{ik}} = \sum_i \alpha_i = 1,$$

with equalities in holding it if and only if $x_{ij} \Big/ \sum\limits_k x_{ik}$ does not depend on i for each j. Verify the equivalence of this condition to the proportionality of the rows of the matrix.

H9.16

Given a hyperplane $P = \{\langle x, v\rangle = h\}$ $(v \perp P)$, we will find, for a vector x ending on P, $|x|^2|v|^2 \ge \langle x, v\rangle^2 = h^2$. Therefore, $|x|_{\min} = |h|/|v|$. The minimum is reached for x proportional to v, that is, $x \perp P$.

H9.17

Let $h_1(x),\ldots, h_k(x)$ be oriented distances from $x \in \mathbb{R}^n$ to the polyhedral faces, which are taken with a "+" ("−") sign when a normal vector from x to the corresponding face (or its extension) reaches it from the side of the interior (resp. exterior) of the polyhedron. Prove the following basic lemma.

Lemma *Given a polyhedron* $P \subset \mathbf{R}_x^n$, *a map* $H : \begin{cases} \mathbf{R}_x^n \to \mathbf{R}_h^k, \\ x \mapsto (h_1(x), \ldots, h_k(x)) \end{cases}$ *is an affine isomorphism (a superposition of a linear isomorphism and a parallel translation) onto an n-dimensional plane* $\Pi_P^n \subset \mathbf{R}_h^k$.

Making $\mathbf{R}_h^k$ a Euclidean space so that the coordinates $h_1,\ldots, h_k$ are orthonormal, $m(P)$ will be the squared distance between the origin and Π_P^n, and $x_m(P)$ will be the orthogonal projection of the origin onto Π_P^n. This proves the unique existence of $x_m(P)$ and, since the plane Π_P^n and map $H^{-1} : \Pi_P^n \to \mathbf{R}_x^n$ may be explicitly

expressed via the faces of P, yields a numerically realizable algorithm for the determination of $m(P)$ and $x_m(P)$. (Fill in the details.)

Show the existence of $x_m(P)$ for any polyhedral-shaped bodies. Prove that the uniqueness holds for, and only for, bodies having at least one vertex. The half-spaces, the strips between parallel hyperplanes, and, generally, bodies with bounding hyperplanes K_i, such that $\bigcap_{j=1}^{l} K_{i_j} = \varnothing$ for $l \geq n$, have infinitely many points $x_m(P)$. (Where are they located?)

Prove the inclusion of the plane Π_P^n in a hyperplane $\Pi^{k-1} = \left\{ \sum_{i=1}^{k} A_i h_i = nV \right\}$. For a simplex P ($k = n + 1$), $\Pi_P^n = \Pi^{k-1}$, so (using, for example, section P9.16**) the solution of the source extreme problem is given by $h_i = nVA_i/\sum A_i^2$ ($i = 1,\ldots, k$). Therefore, $h_i > 0$, $\forall i$, meaning that $x_m(P) \in \text{int } P$. Provide examples of convex polygons with $x_m(P) \notin \text{int } P$.

Next, $f(P) = (nV)^2/\sum A_i^2$, which is the squared distance from the origin to Π^{k-1}, so it cannot exceed a squared distance to Π_P^n because $\Pi_P^n \subseteq \Pi^{k-1}$.

Examples of $m(P) = f(P)$ different from simplexes are given by the cubes and other polyhedrons possessing rich groups of orthogonal symmetries $G(P)$, so that for any two faces F_1, F_2, $\exists T \in G(P)$: $T(F_1) = T(F_2)$, and examples of $m(P) > f(P)$, are given by rectangular parallelepipeds that are not cubes. [Use that $x_m(P)$ is always a fixed point for the orthogonal symmetries of P. Also, remember that $k = 2n$ for the parallelepipeds, and Π_P^n may be defined in $\mathbf{R}_h^k$ by a system of equations $A_i \cdot (h_i + h_{i+n}) = V$, $i = 1,\ldots, n$, relative to the numbering of faces by which the $(i + n)$th face is kept parallel to the ith face.]

H9.18

The Minkowski inequality for a two-element set of indices j and $p = 2$ is a triangle inequality for Euclidean metrics. Hence, the Minkowski inequality with these parameters is equivalent to the CSB inequality (see section P9.16**).

The Minkowski inequality can be derived from the Hölder inequality by Riesz's method (Beckenbach and Bellman 1961; Hardy et al. 1934). We have, denoting $\sum_j x_{ij} = S_i$, $\left(\sum_j S_i^p \right)^{1/p} = S$, that

$$S^p = \sum S_i^p = \sum S_i S_i^{p-1} = \sum_j \sum_i x_{ij} S_i^{p-1},$$

and for $p > 1$, application of the Hölder inequality yields

$$S^p \le \sum_j \left(\sum_i x_{ij}^p\right)^{1/p} \left(\sum_i S_i^p\right)^{1/q} = \sum_j \left(\sum_i x_{ij}^p\right)^{1/p} S^{p-1},$$

where $q = p/(p-1)$ (or, equivalently, $p^{-1} + q^{-1} = 1$), which is equivalent to what is required. For $p < 1$ (and $x_{ij} > 0$), similar arguments using the accompanying inequality for the Hölder inequality (section P9.18**) are applicable.

H9.19

Young's inequality is a special case of a more general one derived with the Legendre transform (section P9.20**). Nevertheless, Young's inequality may be derived directly using similar geometric arguments as in the geometric proof of the Cauchy inequality discussed in section H9.12. That is, the summands on the right-hand side of Young's inequality are equal to areas of the curvilinear triangles on the plane $\mathbf{R}^2_{x,y}$ bounded by a curve $y = \varphi(x)$ and pairs of straight lines, resp. $\{x = a\}$, $\{y = 0\}$ and $\{y = b\}$, $\{x = 0\}$, while the left-hand side is equal to the area of a rectangle with these sides (Fig. 3).

A special case of Young's inequality for $\varphi = x^{p-1}$, together with the accompanying inequality, can be proved using differential calculus.

The special case of the Hölder inequality, and the accompanying inequality, may be derived as follows (Beckenbach and Bellman 1961): denoting $X = \sum x_i^p$, $a = x_i^p/X$, $Y = \sum y_i^q$, $b = y_i^q/Y$, applying Young's (accompanying) inequality, and summing the results we obtain

$$\frac{\sum x_i y_i}{X^{\frac{1}{p}} Y^{\frac{1}{q}}} \le (\ge) \quad \frac{1}{p} \cdot \frac{\sum x_i^p}{X} + \frac{1}{q} \cdot \frac{\sum y_i^q}{Y} = 1, \qquad p>1 (\text{resp. } p<1,\ p \ne 0).$$

QED.

The last of the inequalities presented in section P9.18** is a special case of Young's inequality corresponding to $\varphi(x) = \log(x + 1)$ and $a - 1$ substituted for a.

H9.20

The uniqueness of $x(p)$ (if it exists) follows from the strict convexity of f. (Furnish the details). The convexity of the function $g(p)$ is established as follows:

$$\begin{aligned} g(\alpha p + (1-\alpha)q) &= \max_x \left[(\alpha p + (1-\alpha)q)\cdot x - f(x)\right] \\ &= \max_x \left[\alpha(px - f(x)) + (1-\alpha)(qx - f(x))\right] \\ &\le \max_x \left[\alpha(px - f(x))\right] + \max_x \left[(1-\alpha)(qx - f(x))\right] \\ &= \alpha g(p) + (1-\alpha) g(q)\,. \end{aligned}$$

The Legendre image of $f = kx^2$ is $g = p^2/4\ k$; therefore, we have $x(p) = p/2\ k$, $g(p) = px(p) - f(x(p)) = f(x(p))$; on the other hand, $k = 1/4\ k$, and so f is equal to its Legendre image in the sense that substituting p for x in $f(x)$ brings $g(p)$ for $k = \pm ½$.

The statement that a function $f(x) = \int_0^x \varphi\, dx$ has a Legendre image $g(p) = \int_0^p \varphi^{-1}\, dp$ is equivalent to the equality $p\varphi^{-1}(p) - \int_0^{\varphi^{-1}(p)} \varphi(x)\, dx = \int_0^p \varphi^{-1}(y)\, dy$ (why?), which is simply the case of equality in Young's inequality.

The involution property of the Legendre transform for differentiable functions may be proved as follows. The Legendre image of $g(p)$ is $G(x,p(x))$, where $G(x, p) = xp - g(p)$ and $p(x)$ is determined by the equation $g'(p) = x$, so we must establish that $G(x,p(x)) = f(x)$. For this, verify that $G(x,p)$ has the following simple geometric meaning: it is the ordinate of a point with abscissa x on a tangent to the graph of f at the point $x(p)$. Thus, the equation $y = G(x,p)$ describes a one-parameter family of the tangents to the graph of f (parameterized with a slope p). This graph is the enveloping of this family, so we must prove that the enveloping must satisfy an equation $y = G(x,p(x))$. For a twice-differentiable f, the family of tangents is differentiable, and the usual methods for determining envelopings for differentiable families yield the required proof. Regardless of the twice-differentiability of f, the proof can be completed with simple geometric arguments using strict convexity. (The reader is invited to find those arguments.)

The Legendre images of piecewise linear convex (concave) functions are functions of the same kind, and the vertices and rectilinear segments of the graph of $f(x)$ correspond to, respectively, the rectilinear segments and vertices of the graph of $g(p)$. The Legendre image of a cubic $f = x^3$ is a semicubic $g = \pm 2\cdot(p/3)^{3/2}$.

H9.21

Let $x_n \to 0$ ($x_n \neq 0$), and a positive sequence γ_n also tends to zero, but slower (say, $\gamma_n = |x_n|^{1/2}$). Then $x_n/\gamma_n \to 0$; therefore, $x_n/\gamma_n \in D$, and so $f(x_n/\gamma_n) \le C$ for large n. But since ${\gamma_n}^{-1} \ge 1$ (for large n), $f(x_n)/\gamma_n \le f(x_n/\gamma_n)$; hence, $f(x_n) \le C\gamma_n$ (for large n), so $f(x_n) \to 0$ if, in addition, $f(x_n) \ge 0$.

H9.22

A convex function of a finite number of variables is bounded above at the internal points, because its values in the interior of a simplex do not exceed the maximum over the vertices. Continuity at the internal points follows from section P9.21*, taking into account the existence and nonverticality of the straight lines of support (section P9.8*).

Next, using some elementary arguments, show that a univariate convex function has a limit at each endpoint; moreover, $\lim f(x) \neq -\infty$ (values of a univariate convex function in a bounded domain cannot tend to $-\infty$; why?), and also, if $\lim f(x) < +\infty$, then the extension to the endpoint by continuity is convex.

H9.23

The following inequality follows from the definition of a convex function:

$$\frac{f(x)-f(x_1)}{x-x_1} \leq \frac{f(x_2)-f(x_1)}{x_2-x_1} \leq \frac{f(x_2)-f(x)}{x_2-x} \qquad (x_1<x_2)$$

(provide a figure). Prove that the quotient on the right-hand (resp. left-hand) side decreases, remaining bounded below, as $x_2 \to x$ (resp. increases, remaining bounded above, as $x_1 \to x$). From this, deduce, taking a limit as $x_1, x_2 \to x$, the existence of one-sided derivatives and the inequality $f'_-(x) \leq f'_+(x)$ for internal x, and, passing on to the limits as $x \to x_1$ and $x \to x_2$, the inequality $f'_+(x_1) \leq f'_-(x_2)$ for $x_1 < x_2$. (Therefore, one-sided derivatives at the internal points are finite, but at endpoints the cases of $f'_-(x_{right}) = +\infty$, $f'_+(x_{left}) = -\infty$ are not excluded; in particular, they take place if f is not continuous there.) From these inequalities derive that intervals on the ordinate axis $\left(f'_-(x), f'_+(x)\right)$ cannot intersect for distinct x, so there is at most a countable number of those intervals. (Why?)

From the preceding discussion, the derivative of a (differentiable) convex function on a real interval is monotonically nondecreasing. Prove a conversion: *a nondecreasing function (is integrable and) has its primitive convex.*

H9.24

Verify the meaning of openness of the complement $(D \times \mathbb{R}) \setminus epi(f)$: $\forall x_0 \in D$; if $y < f(x_0)$, then $\exists \varepsilon > 0$ such that $f(x) > y + \varepsilon$ for all x close to x_0. Show that this openness is equivalent to the lower semicontinuity of f.

H9.25

The claim immediately follows from section P9.24*** and the Hahn-Banach theorem.

H9.26

Prove claim (i) using the lemma from section P9.21*, and (ii) follows from (i) – for a convex function and an internal point of its domain.

If a point is not internal, then (ii) may be false (Fig. 5), but (i) will hold in the univariate case since for the endpoint $x, f(x) \geq \lim_{x' \to x} f(x')$ (work out the details using the Jensen inequality). What can be stated regarding (i) in a multivariate case?

You can prove Gelfand's lemma starting from the following statement.

Lemma A *convex function is locally bounded above on the interior of its domain if it is bounded above in a neighborhood of some internal point.*

(Try to find a proof of this claim that would be valid for any topological vector space.)

H9.27

A convex function of a finite number of variables is locally bounded above in the interior of its domain, as discussed in section H9.22. Its continuity in the interior follows from section P9.26***. The existence of the hyperplanes of support follows from sections P9.24*** and P9.25***.

H9.28

The conditions of the lemma from section H9.26 may be weakened by replacing the convexity with the allowed weights property (which is clear from the proof, as discussed in section E9.26): a function allowing weights is locally bounded above in the interior of its domain if it is bounded above in a neighborhood of an internal point. In turn, a function allowing weights and bounded above in a neighborhood of an internal point is continuous at that point, the proof of which is similar to the proof for convex functions (see section P9.26***). Another proof can be found in Hardy et al. (1934).

H9.29

The continuity of a convex functional in a finite-dimensional space follows from section P9.27***, because its domain is an open set (namely, the entire space).

H9.30

By virtue of the claims in section P9.26***, it is sufficient to establish the lower semicontinuity of the supremum.

Explanation

E9.1

Probe the counterexamples with (1) $f = x$, $g = x^2$; (2) $g = x^2 + 1$; (3) $f = x^4$ or $f = x^2 + \sin x$; and (4) $f = x^2$

E9.5

$X^{**} \neq X$ always when X is not convex, or is not closed, or does not contain the origin. (A famous Minkowski theorem states that, conversely, these three properties are sufficient for the equality $X^{**} = X$. This theorem is not so obvious, but the reader may easily verify that $X^{k*} = X^{(k+2)*}$ starting from $k = 1$.) In general, the inclusion $(X \cap Y)^* \subseteq X^* \cup Y^*$ also does not hold; for example, consider one-dimensional vector subspaces $X \neq Y$. A dual set to a vector subspace is an orthogonal vector subspace. (Why?) A dual set to an affine subspace $\{x\colon \langle x, p_i\rangle = 1, \forall i = 1,\ldots, n-k\}$ is an affine half-space $\{\sum\alpha_i p_i\colon \sum\alpha_i \leq 1\}$ of the orthogonal vector subspace (spanned on p_i). (Why?)

A. X^* is an intersection of closed half-spaces and is, therefore, a closed convex set. For a bounded X, X^* contains a neighborhood of the origin (why?) and, hence, has a nonempty interior. Verify that a dual set to a set of the vertices of a simplex that contains the origin in its interior is once again a simplex. Hence, for a good set X, X^* is a finite intersection of closed subspaces, has a nonempty interior, and is bounded (because it is contained in the simplex dual to a subset of X consisting of the vertices of a simplex), and so is a convex polyhedron.

B. To establish the equality $X^* = (X \backslash S^\circ)^*$, use the fact that the points of S° are weighted means of vertices of S. Now, let X be a good set and x one of its points of convexity. Find p belonging to $(X \backslash \{x\})^* \backslash X^*$; so x does plays a part in defining X^*.
C. Y^* is a convex polyhedron by virtue of paragraph **A** because $Y^* = (vert(Y))^*$. The equality $Y^{**} = Y$ is established as follows.

Lemma 1 (on barycentric coordinates). *$k + 1$ points $x_0,\ldots, x_k$ ($k \geq 0$) of a vector space that are not contained in a $k - 1$-dimensional affine plane are contained in a k-dimensional affine plane whose points are linear combinations of x_i with the sums of coefficients* 1.

Outline of proof. Those linear combinations shifted by $- x_0$ fill a k-dimensional vector subspace: $x = \sum \alpha_i x_i$, $\sum \alpha_i = 1 \Leftrightarrow x - x_0 = \sum_{i>0} \alpha_i (x_i - x_0)$ ($\alpha_1,\ldots, \alpha_k \in \mathbb{R}$). Complete the proof.

The parameters α_i are referred to as **barycentric coordinates** on this k-dimensional plane. (Readers may easily determine them via Cartesian (or other linear) coordinates of x and $x_0,\ldots,x_k$.)

For $\varnothing \neq X \subseteq \mathbb{R}^n$, let $\partial X^* := \{p \in \mathbf{R}^n : \langle x, p\rangle = 1,\ \forall x \in X\}$. Then we have $X \subseteq Y \Rightarrow \partial Y^* \subseteq \partial X^*$, $\partial (X \cup Y)^* = \partial X^* \cap \partial Y^*$, $\partial (X \cap Y)^* \supseteq \partial X^* \cup \partial Y^*$ (when $X \cap Y \neq \varnothing$).

Lemma 2 *For a k-dimensional affine plane X, ∂X^* is an $n - k - 1$-dimensional affine plane, and also $\partial (\partial X^*)^* = X$.*

Outline of proof. We have $X = x_0 + X_0$, where X_0 is a k-dimensional vector subspace orthogonal to the x_0. ∂X^* is determined by a system of $k + 1$ independent linear equations $\langle x_0, p\rangle = 1$, $\langle x_i, p\rangle = 0$, $i = 1,\ldots, k$, for a basis $\{x_i\}$ of X_0. The solution is given by $\partial X^* = |x_0|^{-2} \cdot x_0 + P$, where P is an orthogonal complement to a $k + 1$-dimensional vector subspace spanned on x_0, X_0.

Work out the details and verify the duality $\partial (\partial X^*)^* = x_0 + X_0$.

Corollary *For the vertices ($n - 1$-dimensional faces) G of a convex polyhedron Y containing the origin in its interior, ∂G^*, determine $n - 1$-dimensional faces (resp. vertices) of Y^*. [Similarly, one can prove that for the k-dimensional faces (edges) R, ∂R^* determine $n - k - 1$-dimensional faces of Y^*.]*

Outline of proof. For a vertex y, $\partial \{y\}^*$ takes part in defining a boundary of Y^* (according to paragraph **B**), and in addition, an entire neighborhood of a point $|y|^{-2} \cdot y$ in this hyperplane is contained in that boundary (provide a figure), so $\partial \{y\}^*$ determines an $n - 1$-dimensional face of Y^*. For an $n - 1$-dimensional face G of Y, the radius-vector of the point $p = \partial G^*$ is orthogonal to G and has a length equal to a distance from the origin to G, to the power -1 (from the proof of Lemma 2). Evidently, $\langle y, p\rangle \leq 1, \forall y \in Y$ (also, $\langle y, p\rangle < 1, \forall y \in Y \backslash G$; provide a figure!), so $p \in Y^*$ and is placed on its boundary. Now, let $y_1,\ldots,y_n$ be vertices contained in G but not in a plane of a smaller dimension (such vertices exist by section H9.5). The hyperplanes $\partial \{y_i\}^*$ intersect at a point p. (Why? Use Lemma 1 for the answer.) Since $p \in Y^*$ and because of the convexity of Y^*, p is an intersection of the faces of Y^*

determined by these hyperplanes (provide a figure!), so it is a vertex. Furnish the details.

By this corollary, for the vertices $p = \partial G^*$ of Y^*, where G are $n-1$-dimensional faces of Y, the hyperplanes $\partial\{p\}^*$ determine $n-1$-dimensional faces G^{**} in Y^{**}. By Lemma 2, these hyperplanes determine the faces G of Y, and the neighborhoods of the points $p/|p|^2$ in G and G^{**} coincide. The polyhedron Y cannot be extended without violating this condition or convexity (provide a figure), so $Y^{**} = Y$. **QED**.

D. From **B**, the $n-1$-dimensional faces of X^* are determined by hyperplanes $\partial\{x\}^*$ corresponding to the points of convexity in X. Therefore, by the proof of the corollary in **C**, the points of convexity in X are the vertices of X^{**}. (Fill in the details.)

E9.6

C. We have

$$\begin{aligned} f\left(\sum_{j\in J}\frac{\alpha_j x_j}{\sum\limits_{k\in J}\alpha_k}\right) &= f\left(\sum_{j\in J}\alpha_j x_j + \left(1-\sum_{k\in J}\alpha_k\right)\cdot\frac{\sum\limits_{j\in J}\alpha_j x_j}{\sum\limits_{k\in J}\alpha_k}\right) \\ &= f\left(\sum_{j\in J}\alpha_j x_j + \sum_{i\notin J}\alpha_i\cdot\frac{\sum\limits_{j\in J}\alpha_j x_j}{\sum\limits_{k\in J}\alpha_k}\right) \\ &\le \sum_{j\in J}\alpha_j f(x_j) + \sum_{i\notin J}\alpha_i\cdot f\left(\sum_{j\in J}\frac{\alpha_j x_j}{\sum\limits_{k\in J}\alpha_k}\right) \\ &= \sum_{j\in J}\alpha_j f(x_j) + \left(1-\sum_{k\in J}\alpha_k\right)\cdot f\left(\sum_{j\in J}\frac{\alpha_j x_j}{\sum\limits_{k\in J}\alpha_k}\right), \end{aligned}$$

and after a little computation,

$$\left(\sum_{k\in J}\alpha_k\right)\cdot f\left(\sum_{j\in J}\frac{\alpha_j x_j}{\sum\limits_{k\in J}\alpha_k}\right) \le \sum_{j\in J}\alpha_j f(x_j).$$

QED.

Continuous additive functions are monotone or totally zero; hence, zero sets different from $\{0\}$ or $\mathbb{R}$ are inadmissible. Therefore, fixing a basis of $\mathbb{R}$ over $\mathbb{Q}$, setting f equal to zero on some basis element, one on some other and arbitrarily on

the remaining elements, and extending this to all of $\mathbb{R}$ by linearity over $\mathbb{Q}$ produces a discontinuous additive function. (Where is the axiom of choice referred to in this argument?) A closure of the graph of an additive function is a vector subspace of $\mathbf{R}^2_{x,y}$ over the reals (why?), so it may be a nonvertical straight line $y = kx$ or the whole plane. The first possibility relates to continuous functions only. (Why?) Thus, the graphs of discontinuous additive functions are everywhere dense in the plane.

E9.8

Write down an equation $y = kx + b$ of a secant of a graph of f passing through the points $(x_\pm, f(x_\pm))$ $(x_1 \le x_- \le x \le x_+ \le x_2,\ x_- < x_+)$. Using the fact that a point $(x, f(x))$ cannot be placed above this line, show that parameters k and b are bounded. Hence, there are limit positions of a secant as $x_\pm \to x \pm 0$. Take one of them. Prove that the graph of f to the left of x_- and to the right of x_+ cannot go below the secant, and derive from it that for $x' \ne x, f(x')$ is not less than the ordinate of a point on the limit straight line with abscissa x'. Therefore, for the internal x, the limit line is not vertical (lim $k \ne \pm\infty$). Lastly, since a point $(x, f(x))$ cannot be placed above this line (why?), or below it (why?), it must be a line of support for *epi*(f) at point x.

E9.9

For a twice-differentiable f write down a Taylor series exapnsion

$$f(x_0 + h) - f(x_0) - f'_{x_0}(h) = f''_{x_0}(h, h) + o\big(|h|^2\big) \qquad \text{for } h \to 0.$$

If $f''_{x_0}(h, h) < 0$ for some h, then we will have, for small t, $f(x_0 + th) - f(x_0) - f'_{x_0}(th) < 0$ (why?), so f cannot be convex (by the claim just proved). Conversely, let $f''_x \ge 0$. We must verify the convexity of f on the line segments in the domain, so we can let it be a univariate function. Considering $f(x) - f(x_0) - f'_{x_0}(x - x_0)$ in place of $f(x)$, which affects neither the second differentials nor the convexity of f (why?), we will have $f(x_0) = 0$ and must show that $f(x) \ge 0$. This is attained by representing

$$f(x) = \int_{x_0}^{x} du \int_{x_0}^{u} f''_v \, dv,$$

which indeed is a nonnegative value. (Why? Provide the details.)

E9.13

Assuming the validity of (**), and, taking into account that the inequality between different power means is strict unless all R_i are equal to each other, deduce that equality in (*) can hold only for equal R_i. In turn, for equal R_i, (*) takes the form of $\frac{1}{n}\sum \cos\gamma_i \le \cos\frac{\pi}{n}$, which is a strict inequality unless all γ_i are equal to each other, as cosine is strictly concave on $[0, \pi/2]$ (see section H9.10).

Proving (**) for $n = 3$ by method (1). The critical points of L are determined by a system of equations

$$\sum \gamma_i = \pi, \qquad 0 = \partial f/\partial\gamma_i = -a_i \sin\gamma_i + \lambda \quad (i = 1, \ldots, n).$$

For $n = 3$ calculating $f_{critical}$ might avoid having to find λ. We have

$$f_{critical} = \sum a_i \sqrt{1 - (\lambda/a_i)^2} = \sum \sqrt{a_i^2 - \lambda^2}.$$

Since $\sin\gamma_i = \sin\sum_{j\ne i}\gamma_j$ we obtain, using the equations defining a critical point,

$$\frac{\lambda}{a_i} = \frac{\lambda}{a_{(i+1)\bmod 3}\cdot a_{(i+2)\bmod 3}} \cdot \left(\sqrt{a_{(i+1)\bmod 3}^2 - \lambda^2} + \sqrt{a_{(i+2)\bmod 3}^2 - \lambda^2}\right) \qquad (i = 1, 2, 3).$$

Summation of these equations gives $f_{critical} = \frac{1}{2}\left(\frac{a_2a_3}{a_1} + \frac{a_3a_1}{a_2} + \frac{a_1a_2}{a_3}\right)$.. Specifically, for $a_i = \sqrt{R_iR_{(i+1)\bmod 3}}$ we have $f_{critical} = \frac{R_1+R_2+R_3}{2}$, that is, in a critical point (if it exists) we will have equality in (**). On the boundary of Δ we may assume, without loss of generality, that $\gamma_1 + \gamma_2 = \gamma_3 = \pi/2$. A short computation using the CSB inequality yields

$$\begin{aligned}\sqrt{R_1R_2}\cos\gamma_1 + \sqrt{R_2R_3}\cos\gamma_2 &\le \sqrt{R_1R_2 + R_2R_3}\cdot\sqrt{\cos^2\gamma_1 + \cos^2\gamma_2}\\ &= \sqrt{R_1R_2 + R_2R_3} \le \tfrac{R_1+R_2+R_3}{2}\end{aligned}$$

[where the last inequality is equivalent to that of $(R_1 - R_2 + R_3)^2 \ge 0$], which completes the proof.

Proving (***) for $n = 3$ by method (2). The matrix in the usual coordinate basis corresponding to the considered quadratic form is $\frac{1}{2}\begin{pmatrix} -1 & c_1 & c_3 \\ c_1 & -1 & c_2 \\ c_3 & c_2 & -1 \end{pmatrix}$ ($c_i = \cos\gamma_i$). For the first two corner minors we have $\det\frac{1}{2}(-1) = -1/2 < 0$, $\det\frac{1}{2}$

$\begin{pmatrix} -1 & c_1 \\ c_1 & -1 \end{pmatrix} = \frac{1}{4}\left(1 - c_1^2\right) \geq 0$. Equality in the last inequality corresponds to $\gamma_1 = 0$, $\gamma_2 = \gamma_3 = \pi/2$, so our quadratic form turns into $x_1 x_2 - \frac{x_1^2 + x_2^2 + x_3^2}{2}$, which is nonpositive and has a one-dimensional kernel (which one?). Finally, the determinant of the whole matrix equals $\frac{1}{8}\left(-1 + \sum c_i^2 + 2\prod c_i\right)$, which, referring to the lemma from section H9.13, finalizes the proof. (Work out the details.) In turn, this lemma may be proved as follows: $\varphi = 1$ for $\gamma_1 = 0$, $\gamma_2 = \gamma_3 = (k + ½)\,\pi$. Next, a short computation with a routine trigonometric combinatorics shows that $\partial\varphi/\partial\gamma_1 = \partial\varphi/\partial\gamma_2 = \partial\varphi/\partial\gamma_3$ on the plane $\{\sum\gamma_i = (2\,k + 1)\pi\}$. On the other hand, for vectors $\alpha = (\alpha_1,\alpha_2,\alpha_3)$ tangent to this plane, $\sum\alpha_i = 0$. Therefore, for a derivative along any such vector we find

$$\partial\varphi/\partial\alpha = \sum \alpha_i \partial\varphi/\partial\gamma_i = \partial\varphi/\partial\gamma_1 \sum \alpha_i = 0.$$

QED. (We leave it to the reader to fill in the details.)

E9.14

We have

$$\left(\sum \alpha_i x_i^p\right)^{1/p} \leq \left(\sum \alpha_i \max x_i^p\right)^{1/p} \leq \max x_i \cdot \left(\sum \alpha_i\right)^{1/p} = \max x_i;$$

on the other hand, if $\max(X) = x_{i_0}$, then we will have

$$\left(\sum \alpha_i x_i^p\right)^{1/p} = \alpha_{i_0}^{1/p} x_{i_0} \left(\sum \frac{\alpha_i x_i^p}{\alpha_{i_0} x_{i_0}^p}\right)^{1/p} \geq \alpha_{i_0}^{1/p} x_{i_0},$$

because the value in the brackets is not less than 1. The limit of the right-hand side is x_{i_0}. **QED**. Now consider, for the function $X(t)$, a two-indexed sequence of integral sums $M_{p,n} := \left(2^{-n} \sum_{i=1}^{2^n} x_i^p\right)^{1/p}$, with $x_i = \min_{\frac{i-1}{2^n} \leq t \leq \frac{i}{2^n}} X(t)$. From sections P9.10** and P9.12*, it monotonically converges on p, for fixed $n \leq \infty$ $\left[M_{p,\infty} = \lim_{n\to\infty}\left(2^{-n}\sum_{i=1}^{2^n} x_i^p\right)^{1/p} = \left(\int_0^1 [X(t)]^p dt\right)^{1/p}\right]$. Also, it monotonically converges on n, for fixed $0 \leq p \leq \infty$ ($M_{\infty,n} = \max_i \min_{\frac{i-1}{2^n} \leq t \leq \frac{i}{2^n}} X(t)$). By virtue

of the continuity of $X(t)$, $\lim_{n\to\infty} M_{\infty,n} = \max(X)$. Therefore, we must establish that $\lim_{p\to\infty} M_{p,\infty} = \lim_{n\to\infty} M_{\infty,n}$. Lastly, this equality holds since both its sides are equal to $\sup_{p,n<\infty} M_{p,n}$, completing the proof.

E9.15

If the rows of matrix (x_{ij}) are proportional, then x_{ij}/x_{ik} depends only on j and k, and thus $x_{ij}\Big/\sum_k x_{ik}$ depends only on j. Conversely, term-by-term division of the equalities $x_{i_1 j_1}\Big/\sum_k x_{i_1 k} = x_{i_2 j_1}\Big/\sum_k x_{i_2 k}$, $x_{i_1 j_2}\Big/\sum_k x_{i_1 k} = x_{i_2 j_2}\Big/\sum_k x_{i_2 k}$ yields the desired proportionality $x_{i_1 j_1}/x_{i_2 j_1} = x_{i_1 j_2}/x_{i_2 j_2}$. **QED**.

E9.17

Proof of the lemma. Let the hyperplanes determining the faces of P be defined by the equations $\{\langle x, v_i\rangle = c_i\}$ $(i = 1,\ldots, k)$. Then $h_i(x) = \frac{\langle x, v_i\rangle - c_i}{|v_i|}$, so H is an affine map. Without loss of generality, let indices $1,\ldots,n$ correspond to faces in a general position [which means that $F_1,\ldots,F_n$ have a point intersection (a vertex)].[10] Then, the right-hand sides of the first,$\ldots,n^{\text{th}}$ of the foregoing k equations will be linearly independent, and so the remaining equations determining the plane im H will be linear combinations of these n. Hence, dim im $H = k - (n - k) = n$. **QED**. (A nonlinear version of this argument yields a proof of the implicit function theorem.)

The existence of $x_m(P)$ for any polyhedral-shaped bodies follows from the finiteness of the number of faces; indeed, quadratic forms in finite-dimensional spaces are brought to principal axes; hence, obviously, positive semidefinite quadratic forms have minima. If P does not have a vertex, then it contains an m-dimensional $(m > 0)$ subspace K parallel to all its faces (why?); hence, a set of the points $x_m(P)$ contains an m-dimensional affine plane $x_0 + K$. Readers can prove that $\dim\{x_m(P)\}$ is actually equal to the minimum of the dimensions of all nonempty intersections of P's faces.

Denote by $S = S(x,P)$ an algebraic sum of volumes of pyramids with the vertex x based on the faces of P; in this sum, the pyramid's volume is denoted by a "+"

[10] Hyperplanes $F_1,\ldots,F_m$ are in a general position when codim $\bigcap F_i = m$, or, in other words (why?), the vectors orthogonal to F_i are linearly independent. Among the faces containing a vertex, there are always n faces in a general position, because otherwise the intersection would have a positive dimension and not be a vertex.

("−") sign if the height of the pyramid reaches the base from the side of the interior (resp. exterior) of the polyhedron. Geometrically, the inclusions $\Pi_P^n = im\ H \subseteq \Pi^{k-1}$ for all polyhedrons P mean simply the identical equality of $S(x,P)$ to P's volume: $S = V, \forall x,P$. This equality, almost obvious for simplexes, in fact expresses a fundamental topological feature allowing generalizations. Readers may prove this equality as follows. Let P first be a simplex. S is affine in x (why?) and, obviously, $S = V$ for $x \in P$. In particular, S is constant on an open subset of $\mathbb{R}^n$, so it is globally constant. (Why?) Next, verify the additivity of S with respect to polyhedrons: $S(x, P_1 \cup P_2) = S(x, P_1) + S(x, P_2)$ when P_1, P_2 intersect at most by their boundaries. The proof is completed by applying the theorem that polyhedrons allow for **triangulation**, that is, can be represented as unions of simplexes intersecting at most by their boundaries. (A proof of the triangulation theorem is beyond the scope of this book, so readers not familiar with this theorem may take us at our word or use a handbook on topology.)

The remaining claims in section H9.17 can be proved straightforwardly.

E9.19

(Beckenbach and Bellman 1961). The substitution $x = a/b^{q/p}$ converts Young's inequality corresponding to $\varphi = x^{p-1}$ and the accompanying inequality into the inequalities $x \leq (\geq) \frac{x^p}{p} + \frac{1}{q}$ $(x > 0)$. Consider a function $F(x) := x - \frac{x^p}{p} - \frac{1}{q}$ on $\mathbb{R}^{+*} = \{x > 0\}$. We have $F(1) = F'(1) = 0$, and $F''(x)$ is positive (negative) when $p < 1$ (resp. $p > 1$). Therefore, F behaves like a quadratic function (provide a figure) and, in particular, has a unique maximum (resp. minimum) at $x = 1$ and is negative (resp. positive) at the rest of the points. **QED**.

E9.20

Finalization of proof of involution property of Legendre transform (Arnol'd 1989). We must show that $G(x,p(x)) = f(x)$, without applying the second f's derivative. Fix $x = x_0$ and vary p. The values of $G(x,p)$ will be the ordinates of intersections of tangents to the graph of f with a vertical line $x = x_0$. Those tangents go below this graph (as discussed in section P9.9*), so the maximum of $G(x,p)$, for a fixed $x_0 = x$ (p_0), is equal to $f(x_0)$ [and is attained for $p_0 = p(x_0) = f'(x_0)$]. **QED**. (Provide a figure.)

E9.22

Without loss of generality (as discussed in section P9.1*), we suggest dealing with $x = 0$ for the internal point. Next, without affecting the claim being proved, we may subtract from f a linear function that has a graph that coincides with the hyperplane of support for f's graph at the point $(0, f(0))$, so we have $f \geq f(0) = 0$. Thus, we must show that $x \to 0 \Rightarrow f(x) \to 0$. Using Jensen's inequality, $f(\alpha x) \leq \alpha f(x)$, $\forall x \in D$, $\forall \alpha \in [0,1]$, so $f(\alpha x) \geq \alpha f(x)$, $\forall \alpha \geq 1$, $\forall x$ *with* $\alpha x \in D$. (Why?) Now the proof is complete with the lemma from section P9.21*.

Also, continuity of the univariate convex functions in the interiors of their domains may be derived as a special case of the claim in section P9.27*** below. We suggest experienced readers pay attention to the nuances of this claim's proof.

E9.24

If, for a lower semicontinuous f, for some $x_0 \in D$ and $y < f(x_0)$, $\forall \varepsilon > 0$ $\exists$ x arbitrarily close to x_0 with $f(x) \leq y + \varepsilon$, then we will have a contradiction because $f(x_0) \leq f(x) + \varepsilon \leq y + 2\varepsilon$, so $f(x_0) \leq y$ by virtue of the arbitrariness of ε. Conversely, if $(\mathrm{D} \times \mathbb{R}) \backslash epi(f)$ is open but for some $x_0 \in D$ and $\varepsilon > 0$ $\exists$ x arbitrarily close to x_0 with $f(x) \leq f(x_0) - \varepsilon$, then we will have a contradiction because $f(x) \geq f(x_0)$.

E9.26

Proving claim (i) we must show that $f(x) - f(x_0) \leq \varepsilon$, for $x_0 \in D$, $\varepsilon > 0$, and x close to x_0, which is trivial if $f(x) < f(x_0)$ and for $f(x) \geq f(x_0)$ follows from the lemma in section P9.21* applied (as in section E9.22) to a function $f' = f - f(x_0)$ of an argument $x' = x - x_0$ instead of f [which is applicable because, due to convexity, $f'(\alpha x') \leq \alpha f'(x')$ when $0 \leq \alpha \leq 1$, or – with the substitution of $\alpha x'$ and $\alpha' = \alpha^{-1}$ for, respectively, x' and $\alpha - \alpha' f'(x') \leq f'(\alpha' x')$ when $1 \leq \alpha'$]. (We leave it to the reader to work out all details.)

Proof of the fact that for a convex function and an internal point of its domain, (i) $\Rightarrow$ (ii). For $x \to x_0$ such that $f(x) \leq f(x_0)$, we have $x' \to x_0$ and $2f(x_0) \leq f(x) + f(x')$, where $x' = 2x_0 - x$; hence, $f(x_0) \leq f(x')$ and $0 \leq f(x_0) - f(x) \leq f(x') - f(x_0) \to 0$, and so $f(x) \to f(x_0)$. (We leave it to the reader to fill in the details.)

In the multivariate case, (i) may be false if the point is not internal. (Find an example with two variables.)

Proof of lemma from section H9.26. Let f be bounded above on an open neighborhood of the origin U. Then f will be bounded on a neighborhood of a point $x \in \operatorname{int} D$, that is, on $(x + n^{-1}U) \cap D$ with a proper natural n. Indeed, for a large n so that $nx/(n-1) \in D$ we find

$$f\left(x+\frac{h}{n}\right)=f\left(\underbrace{\frac{1}{n}\cdot\frac{nx}{n-1}+\ldots+\frac{1}{n}\cdot\frac{nx}{n-1}}_{n-1}+\frac{h}{n}\right)\le\frac{n-1}{n}\cdot f\left(\frac{nx}{n-1}\right)+\frac{1}{n}\cdot f(h)$$

with $h\in U$, which completes the proof.

Now complete the proof of Gelfand's lemma using the following argument: if, for a convex function f, lower semicontinuous on the interior, Gelfand's lemma is false, then one can choose with the lemma from section H9.26 a sequence of closed balls $B_1 \supseteq B_2 \supseteq \ldots$ of radii $\le 1, 2^{-1},\ldots$ such that $f|_{B_n} > n$, so f could not obtain a finite value on the intersection $\bigcap B_n$.

E9.28

The implication "*boundedness above* $\Rightarrow$ *upper semicontinuity at internal points*" is proved with a lemma similar to that from section P9.21*.

Lemma *Let* $f(0)=0$, f *be bounded above in a neighborhood of* $x=0$ *and* $f(\alpha x)\ge\alpha f(x)$, $\forall\alpha\in\mathbb{Q}$, $\alpha\ge 1$, $\forall x$ *with* $\alpha x\in D$. *Then* $x\to 0, f(x)\ge 0\Rightarrow f(x)\to 0$.

(Work out the details.) The proof of the implication of "*boundedness above* $\Rightarrow$ *lower semicontinuity at internal points*" is similar to that in section E9.26 because x' can be selected such that $a=|x_0-x'|\in[a_0-\sigma, a_0+\sigma]$ with a rational quotient $a{:}\delta$. (Work out the details.)

E9.30

For $f(x)=\sup f_i(x)$ and a point x_0, let i be such that $f_i(x_0)\ge f(x_0)-\varepsilon$. Then for x close to x_0, so that $f_i(x)\ge f_i(x_0)-\varepsilon$, we have

$$f(x)-f(x_0)\ \ge\ f_i(x)-f(x_0)\ \ge\ f_i(x)-f_i(x_0)-\varepsilon\ \ \ge -2\varepsilon,$$

which completes the proof.

Completing the Solution

S9.2

If the points $x_0,\ldots,x_n$ of a convex set $D\subseteq\mathbb{R}^k$ do not lie in a proper subspace, then $n\ge k$, and there are $k+1$ points among them, say, $x_0,\ldots,x_k$, that do not lie in a proper subspace. (Why?) Let S be the simplex of those vertices. Show that $D\supseteq S$,

using an analytic definition of a simplex as the **convex hull** of its vertices [which are $k + 1$ points in a general position (not contained in a hyperplane)], that is, the set of all linear combinations $\sum_{i=0}^{k} \alpha_i x_i$ with $1 \geq \alpha_i \geq 0$ and $\sum \alpha_i = 1$ (x_i are the vertices). (Such linear combinations are referred to as **convex**; the reader may take a look at sections P9.5* and H9.5.) In turn, a simplex has a nonempty interior, as follows from the next lemma.

Lemma 1 *The points of a simplex with $\alpha_i > 0$, $\forall i$ (and only those points) are internal.*

Proof of Lemma 1. Convex linear combinations may be presented as $x_0 + \sum_{j=1}^{k} \beta_j (x_j - x_0)$ with $1 \geq \beta_j \geq 0$ and $\sum \beta_j \leq 1$. Because the points $x_0,\ldots,x_k$ are the vertices of a simplex, the vectors $v_j = x_j - x_0$ are linearly independent, so the map $\begin{cases} \mathbf{R}^k & \to & \mathbf{R}^k \\ (\beta_1, \ldots, \beta_k) & \mapsto & x_0 + \sum \beta_j v_j \end{cases}$ is an affine isomorphism, and so it is a homeomorphism. Therefore, since for a k-element sequence $(\beta_1,\ldots, \beta_k)$ of $\beta_j > 0$, $\forall j$, and $\sum \beta_j < 1$ all close sequences $(\beta_1',\ldots, \beta_k')$ have the same properties, the points $x_0 + \sum \beta_j' v_j$ belong to the simplex's interior. On the other hand, if $(\beta_1,\ldots, \beta_k)$ does not have some of those properties, then there are arbitrarily close sequences $(\beta_1',\ldots, \beta_k')$ such that the points $x_0 + \sum \beta_j' v_j$ are not convex linear combinations of x_i, so they are not points of the simplex. **QED**. (We leave it to the reader to fill in the details.)

Also, readers may find the simplex's interior by employing a geometric definition of a simplex in $\mathbb{R}^k$ (or of dimension k) as a bounded set, which is an intersection of $k + 1$ closed affine half-spaces, such that the hyperplanes that are the borders of each k are in a general position (have a point intersection). Therefore, to proceed by this method, we must establish the following lemma.

Lemma 2 *The preceding analytic and geometric definitions of a simplex are equivalent to each other.*[11]

Proof of Lemma 2. An intersection of closed half-spaces is a convex set (as discussed in section P9.2*). Therefore, we must verify that the points of a "geometric" simplex $S = \bigcap_{i=0}^{k} P_i$ are convex linear combinations of $k + 1$ vertices $x_0,\ldots,x_k$, which are defined as the points of intersections of k hyperplanar borders ∂P_i. Apply an induction on the spatial dimension. Since the interior of an intersection is the intersection of the interiors, the boundary of our "geometric" simplex, ∂S, is the union of $k + 1$ hyperplanar pieces lying on ∂P_i, and they themselves are "geometric" simplexes of dimension $k - 1$. (Why?) By the inductive hypothesis, the points

[11] Actually, those definitions are **dual** with each other, as discussed in section H9.5, so their equivalence means self-duality (autoduality) of the simplicial shape.

of ∂S are convex linear combinations of the vertices. In turn, a straight line passing through $x \in \text{int } S$ meets ∂S on both sides of x, so x is a convex linear combination of some points of ∂S, and so, using the inductive hypothesis and the transitivity of a convex linear combination, x is a convex linear combination of the vertices. **QED**. (We leave it to the reader to fill in the details; alternatively, readers may take a look at section H9.5.)

A convex set D on a straight line contains all points between $\sup D$ and $\inf D$, as is evident from the explicit description in section P9.2*.

S9.4

A transition step of the induction can be done as follows. For $n \geq 2$ we will obtain from Jensen's inequality in section P9.1* and the inductive hypothesis, denoting $x_i' = x_i$, $\alpha_i' = \alpha_i$ for $i = 1,\ldots, n-2$ and $x'_{n-1} = \frac{\alpha_{n-1}x_{n-1}+\alpha_n x_n}{\alpha_{n-1}+\alpha_n}$, $\alpha'_{n-1} = \alpha_{n-1} + \alpha_n$, that $x'_{n-1} \in D$ (why?), so

$$
\begin{aligned}
f\left(\sum_{i=1}^{n} \alpha_i x_i\right) &= f\left(\sum_{i=1}^{n-1} \alpha'_i x'_i\right) \leq \sum_{i=1}^{n-1} \alpha'_i f(x'_i) \\
&= \sum_{i=1}^{n-2} \alpha_i f(x_i) + (\alpha_{n-1} + \alpha_n) f\left(\frac{\alpha_{n-1}x_{n-1} + \alpha_n x_n}{\alpha_{n-1} + \alpha_n}\right) \\
&\leq \sum_{i=1}^{n-2} \alpha_i f(x_i) + (\alpha_{n-1} + \alpha_n)\left[\frac{\alpha_{n-1}}{\alpha_{n-1} + \alpha_n} f(x_{n-1}) + \frac{\alpha_n}{\alpha_{n-1} + \alpha_n} f(x_n)\right] \\
&= \sum_{i=1}^{n} \alpha_i f(x_i).
\end{aligned}
$$

QED.
Readers are encouraged to produce a geometric proof themselves using the geometric definition.

S9.5

$X^{k*} = X^{(k+2)*}$ starting from $k = 1$, as $Y^{**} \supseteq Y$, and so $Y^{***} = (Y^{**})^* \subseteq Y^*$, but on the other hand, $Y^{***} = (Y^*)^{**} \supseteq Y^*$.

For a vector subspace $L = \{x : \langle x, p_i\rangle = 0,\ i = 1,\ldots,n-k\}$, obviously, $L^\perp \subseteq L^*$ (where $L^\perp$ is spanned on p_i); on the other hand, if $p \notin L^\perp$, then we have $\langle p, x\rangle \neq 0$ for some $x \in L$ (for example, for the orthogonal projection of p onto L), so $\langle p, kx\rangle > 1$ for large positive or large negative k, and thus $L^\perp = L^*$. (Create a figure.)

For the points of a k-dimensional affine plane $L_1 = \{x : \langle x, p_i\rangle = 1,\ i = 1,\ldots, n-k\}$, we have $\langle \sum \alpha_i p_i,\ x\rangle = \sum \alpha_i$, which proves that $L_1^* \cap L^\perp = L_-^\perp :=$

$\{\sum \alpha_i p_i : \sum \alpha_i \le 1\}$ (with the foregoing $L^\perp$). Now, let $p \notin L^\perp$. We have $L_1 = p^0 + L$, with the foregoing L and the point p^0: $\{p^0\} = L_1 \cap L^\perp$. [Express the coefficients of a linear combination $p^0 = \sum \alpha_i^0 p_i$ via the Gram matrix $\left(\langle p_i, p_j\rangle\right)_{i,j}$.] The orthogonal decomposition $p = u + v$ ($u \in L$, $v \in L^\perp$) shows that $\langle p, p^0 + kv\rangle = \langle u, p^0\rangle + k|v|^2 \to \infty$ as $k \to \infty$, which gives $L_1^* \subseteq L^\perp$, and thus $L_1^* = L_-^\perp$. **QED**. (Create a figure.)

A. By the CSB inequality, $|\langle p, x\rangle| \le |p||x|$, $\langle p, x\rangle \le 1$ for $|p| \le \varepsilon$ and $|x| \le 1/\varepsilon$, which shows that for a bounded X, X^* contains a neighborhood of the origin.[12]

For a single point $a \ne 0$, the preceding description of the dual set shows that $\{a\}^*$ is a half-space bounded by an affine hyperplane passing through the endpoint of vector $a/|a|^2$ and is orthogonal to this vector. Because of the equality $(X \cup Y)^* = X^* \cap Y^*$, the set dual to the set of the vertices of a simplex in $\mathbb{R}^n$ is an intersection of $n + 1$ closed affine half-spaces. Hyperplanar borders of each of n half-spaces from these $n + 1$ are in a general position (have a point intersection) since vectors with endpoints at the corresponding vertices are linearly independent. Considering now the locations of the borders we conclude that a set that is dual to a set of the vertices of a simplex that contains the origin in its interior is a bounded set and, thus, a simplex (by Lemma 2 from section S9.2). Indeed, if this set were not bounded, there would exist a straight line parallel to n of the $n + 1$ hyperplanar borders, which is impossible since these hyperplanes are in a general position. QED. [Create a figure and work out the details. Readers preferring a different approach using computations may proceed as follows. Let a_i ($i = 0,\ldots, n$) be the vertices of the simplex (or vectors with endpoints at these vertices). Using Lemma 1 from section S9.2 we have $0 = \sum \beta_i a_i$, with positive coefficients β_i (which may be scaled by any common factor). Subtracting, if necessary, this properly scaled equation from the equation $x = \sum \alpha_i a_i$ with nonnegative coefficients α_i we will arrive at a similar equation for x having some coefficients equal to zero; thus, $\mathbb{R}^n$ is a union of $n + 1$ "orthants" $O_i = \left\{\sum_{j \ne i} \alpha_j a_j : \alpha_j \ge 0, \forall j\right\}$. For $p \in O_i$, the inequalities $\langle p, a_j\rangle \le 0$ cannot hold for all $j \ne i$ unless $p = 0$ (indeed, $|p|^2 = \sum \alpha_j \langle p, a_j\rangle \le 0 \Rightarrow p = 0$). Therefore – by virtue of the compactness of an $n - 1$-dimensional simplex O_i' spanned on the set $\{a_j: j \ne i\}$, the continuity of α_j as functions on it and the fact that the (pointwise) maximum of a finite number of continuous functions is also a continuous function {which follows from the identity $\max(x,y) = \frac{1}{2}(x + y + |x - y|)$} – $\exists c > 0$, $\forall p \in O_i'$, $\exists j \ne i$: $\langle p, a_j\rangle \ge c$; or, for $p \in k \cdot O_i'$ with $k > 0$, we have $\langle p, a_j\rangle \ge kc$, using some $j \ne i$. Hence, $p \in O_i$ & $\langle p, a_j\rangle \le 1$, $\forall j \ne i \Rightarrow p$ belongs to the union of $k \cdot O_i'$ with $0 \le k \le c^{-1}$}.

B. Start to prove that X^* can be properly defined using only all of the points of convexity of X as defined in section H9.5 from the following theorem.

Rough triangulation theorem. *The convex hull of a finite set of points X not contained in a hyperplane is a union of the simplexes with vertices in X.*

[12] A similar result (without exact estimates) may be obtained by using the continuity of a scalar product.

This can be proved using the following lemma.

Lemma *The boundary of the convex hull of a finite set of points X not contained in a hyperplane belongs to a union of convex hulls of the subsets of X, which are contained in hyperplanes.*

Proof of lemma. For the points of *conv* (*X*) that are not contained in the convex hull of any subset of *X* belonging to a hyperplane, a presentation as convex linear combinations $\sum \alpha_i x_i$ (where all x_i are elements of *X*) is such that $\alpha_i > 0$, at least for $n + 1$ indices *i*. (Why?) Thus, any such point can be presented, after a proper index rearrangement, as $x_0 + \sum_{j=1}^{n} \beta_j (x_j - x_0) + a$, with $\beta_j > 0$, $\forall j = 1,\ldots, n$, and $\sum_{j=1}^{n} \beta_j < 1$ [where $a = \sum_j \beta_j (x_j - x_0)$]. This is an internal point of *conv* (*X*), which may be proved similarly to Lemma 1 from section S9.2, considering the map

$$\begin{cases} \mathbf{R}^n & \to & \mathbf{R}^n \\ (\beta_1, \ldots, \beta_n) & \mapsto & x_0 + \sum_{j=1}^{n} \beta_j v_j + a \end{cases}$$ (where $v_j = x_j - x_0$). **QED**. (We leave it to the reader to fill in the details.)

Proof of theorem. Apply induction on the spatial dimension. A straight line passing through a point $x \in conv\ (X)$ and $x_0 \in X$ meets the boundary at some point x' lying on a different side of the line from *x* than x_0. (Why?) By the inductive hypothesis and the preceding lemma, x' belongs to a simplex of dimension $n - 1$ with vertices in *X*, so *x* belongs to a simplex of dimension *n* with vertices in *X*. QED. (We leave it to the reader to fill in the details.)

Now, for a good (as defined in section H9.5) set *X*, let $x \in X$ be a point of convexity. By the rough triangulation theorem, $x \notin C$, where $C := conv(X \setminus \{x\})$. By virtue of the compactness of *C*, there exists a point $c \in C$, closest to *x*. Evidently, this point belongs to the border of *C* (why?), and an affine hyperplane *P* passing through the point *c* orthogonally to the vector $x - c$ is a **hyperplane of support** for the set *C* (which means that *C* lies entirely in a closed half-space having *P* as its border). Indeed, if a point $y \in C$ does not belong to that half-space, one could find a point $c' \in C \cap L$, closer to *x* than *c*, where *L* denotes the straight line passing through *c* and *y*. (Create a figure and work out the details.[13]) Obviously, $0 \notin P$ (as by

[13] The famous Hahn-Banach theorem states, in one of its geometrical versions, that a closed convex set has hyperplanes of support at all its face points. It can be proved with no regard for the compactness, which allows multiple applications in infinite-dimensional analysis, where *boundedness + closure ≠ compactness*. (In infinite-dimensional problems, mostly closed hyperplanes of support are considered.) For this reason, the Hahn-Banach theorem is usually learned in courses on functional analysis. Readers seeking more information may consult Dunford and Schwartz (1957), Hille and Phillips (1957), Yosida (1965), (Riesz and Nagy 1972), Edwards (1965), Reed and Simon (1972), Rudin (1973), Kolmogorov and Fomin (1976), Rockafellar (1970), and Kutateladze and Rubinov (1976).

assumption $0 \in \text{int } C$; work out the details), so $P = \{y : \langle p, y\rangle = 1\}$, with the definite p collinear with $x - c$ (how is this p defined?), and, evidently, $p \in (X \setminus \{x\})^* \setminus X^*$(why?). **QED**.

An important remark The vector p', proportional to $x - c$ and normed so that $\langle p', x\rangle = 1$, belongs to the interior of $(X\setminus\{x\})^*$. (Why?) Therefore, this interior contains p' together with its neighborhood, and so a boundary of X^* contains a neighborhood of p' in the affine hyperplane $Q = \{q : \langle x, q\rangle = 1\}$ (complete the details), which is used in the proof of the corollary in section E9.5(**C**), and in the proof in section E9.5(**D**).

S9.6

A. For a partition $I = I_1 \cup \ldots \cup I_k$ ($I_l \cap I_m = \varnothing$ for $l \neq m$), we will have, considering x_i such that $x_i = x_j$ when i and j belong to the same I_l and denoting them y_l:

$$f\left(\sum_l \left(\sum_{i\in I_l} \alpha_i\right) y_l\right) = f\left(\sum_l \sum_{i\in I_l} \alpha_i x_i\right)$$

$$= f\left(\sum \alpha_i x_i\right) \leq \sum \alpha_i f(x_i) = \sum_l \left(\sum_{i\in I_l} \alpha_i\right) f(y_l),$$

QED.

B. We have, applying the Jensen inequality twice, with respect to $\{\alpha_i\}$ and $\{\beta_j\}$:

$$f\left(\sum_{i,j} \alpha_i \beta_j x_{ij}\right) = f\left(\sum_i \alpha_i \sum_j \beta_j x_{ij}\right) \leq \sum_i \alpha_i f\left(\sum_j \beta_j x_{ij}\right) \leq \sum_{i,j} \alpha_i \beta_j f(x_{ij}),$$

QED. (We leave it to the reader to fill in the details: do the points $\sum_j \beta_j x_{ij}$ belong to the domain D?)

C. Readers have to verify that all points put in f in the computation from section E9.6 (**C**) belong to D.

D. This claim follows directly from **B** by induction (We leave it to the reader to complete the details).

Vector space $\mathbb{R}$ over the field $\mathbb{Q}$ is infinite-dimensional. (Why? What are cardinalities of $\mathbb{R}$ and $\mathbb{Q}$, respectively?) A basis of a infinite-dimensional vector space $\mathbb{R}$ over $\mathbb{Q}$ "is fixed" using the axiom of choice (but cannot be found without it).

S9.8

A point $(x_-, f(x_-))$ is placed on or below the secant l_1 passing through the points $(x_1, f(x_1))$ and $(x, f(x))$. In turn, a point $(x_+, f(x_+))$ is placed on or above l_1. (Why?) Therefore, $k \geq k_1$, where k_1 is a slope of l_1. (Why? Make a figure.) Similarly, $k \leq k_2$, where k_2 is a slope of the secant passing through the points $(x_2, f(x_2))$ and $(x, f(x))$, so k is bounded: $k_2 \geq k \geq k_1$. Considering the secant $l_0 = \{y = k_0 x + b_0\}$ passing through $(x_1, f(x_1))$ and $(x_2, f(x_2))$ brings inequalities $f(x) \leq kx + b \leq k_0 x + b_0$ showing boundedness of b. (Make a figure and complete the details.)

Next, since the points $(x', f(x'))$ for $x' \notin (x_-, x_+)$ are placed on or above the secant passing through $(x_\pm, f(x_\pm))$ (why?), such points of $x' \neq x$ will be placed on or above any limit straight line. A similar argument shows that the point $(x, f(x))$ cannot be placed above the limit line; but it cannot be placed below it either (why?), which finalizes the proof. (We leave it to the reader to work out the details.)

The simplest example of a nonconvex function satisfying the local support condition comes from $f(x) = \begin{cases} 1, & \text{for } x \neq 0, \\ 0, & \text{for } x = 0. \end{cases}$ For a continuous function, the local support condition is equivalent to the global support condition and, thus, to convexity.

Indeed, let $x_1 = \sup \{x' > x$: $(x'', f(x''))$ *are placed on or above l for* $x \leq x'' \leq x'\}$, for a fixed straight line l of local support at $(x, f(x))$, belong to the domain of f. Because of the continuity of f, the point $(x_1, f(x_1))$ cannot be placed below l, so $(x_1, f(x_1)) \in l$ or x_1 is an endpoint. Let it not be an endpoint. Subtracting from f the linear function that has l as its graph (which does not affect the claim) makes it possible to identify l with the abscissas' axis. Consider the maximum of f on the segment $[x, x_1]$. Let x_2 be the last point of this segment where $f = f_{\max}$. Verify the absence of a straight line of local support at $(x_2, f(x_2))$. A similar consideration is performed for the segment to the left of x, which will finalize the proof. (We leave it to the reader to work out the details.)

S9.9

If $f''_{x_0}(h, h) < 0$ for some h, then we will have, for small $t \neq 0$,

$$f(x_0 + th) - f(x_0) - f'_{x_0}(th) = t^2 \left(f''_{x_0}(h, h) + o(1)\right) < -\varepsilon t^2 < 0$$

(using some $\varepsilon > 0$). **QED**. Conversely, for $f''_v \geq 0$, the internal integral in $\int_{x_0}^{x} du \int_{x_0}^{u} f''_v \, dv$ takes values of the same sign as $u - x_0$, and thus as $x - x_0$; in turn, the

differential in the external integral takes values of the same sign, as $x - x_0$; therefore, the iterated integral takes nonnegative values. **QED**.

S9.10

The iterated integral considered in section S9.9 takes positive values on $x \neq x_0$ when the second differential is a positive definite quadratic form, $f''_v > 0$, at all internal points. This implies that f is strictly convex, as a point $(x, f(x))$ obviously is placed below the straight line passing through $(x', f(x'))$ and $(x'', f(x''))$ for $x' < x < x''$. (Create a figure.) (However, the strict convexity of a twice-differentiable function does not imply the positive definiteness of its second differential at all internal points, as is clear considering the function $y = x^4$.)

The inequalities between "discrete" power means are obtained from the corresponding integral means by considering piecewise constant functions that take values x_i along the whole intervals of lengths α_i.

Proving these inequalities with a different technique, one may restrict oneself to the cases of equal weights since *allowing equal weights + continuity = allowing any weights* (in accordance with section P9.6* above; fill in the details).

S9.14

To find $\lim\limits_{n\to\infty} M_{\infty,n}$, note that, obviously, it does not exceed max (X), but on the other hand, for large n, the minimum of X on a segment $\left[\frac{i-1}{2^n}, \frac{i}{2^n}\right]$ containing a maximum point is arbitrarily close to max (X) due to the continuity of X, **QED**.

Next, $\lim\limits_{p\to\infty} M_{p,\infty} = \lim\limits_{n\to\infty} M_{\infty,n} = \sup\limits_{p,n<\infty} M_{p,n}$ due to the estimates

$$M_{p,n} \le M_{\infty,n}, M_{p,\infty} \le \sup_{p,n<\infty} M_{p,n}, \qquad \sup_{p,n<\infty} M_{p,n} - M_{p,n} \to 0 \quad \text{as } p, n \to \infty.$$

QED.

Defining the sequence $M_{p,n}$ in a simpler way, such as $M_{p,n} := \left(n^{-1}\sum\limits_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$, generally we would lose the monotonicity on n. (Readers may provide examples.) In this case, the proof might be completed using the following fine theorem.

Young-Dini theorem. *Let a sequence $a_{m,n}$ converge with respect to* both *its indices separately:* $\forall n \le \infty$, $\exists \lim\limits_{m\to\infty} a_{m,n}$, $\forall m \le \infty$, $\exists \lim\limits_{n\to\infty} a_{m,n}$. *If one of these convergences is monotone, say, $a_{m,n} \le a_{m+1,n}$, $\forall n \le \infty$, then there exists a double limit* $\lim\limits_{m,n\to\infty} a_{m,n}$.

(Proving this claim is a very nice exercise in analysis, so readers are encouraged to do it.)

Also, readers may establish the equality $\lim_{p\to\infty} M_p(X) = \max(X)$ by the following arguments. Denote $x_i = \min_{\frac{i-1}{n} \le t \le \frac{i}{n}} X(t)$. Choose, for $\varepsilon > 0$, so large n that $x_{i_0} > \max(X) - \varepsilon$, where $x_{i_0} = \max_{i=1,\ldots,n} x_i$. As with any $p \ge 0$, we have

$$M_p(X) \ge \left(n^{-1}\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}} = x_{i_0}\left(n^{-1}\sum_{i=1}^{n}(x_i/x_{i_0})^p\right)^{\frac{1}{p}} \ge x_{i_0} n^{-\frac{1}{p}},$$

which cannot be less than $x_{i_0} - \varepsilon$ for large p (keeping n fixed). Finally, $\lim M_p(X) > \max(X) - 2\varepsilon$, completing the proof.

S9.17

The oriented volume of a pyramid with a fixed basic face is an affine (linear plus constant) function of the opposite vertex,because the oriented height of this vertex over the base is such a function.

An affine function will be equal to zero identically if it equals zero on an open set because that set has sufficiently many linearly independent elements to form a spatial basis. (We leave it to the reader to work out the details.)[14]

S9.20

The uniqueness of $x(p)$ for strictly convex functions f may be proved as follows. A function $F(x,p) = px - f(x)$ with respect to x, keeping p fixed, is strictly concave [by section P9.1* because, in fact, adding a linear function does not affect strict convexity (concavity)], so if $x_1(p) \ne x_2(p)$ were its points of maximum, then its graph between these maxima would be placed above the horizontal straight line connecting them, so actually they would not be maxima. (Create a figure.)

[14] Advanced readers probably know that this statement is generalized for any analytic function (a one, representable by the sum of a convergent power series in a neighborhood of any point of its domain): an analytic function on an open connected set will be equal to zero on the entire domain if it equals zero on its open subset. Interested readers will find more details in Dieudonné (1960).

The simplest method of establishing the identity $p\varphi^{-1}(p) - \int_0^{\varphi^{-1}(p)} \varphi(x)dx = \int_0^p \varphi^{-1}(y)dy$ [for monotone continuous functions φ with $\varphi(0) = 0$] is geometric, by adding areas under the graphs of $\varphi(x)$ and $\varphi^{-1}(y)$. (Create a figure.)

A different method of establishing this identity, for differentiable φ, consists in integrating the right-hand side by parts and then making a change to the variables $y \mapsto x := \varphi^{-1}(y)$ in the integral. (Fill in the details.)

A similar approach is applicable to any φ using the following version of the integration-by-parts formula:

Theorem *Let f and g be integrable functions on* [*a*,*b*]. *Then fg is also integrable, and, denoting* $G(x) = \int_a^x g(t)\,dt$,

$$\begin{aligned} \int_a^b f(x)g(x)\,dx &= G(b)f(b) - G(a)f(a) - \lim \sum_{i=1}^{n} G(x_i)(f(x_i) - f(x_{i-1})) \\ &= G(b)f(b) - G(a)f(a) - \lim \sum_{i=0}^{n-1} G(x_i)(f(x_{i+1}) - f(x_i)), \end{aligned}$$

with the limits taken over all finite partitions $a = x_0 < \ldots < x_n = b$, *as max* $(\Delta x_i) \to 0$.

Readers are invited to prove this theorem and then prove the foregoing stated Young's identity.

Another corollary of this theorem is the classic second integral intermediate value theorem, originally established by O. Bonnet in 1849, which has multiple applications, as follows.

Bonnet's theorem. *Let f be a monotone function, and g an integrable one, on* [*a*,*b*]. *There exists* $\xi \in (a,b)$ *such that*

$$\int_a^b f(x)g(x)dx = f(a)\int_a^\xi g(x)\,dx + f(b)\int_\xi^b g(x)\,dx.$$

Derive Bonnet's formula from the previous theorem. Also, readers may find such a proof in classic analysis handbooks (e.g., Fichtengoltz 1970).

S9.22

As described in section H9.22, the asymptotic behavior near the endpoints of the closure of D (when $D \neq \mathbb{R}$) of a convex function f, defined on a convex domain $D \subseteq \mathbb{R}$, may be established as follows. First, note the boundedness of f from below in a neighborhood of the endpoint. Indeed, otherwise three points $x_1 < x_0 < x_2$

could be found, with $(x_0, f(x_0))$ lying above the straight line passing through $(x_1, f(x_1))$ and $(x_2, f(x_2))$. [Create a figure. Alternatively, a lower bound can be determined using a straight line of support (section P9.8*).] Next, note that only one limit value exists. (Use similar arguments; create a figure). Finally, the extension of f by continuity to an endpoint x is a convex function, regardless of whether f has a finite or infinite limit at x, because

$$f(\alpha x + (1-\alpha)y) = \lim_{x_n \in D,\ x_n \to x} f(\alpha x_n + (1-\alpha)y) \le \lim_{x_n \in D,\ x_n \to x} [\alpha f(x_n) + (1-\alpha)f(y)]$$
$$= \alpha \left[\lim_{x_n \in D,\ x_n \to x} f(x_n) \right] + (1-\alpha)f(y) = \alpha f(x) + (1-\alpha)f(y).$$

QED.

S9.23

For a convex function f defined on a convex domain $D \subseteq \mathbb{R}$, the same arguments as in section H9.23 show the existence of the corresponding one-sided derivatives at the endpoints of D (if it has endpoints at all). Therefore, at the endpoints of the closure of D (when $D \neq \mathbb{R}$) there exist one-sided derivatives of the extension of f by continuity to this closure. (Indeed, this extension is convex, as discussed in sections H9.22 and S9.22.)

There may be at most a countably infinite set of points of nondifferentiability of f (due to unequal one-sided derivatives) because there may be found at most a countably infinite set of nonintersecting intervals on a straight line. (Indeed, any interval contains rational points.)

Finally, discontinuous, but lower (upper) continuous, functions are the famous **Heaviside step functions**: $\theta(x) = \begin{cases} 0 & \text{when } x \le 0, \\ 1 & \text{when } x > 0 \end{cases}$ $\left(\text{resp. } \theta(x) = \begin{cases} 0 & \text{when } x < 0, \\ 1 & \text{when } x \ge 0 \end{cases} \right)$.

Convexity of the primitive of a monotonically nondecreasing function on a real interval. The integrability of a nondecreasing function f is implied by a **theorem by Lebesgue** and the countablity of the number of discontinuities. (Work out the details.) For a differentiable f, the convexity of the primitive follows from section P9.9*. (How?) In the general case, it can be proved as follows. Without loss of generality, we are dealing with a function on the segment $[y,x] = [0,1]$, so we must prove that $\int_0^\alpha f(x)\,dx \le \alpha \int_0^1 f(x)\,dx$ $(0 \le \alpha \le 1)$, or, equivalently, that $(1-\alpha)\int_0^\alpha f(x)\,dx \le \alpha \int_\alpha^1 f(x)\,dx$. (Why? Answer by creating a figure.) But we have

$$(1-\alpha)\int_0^\alpha f(x)\,dx \ \le\ (1-\alpha)\alpha f(\alpha) = \alpha(1-\alpha)f(\alpha) \ \le\ \alpha \int_\alpha^1 f(x)\,dx,$$

by virtue of the monotonicity of f. **QED**. (We leave it to the reader to fill in the details.)

S9.26

Completing the proof of Gelfand's lemma. If Gelfand's lemma is false, then by the lemma from section H9.26, there is a point x_1 such that $f(x_1) > 1$, and, due to the lower semicontinuity at x_1, we find a ball B_1 of radius ≤ 1 with its center at x_1 on which $f > 1$. By induction we can use the lemma from section H9.26 for a point $x_{n+1} \in B_n$ such that $f(x_{n+1}) > n + 1$, and, due to the lower semicontinuity at x_{n+1}, we find a ball B_{n+1} of radius $\leq (n + 1)^{-1}$ with its center at x_{n+1} on which $f > n + 1$. Hence, f could not get a finite value on an intersection $\bigcap B_n$.

A bounded convex function may not be upper semicontinuous: for example, such is the case in $f(x,y) = \begin{cases} y^2/x & \text{if } (x,y) \neq (0,0), \\ 0 & \text{otherwise} \end{cases}$ on a convex domain $D = \{x \geq 0,\ |y| \leq \sqrt{x}\}$ ($0 \leq f \leq 1$: draw a figure). This function is homogeneous of degree one, and so it is linear on any ray with the vertex (0,0). So the second differential f'' is positive semidefinite everywhere including the boundary points, except (0,0). Hence, f is convex (use section P9.9*), but it is not upper semicontinuous at (0,0). (We leave it to the reader to work out all the details.)

S9.28

The proof of the lemma from section P9.21**, discussed in section H9.21, will be applicable to the lemma in section E9.28, literally, if rational γ_n is used.

S9.29

We have $f(0) = f(2 \cdot 0) = 2 \cdot f(0)$, whence $f(0) = 0$. Therefore,

$$0 = f(0) = f(x - x) \qquad f(x) + f(x) = 2f(x),$$

so $f(x) \geq 0$, $\forall x$. Hence, $0 \leq f(x + y) \leq f(x) + f(y)$, so $f(x) = f(y) = 0 \Rightarrow f(x + y) = 0$, so the zero set is a subspace. (Work out the details.)

A proof of the continuity of a convex functional in a finite-dimensional space, different from that given in section H9.29, can be completed by considering the "balls" $B_{f,k} = \{f \leq .k\}$ for the factor-functional on the quotient space modulo zero subspace (for which we will use the same notation "f"). These are homothetic convex sets, symmetric relative to the origin, containing a segment of positive

length in any spatial direction. (Why?) Prove the boundedness of the balls (with respect to a common Euclidean norm). To do this, use the representation of vectors in spherical (using a fixed Euclidean norm) coordinates $v = r{\cdot}\sigma$ ($r = |v|$, $\sigma \in \mathbf{S}^{n-1}$, and n is the spatial dimension). If $v_n \in B_{f,1}$ and $r_n \to \infty$, then we obtain $f(\sigma_n) \to 0$, which contradicts the convexity of $B_{f,1}$; readers will easily find this contradiction if they consider a simplex, contained in $B_{f,1}$,, which, in turn, contains the origin in its interior. As follows directly from the boundedness of $B_{f,k}$, $f(y)$ is small for small y, which shows the smallness of the difference between $f(x + y)$ and $f(x)$, provided that y is small:

$$f(x+y) - f(x) \text{ is small } f(y), \qquad f(x) - f(x+y) = f(x+y-y) - f(x+y) \text{ is small } f(-y) = f(y).$$

QED. (We leave it to the reader to fill in all the details of this proof.)

One-Parameter Groups of Linear Transformations

Problems

P10.0[*]

Preliminaries. A one-parameter linear group in a vector space L is a continuous map $g: t \mapsto g^t = g(t)$ on the space of parameters $\mathbb{R}$, taking values in a space of the linear operators on L and having the following properties: (1) $g^0 = E$ and (2) the sum of parameter values corresponds to the composition $g^{s+t} = g^s \circ g^t$.

`Verify` that property (1) may be substituted by $(1)'$ – all g^t are invertible operators. Thus, a one-parameter linear group is a group homomorphism g: $\langle \mathbb{R}, + \rangle \to \langle GL(L), \circ \rangle$, which is continuous as well: $t \to t_0 \Rightarrow g^t \to g^{t_0}$. Readers can define this continuity themselves using the usual terminology with which they are familiar.[1]

Let us consider the details of the simplest case, where $L = \mathbb{R}$. As readers know, linear operators on a one-dimensional space are multiplications by scalars. (`When` are these operators invertible?) In particular, we have a natural isomorphism between $GL(1,\mathbb{R})$ and a multiplicative group $\langle \mathbb{R}^*, \rangle$. Thus, one-parameter linear groups in $\mathbb{R}$ are the continuous homomorphisms g: $\langle \mathbb{R}, + \rangle \to \langle \mathbb{R}^*, \rangle$, or, in other words, the continuous solutions of the functional equations

[1] The fact that elements $g^t - g^{t_0}$ in the space of linear operators $GL(L)$ tend to 0 for $t \to t_0$ may be expressed in terms of by-coordinate tending to zero, that is, for all of $(\dim L)^2$ coordinates with respect to any fixed basis in $GL(L)$. However, readers may show that for a finite-dimensional L the continuity of a group is **equivalent to continuities** of the orbits: $t \to t_0 \Rightarrow g^t \to g^{t_0}$ if (and, evidently, only if) $t \to t_0 \Rightarrow g^t x \to g^{t_0} x, \ \forall x \in L$. (Proving a similar claim for infinite-dimensional spaces requires stipulating a kind of topological **completeness** of the space (the **Banach-Steinhaus equicontinuity theorem**). Interested readers may apply to Dunford and Schwartz (1957), Hille and Phillips (1957), Yosida (1965), Halmos (1967), Riesz and Sz.-Nagy (1972), Reed and Simon (1972), Rudin (1973), and references therein.)

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_10,

$$g(0) = 1, \qquad g(s+t) = g(s)g(t). \tag{*}$$

As is well known, exponential functions satisfy these equations. Having solved problems P10.1*, P10.2*, P10.3*, P10.4*, P10.5*, P10.6*, P10.7*, P10.8*, P10.9*, P10.10*, and P10.11*, readers will find that the continuous solutions of (*) are all exponential and so will extract the usual properties of exponential functions from functional equations (*) (which corresponds to a real historical development). Readers will proceed along with the exponential power series. In this problem group, prior knowledge of power series is not assumed; therefore, in the "Hint" section of the book we provide the necessary definitions and ask readers to prove the related analytical claims, for example, the famous **Abel's lemma**.

The second part of this problem group (sections P10.13, P10.14**, P10.15**, P10.16**, P10.17**, P10.18**, P10.19**, P10.29**, P10.30**) is devoted to proper generalizations of the statements in sections P10.1*, P10.2*, P10.3*, P10.4*, P10.5*, P10.6*, P10.7*, P10.8*, P10.9*, P10.10*, and P10.11* for one-parameter matrix groups. Readers will find that all of them are given by the matrix exponential series, as in the one-dimensional case. However, a multidimensional picture looks much more complicated. Working on these problems will familiarize readers with matrix analysis. For this, in the "Hint" and "Explanation" sections of the book, we provide readers with elements of multivariate differential calculus (including the implicit function theorem); however, readers are assumed to have a working knowledge of eigenvalues and the Jordan canonical form of a matrix.

In section P10.13 we analyze the structure of one-parameter matrix groups. For the reader's convenience, this section is split into six connected problems, A–F. While working on these problems readers may need familiarity with some features of number theory and theory of differential equations; therefore, in the "Hint", "Explanation," and "Completing the Solution" sections we discuss the appropriate tools: **Euclid's algorithm**, **Chinese remainder theorem,** and **Poincaré's recurrence theorem**.

In the next part (sections P10.20**, P10.21***, P10.22**, P10.23**, P10.24**, P10.25**, P10.26*, P10.31**) readers will find some elementary classic applications of one-parameter matrix groups in differential equation theory (Liouville formula), in complex analysis (e.g., complex exponential functions, Euler's formula), for finite-dimensional functional spaces (spaces of quasipolynomials), and others. Readers will learn how to deal with these problems using elementary tools of analysis and linear algebra.

There is a part of this problem group that formally contains two problems (sections P10.27**, P10.28**), but in fact these include a large set of connected problems. We have compiled these problems for readers possessing more advanced prior experience (although we introduce and discuss the necessary tools). While working on them, readers will come across many interesting aspects of analysis and linear algebra and become acquainted with important concepts (e.g., symplectic forms). In the "Hint" section (H10.27) readers will find an excellent classic real-life application to theoretical physics (phenomenological thermodynamics).

Readers will encounter far-reaching advances and applications of the subjects considered in the present problem group in powerful mathematical theories: **differential equations**, **Lie groups,** and **group representations**.[2]

P10.1[*]

`Prove` that a one-parameter linear group in $\mathbb{R}$ takes on positive values. [The results of sections P10.6[*] and P10.8[*] below show that the image of such a group is $\{1\}$ or $\mathbb{R}^{+*} = (0,\infty)$.]

P10.2[*]

`Prove` that for any $a > 0$ and $t_0 \neq 0$ there is at most one one-parameter linear group g in $\mathbb{R}$ such that $g(t_0) = a$.

P10.3[*]

`Show` that a differentiable solution of the equations (*) from section P10.0[*] satisfies a differential equation $\dot{g}(t) = kg(t)$, $k = const = \dot{g}(0)$ ($\dot{g} = dg/dt$).

---

[2] An ordinary autonomous differential equation, such that the solutions do not become infinite for a finite time, defines a one-parameter group of **diffeomorphisms** of the phase space, g^t, called the **phase flow**: $t \mapsto g^t x$ ($t \in \mathbb{R}$) are the solutions of this equation under the initial conditions $g^0 x = x$. (The transformations g^t are linear when the equation is linear.) A Lie group is a group and simultaneously a smooth manifold such that the group operations are smooth maps; in fact, it is a multiparameter, or multidimensional, group $g^{t,u,\cdots}$, but of course an identity such as $g^{t_1+t_2,u_1+u_2,\cdots} = g^{t_1,u_1,\cdots} g^{t_2,u_2,\cdots}$ may hold only for Abelian (commutative) Lie groups. Note that discrete, e.g., finite **Hausdorff**, **groups** are Lie groups. (`What` are their dimensions?) In the theory of group representations, abstract groups and functions on them are studied via homomorphisms (representations) of those groups onto groups of linear transformations of vector spaces. In this way various group-theoretic problems are reduced to problems in linear algebra; this is also very important for applications, e.g., when dealing with symmetry groups of physical systems. A detailed discussion of the contents of all these mathematical theories is beyond the scope of this problem book; and interested readers may consult Arnol'd (1975, 1978, 1989), Weyl (1939, 1952), Chevalley (1946–1955), Lang (1965), Vilenkin (1965), Nomizu (1956), Serre (1965), Adams (1969), Weil (1940), Bredon (1972), Kirillov (1972), Humphreys (1973), Naymark (1976), Helgason (1962), Kobayashi and Nomizu (1963, 1969), Bourbaki (1968–1982), Warner (1983), Fulton and Harris (1991), Alperin and Bell (1995), and references therein; in addition, readers may turn to the vast literature devoted to applications of group theory to differential geometry, functional analysis, and theoretical physics (theory of relativity, quantum mechanics, theory of solids, and other branches).

P10.4[*]

A formal solution to the algebraic equation $F(t,g) := \sum F_{ij} t^i g^j = 0$ is a formal power series $g(t) = \sum_{n=0}^{\infty} a_n t^n$ such that inserting it into the equation yields the resulting formal series in t on the left-hand side to zero (and, similarly, a formal solution of a differential equation is defined using formal term-by-term differentiation $\dot{g}(t) = \sum_{n=1}^{\infty} n a_n t^{n-1}$). An **initial value problem** (**Cauchy problem**) is formulated as finding $g(t)$ satisfying a given differential equation and having a given value $g(0)$. `Prove` that the Cauchy problem for the differential equation from section P10.3[*] using initial value $g(0) = 1$ has a unique formal solution, namely, $g(t) = \sum (n!)^{-1} (kt)^n$. (This series is referred to as **exponential**.)

P10.5[*]

A numerical series $\sum_{n=0}^{\infty} a_n$ is said to **converge** when $\lim_{n\to\infty} \sum_{k=0}^{n} a_k$ exists, and this limit is called the series' sum. Suppose we have proved that (1) the exponential power series $g(t)$ from section P10.4[*] converges, (2) a series of its term-by-term derivatives $\dot{g}(t) = k \sum [(n-1)!]^{-1} (kt)^{n-1}$ also converges, and (3) the sum of $g(t)$ is differentiable, having a derivative equal to the sum of $\dot{g}(t)$. Then the sum of $g(t)$ itself will be a solution of the Cauchy problem in section P10.4[*]. (`Why`?) Therefore, `prove` claims (1)–(3). (For additional definitions or tools related to this problem, see the corresponding hint in the "Hint" section.)

P10.6[*]

`Prove` that the sums of the exponential series satisfy the equations (*) from section P10.0[*]. From the foregoing statements `derive` that those sums are the only continuous solutions of (*).

P10.7[*]

`Prove` that $\lim_{n\to\infty} (1 + t/n)^n$ exists for any $t \in \mathbb{R}$ and is equal to $\sum (n!)^{-1} t^n$. [Readers know that this function of t is referred to as exponential and denoted e^t or $\exp(t)$. Multidimensional versions of sections P10.1[*], P10.2[*], P10.3[*], P10.4[*], P10.5[*],

P10.6*, and P10.7* are discussed in sections P10.13, P10.14**, P10.15**, P10.16**, P10.17**, P10.19**, P10.20**.]

*P10.8**

`Show` that (1) $e^{-t} = 1/e^t$, (2) $\lim_{t\to\infty} e^t = \infty$, (3) $\lim_{t\to-\infty} e^t = 0$, (4) e^t is a monotonically growing function. Therefore, for any $a > 0$ an equation $e^t = a$ has a unique solution. (`Prove.`) [Readers know that this function of a is referred to as (natural) logarithmic and denoted $\log a$ or $\ln a$.]

*P10.9**

For $a > 0$, define $a^t := e^{kt}$, where $k = \ln a$. [`Verify` that this definition extends a definition of integral powers as $a^n = \underbrace{a \cdot \ldots \cdot a}_{n}$, $a^{-n} = 1/a^n$ $(n > 0)$, $a^0 = 1$.] `Show` that $a^{st} = (a^s)^t$. Therefore, $\lim_{n\to\infty} (1 + kt/n)^n = \sum (n!)^{-1}(kt)^n = a^t$, where $a = e^k$.

`Establish` the claims from section P10.8* for a^t with $a \neq 1$. (The logarithms to base a are denoted $\log_a$.)

*P10.10**

`Prove` that the constant zero function $f \equiv 0$ and the logarithmic functions $f(t) = \log_a t$ $(a>0,\ a \neq 1)$ are the only continuous solutions of the functional equations

$$f(1) = 0, \qquad f(st) = f(s) + f(t), \tag{**}$$

on $\mathbb{R}^{+*} = (0, \infty)$. `Show` that the function f can be uniquely extended to the domain $\mathbb{R}^* = \mathbb{R}\backslash\{0\}$, retaining the property of being a solution of (**) as an even function $f(|t|)$. `Establish` the smoothness of the logarithms and derive for them a differential equation $\dot{f}(t) = 1/kt$.

*P10.11**

`Verify` that all solutions of the equations (*) in section P10.0* have a form $g(t) = e^{\varphi(t)}$, for appropriate homomorphisms φ: $\langle\mathbb{R},+\rangle \to \langle\mathbb{R},+\rangle$; the continuous

solutions correspond to linear functions $\varphi : \begin{cases} \mathbf{R} \to \mathbf{R} \\ t \mapsto kt \end{cases}$. Give examples of discontinuous solutions, that is, discontinuous homomorphisms g: $\langle \mathbb{R}, + \rangle \to \langle \mathbb{R}^*, \rangle$. (Therefore, the remaining properties of a one-parameter linear group do not imply the continuity.)

P10.12[*]

A brief historical excursion. John Napier (1550–1617) introduced tables of logarithms in the second half of the sixteenth century for accelerating computations using summation instead of multiplication. Therefore, the functional equations (**) were originally laid as a basis of notions of logarithms. (The same tables might be used in the opposite direction, which would correspond to an exponential function, but it would not be practical!)

Isaac Newton (1642–1727) introduced the exponential series and the differential equations presented in sections P10.3[*] and P10.10[*]. His aim was to convert a function defined by Merkator's series $\ln \frac{1+t}{1-t} = 2t\left(1 + \frac{t^2}{3} + \frac{t^4}{5} + \ldots + \frac{t^{2n}}{2n+1} + \ldots\right)$. Newton was the first mathematician to systematically use such a remarkable technique as power-series expansions for solving mathematical and physical problems (later on, those series were christened "Taylor series") (Arnol'd 1989 Historical). In addition, Newton was the first to use differential equations to describe various physical processes. Combining both approaches, he was able to obtain a differential equation for logarithmic functions, convert it into a differential equation for exponential functions, and find formal solutions for the last one using power series.

"Napier's number" $e = \sum (n!)^{-1}$ was in fact introduced by Leonhard Euler (1707–1783), who also introduced a famous constant $c = \lim\limits_{n \to \infty} \left(\sum\limits_{k=1}^{n} k^{-1} - \ln(n+1) \right) = 0.57721566490\ldots$.[3]

Finally, a curious anecdote closely touching on our subject is ascribed to Evariste Galois (1811–1832); interested readers may consult his biography by Infeld (1948).

P10.13. A[*]

Prove that the elements of one-parameter linear groups in finite-dimensional vector spaces over the reals preserve orientation, that is, have positive determinants.

[3] The arithmetical nature of this constant has not been investigated. Specifically, it is not known if it is a rational number or not (1984) (Vinogradov 1977–1984).

P10.13. B[**]

`Prove` that a one-parameter group that is not a constant map $g(t) \equiv E$ belongs to one of the two following types:

Type 1: $g(\mathbb{R})$ is isomorphic to the group of rotations of a circle that has the usual representative forms: either an additive, a group $\mathbb{R}$ mod 2π, or a multiplicative, a group of 2×2-matrices $R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ ($\theta \in \mathbb{R}$ mod 2π);

Type 2: $g(\mathbb{R})$ is isomorphic to $\langle \mathbb{R}, + \rangle$.

P10.13. C[***]

`Show` that the total eigenspace corresponding to any eigenvalue of any element $g(t_0)$ of a one-parameter linear group in $\mathbb{C}^n$ contains a common eigenspace for all elements $g(t)$. Let $\lambda_j(t) = \rho_j(t)\big(\cos\theta_j(t) + i\sin\theta_j(t)\big)$ be their eigenvalues on it; `prove` that the moduli of the eigenvalues depend on the parameter t exponentially and the arguments depend linearly: $\rho_j(t) = e^{k_j t}$, $\theta_j(t) = (\omega_j t) \bmod 2\pi$.

Is the type of a one-parameter linear group g completely determined, given $g(t_0)$ for some $t_0 \neq 0$? This question is discussed in sections P10.13. D[***], P10.13. E[***], and P10.13. F[***] below.

P10.13. D[***]

Let us agree to say that real numbers a_1, a_2 that are not integer multiples of $b \neq 0$ (usually, $b = 2\pi$) are **commensurable modulo** b if $m_1 a_1 - m_2 a_2$ is an integer multiple of b for some nonzero *relatively prime* integers m_1, m_2. `Show` the equivalence of this condition to the following one: $\frac{a_1 + n_1 b}{k_1} = \frac{a_2 + n_2 b}{k_2}$, for proper integers $k_1 \neq 0$, $k_2 \neq 0$, n_1, n_2.

`Prove` that a linear operator B can be included in a one-parameter linear group g in $\mathbb{C}^n$ [that is, become $g(t_0)$ for some $t_0 \neq 0$] such that $g(\mathbb{R}) = \{E\}$ or $g(\mathbb{R}) \cong \mathbb{R}$ mod 2π if and only if the following conditions hold:

(a) B has its eigenvalues placed on the unit circle $\mathbf{S}^1 \subset \mathbb{C}$, that is, has all of them in the form $\cos\theta + i\sin\theta$ ($0 \leq \theta < 2\pi$);
(b) B is diagonalized in some basis (that is, all of its root subspaces are eigenspaces);
(c) Any pair of complex arguments of the eigenvalues distinct from one are commensurable modulo 2π.[4]

[4] The necessary and sufficient conditions for inclusion of an invertible linear operator in a one-parameter linear group are discussed in section P10.29[**] below. Here, a special kind of these groups is considered.

P10.13. E***

For a one-parameter linear group g with the elements satisfying conditions (a) and (b) from section P10.13. D***, `prove` that $g(\mathbb{R}) \cong \mathbb{R} \bmod 2\pi$ if and only if all pairs of "angular velocity" components ω_j (defined in section P10.13. C***) are linearly dependent with integral coefficients.

P10.13. F***

`Show` that a linear operator B, satisfying conditions (a) and (b) from section P10.13. D*** and having at least two eigenvalues of complex arguments incommensurable with 2π, can be included in a one-parameter linear group g such that $g(\mathbb{R}) \cong \langle \mathbb{R}, + \rangle$, even if B satisfies the aforementioned condition (c) from section P10.13. D***.[5]

Combining sections P10.13. D***, P10.13. F***, it follows that B can be included in two different types of groups.

P10.14**

`Prove` the existence of the unique one-parameter group $g(\mathbb{R}) \subseteq GL(n, \mathbb{R})$ having a given linear operator B as $g(t_0)$ for a fixed $t_0 \neq 0$, provided that B has positive and simple (nonmultiple) eigenvalues only. `Show` that, on the contrary, for B that is a dilatation operator λE_n (where $\lambda > 0$ or $\lambda < 0$ and n is even) there exist infinitely many such one-parameter groups for $n > 1$. `Provide` more examples of the uniqueness/nonuniqueness. (A general case is discussed in section P10.30** below.)

P10.15**

`Show` that a differentiable one-parameter linear group g [which means that $g(t)$ is differentiable with respect to t] satisfies a differential equation $\dot{g}(t) = Ag(t)$ $[A = \dot{g}(0)]$.

---

[5] In **ergodicity theory** it is proved that the trajectory of a uniform motion $t \mapsto (\omega_1 t, \ldots, \omega_n t)$ along an n-dimensional torus $\mathbb{R}^n \bmod 2\pi = \{(\theta_1, \ldots, \theta_n)\}$ is dense everywhere when ω_i are linearly independent over the integers. This shows that the topology of a subgroup $g(\mathbb{R}) \cong \langle \mathbb{R}, + \rangle$ within the group $GL(L)$ can be very complex. It is very different from the topology of a linear subspace within a vector space.

P10.16[**]

An exponential series for matrices is defined in the same way as for the numbers: $e^{At} = E + At + \ldots + (n!)^{-1}(At)^n + \ldots$. `Prove` that the Cauchy problem for the differential equation in section P10.15[**] using the initial condition $g(0) = E$ has a unique formal solution, namely, an exponential one.

Agreement. *In what follows, we will use the same designations for the exponential series and their sums*.

P10.17[**]

A series $\sum_{n=0}^{\infty} a_n$ of elements of a topological vector space L is said to converge when $\lim_{n\to\infty} \sum_{k=0}^{n} a_k$ exists, and this limit is called the series' sum. This means, for a finite-dimensional space, the convergence of the numerical series of coordinates with respect to a fixed basis $\lim_{n\to\infty} \sum_{k=0}^{n} a_{ki}$ exists for $i = 1, \ldots, \dim L$. The choice of the basis is unimportant: the series of the coordinates with respect to some basis converge if and only if the same is true for any basis. (`Prove.`) Therefore, the exponential series converges if the corresponding series of the matrix entries converge. Define convergence of exponential series for linear operators on a finite-dimensional space via the matrix representation with respect to a fixed basis of this space. `Verify` that a choice of the basis is unimportant.

`Prove,` for a series $g(t) = e^{tA}$, claims (1)–(3) from section P10.5[*]. `Derive` from this that the sum of this series is a solution of the Cauchy problem from section P10.16[**]: $\dot{g}(t) = Ag(t)$, $g(0) = E$.

`Find` the sums of exponential series explicitly for a diagonal matrix and for a nilpotent Jordan box $N = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}$.

P10.18[**]

`Show` that $e^{A+B} = e^A e^B$ if A and B commute. `Derive` from this that the sums of exponential series e^{tA} form one-parameter linear groups.

`Find` e^A explicitly for a Jordan box $A = \lambda E + N = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}$.

`Prove` that a map $A \mapsto e^A$ realizes a diffeomorphism of a neighborhood of zero in the space of linear operators on a finite-dimensional space onto a neighborhood of E.

`Show` that all one-parameter linear groups have a form of e^{tA} for some A, so they are smooth (infinitely differentiable).

The preceding equality for commuting matrices gives $e^A e^B e^{-A} e^{-B} = e^{A+B} e^{-(A+B)} = E = e^0 = e^{[A,B]}$, where $[A, B] = A \circ B - B \circ A$ (which is zero for commuting A, B). Therefore, e^A, e^B commute when A, B commute (which also may be verified directly; how?). `Establish` an extension of this relationship for arbitrary A, B:

$$e^{sA} e^{tB} e^{-sA} e^{-tB} = e^{st[A,B]} + o(s^2 + t^2) \qquad \text{as} \qquad s, t \to 0.$$

(The multiplicative commutator of exponents of matrices differs from the exponent of their additive commutator by a small value of order greater than two.) `Derive` that A commutes with B if $e^{sA} e^{tB} e^{-sA} e^{-tB} = E + o(s^2 + t^2)$ as $s,t \to 0$. `Determine` whether A, B always commute when e^A, e^B commute (that is, $e^A e^B e^{-A} e^{-B} = E$) or when A^m, B^n commute, for some positive integers m, n.

*P10.19***

`Prove` that $\lim_{n\to\infty} (E + A/n)^n$ exists for any square matrix and equals the sum of the exponential series e^A.

*P10.20***

A nice example of the exponential representation of a one-parameter group is the Taylor-series expansion. Consider a finite-dimensional vector space F consisting of functions f: $\mathbb{R} \to \mathbb{R}$ that are the sums of convergent power series $f(t) = a_{f,0} + a_{f,1}\, t + a_{f,2}\, t^2 + \ldots$, $\forall t$. These f are infinitely differentiable, $a_{f,n} = f_{[n]}(0)/n!$, $\forall n = 0,1,\ldots$ and, moreover, $f(t + s)$ equals the sum of the Taylor series $f(t) + f'(t) \cdot s + f''(t) \cdot s^2/2! + \ldots$. for any $s,t \in \mathbb{R}$. (`Work out` the details using section H10.5.)

(1) `Prove` that F is closed with respect to the shifts of the argument, $\forall s$: $f(.) \in F \Rightarrow g^s f(.) := f(. + s) \in F$, if and only if F is closed with respect to the

differentiation, $f(.) \in F \Rightarrow df/dt = \dot{f}(.) \in F$, and in this case there is the following relationship between linear operators $d/dt: F \to F$ and $g^t: F \to F$: $e^{s\frac{d}{dt}} = g^s$.

Examples of spaces F satisfying the conditions of this statement are spaces of the polynomials of degrees not exceeding a given number N and, more generally, spaces of **quasipolynomials**; these have the form $f = \sum \left(p_j(t)\cos\omega_j t + q_j(t)\sin\omega_j t\right)e^{\lambda_j t}$, where the sum is taken over a finite set of complex numbers $\lambda_j + i\omega_j$ (the same set of these numbers corresponds to all elements of F) and p_j, q_j are arbitrary polynomials of degrees not exceeding a given N_j (the numbers N_j are the same for all elements of F). (Below more advanced readers will be asked to establish that if F satisfies the conditions of this statement, then F is a space of quasipolynomials.)

(2) For the space of polynomials `calculate` matrices of the differentiation and shift operators with respect to the usual monomial basis. In particular, you will find that $\mathrm{tr}(d/dt) = 0$ and $\det g^t = 1$.

(3) For a space of trigonometric polynomials (which corresponds to $\omega_j = j$, $\lambda_j = 0, N_j = 0, \forall j$) `calculate` matrices of the differentiation and shift operators with respect to the basis consisting of trigonometric monomials $\{\cos jt, \sin jt\}_{j=0,\ldots}$ (this basis is orthogonal using a scalar product $\langle f_1, f_2\rangle = \int_0^{2\pi} f_1(t)f_2(t)\,dt$). Having completed these calculations, you will find the skew symmetry of d/dt and the orthogonality of g^t.

`Try` to generalize the observed correlations of the properties of d/dt and g^t without looking through the subsequent problems.

P10.21[***]

This problem is addressed to readers familiar with ordinary linear differential equations. `Prove` the following generalization of the claim in section P10.6[*], which is also a refinement of the statement in section P10.20[**]:

Proposition. *The following conditions on a finite-dimensional space F of functions* $\mathbb{R} \to \mathbb{R}$ *are equivalent:*

(i) *F consists of continuous functions and is closed with respect to the shifts of the argument.*
(ii) *F consists of differentiable functions and is closed with respect to the differentiation.*
(iii) *F is a space of quasipolynomials corresponding to some set of* $\lambda_j + i\omega_j$ *and* N_j (see the definitions in section P10.20[**]).

`Characterize` the finite-dimensional spaces of continuous functions on a unit circle closed with respect to the shifts of the argument modulo 2π.

P10.22[**]

For a linear operator A in a finite-dimensional space (or a matrix), `prove` that $\det(E + \varepsilon A) = 1 + \varepsilon \cdot tr\, A + O(\varepsilon^2)$ as $\varepsilon \to 0$.

More generally, let $W(\varepsilon)$ be a matrix-valued function such that $W(0) = E$ and $\frac{dW}{d\varepsilon}(0)$ exists; `prove` that $\det W(\varepsilon) = 1 + \varepsilon \cdot tr \frac{dW}{d\varepsilon}(0) + o(\varepsilon)$ as $\varepsilon \to 0$. From the preceding statement, `derive` for a differentiable function $W(t)$ taking values in invertible matrices that

$$\frac{d(\det W)}{dt} = tr(\dot{W} \circ W^{-1}) \cdot \det W. \qquad (*)$$

Readers familiar with ordinary linear differential equations will immediately recognize in the last equality the classic Liouville theorem about the Wronskian of a basic set of solutions. In geometric language, this theorem states that a momentary relative change of a spatial volume under the phase flow's transformation caused by a linear (time-dependent) **vector field** $v(t,x) = (\dot{W} \circ W^{-1})x$ equals the spatial divergence of this field. This geometric version holds for the nonlinear vector fields also and finds wide application in, for example, **Hamiltonian mechanics**, **ergodicity theory**, and the **physics of continuous mediums**. In its integral form, it describes changes of volumes (or, in other cases- masses, charges, and so on) under the phase flow's transformations:

$$\frac{d}{dt}\Big|_{t=t_0} \int\limits_{g^t_{t_0}U} \rho(t,x)\, dV(x) = \int\limits_{U} \left[\frac{\partial \rho}{\partial t}(t_0,x) + div(\rho(t_0,x) \cdot v(t_0,x))\right] dV(x),$$

for any time-dependent continuously differentiable density ρ and measurable spatial region U of finite volume; here $g^t_{t_0}$ is the spatial phase flow's transformation from moment of time t_0 till moment t, caused by vector field $v(t,x)$. Interested readers will find information about far-reaching further developments and discussions of related topics in Arnol'd (1989) (including a supplementary chapter "Poisson Structures") and references therein and multiple sources devoted to ergodicity theory and its applications.

P10.23[**]

As readers probably know, the space of matrices $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ with real entries is closed, with respect to the matrix composition, so it is a subalgebra of the algebra of the 2×2 matrices, and in fact it is a field isomorphic to the field of complex numbers $\mathbb{C} = \{a + bi\colon a, b \in \mathbb{R}\}$.[6] (`What` corresponds to the modulus and the

[6] Readers not familiar with this fact may prove it by themselves or turn for the discussion to section H6.9 (in the "Polar and Singular Value Decomposition Theorems" problem group above).

argument of a complex number?) Show that $e^{\begin{pmatrix} a & -b \\ b & a \end{pmatrix}}$ corresponds to e^{a+bi}, where the exponential function is defined for the complex numbers as a sum of the exponential series, in the same way as for the reals.

P10.24[**]

Prove that for $n \to \infty$

$$\arg(1 + z/n) = \mathrm{Im}z/n + o(1/n), \qquad |1 + z/n| = 1 + \mathrm{Re}z/n + o(1/n).$$

Derive that $\lim\limits_{n\to\infty} n \cdot \arg(1 + z/n) = \mathrm{Im}z$ and $\lim\limits_{n\to\infty} |1 + z/n|^n = e^{\mathrm{Re}z}$. Using these limit relationships and the isomorphism $\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \leftrightarrow a + bi$ discussed in section P10.23[**], establish the famous **Euler's formula** $e^z = e^{\mathrm{Re}z}(\cos \mathrm{Im}z + i \sin \mathrm{Im}z)$ (Arnol'd 1975).

Also, establish this formula in a different way using the preceding isomorphism from section P10.23[**] and the statement in section P10.13. C[***].

In addition, Euler's formula may be established using exponential series and Taylor-series expansions of cosine and sine. (Work out the details.[7])

Describe the set $\mathrm{Log}\, u = \{z \in \mathbf{C} : \ e^z = u\}$ $(u \in \mathbf{C}^* = \mathbf{C}\backslash\{0\})$.

P10.25[**]

Alternatively, Euler's formula in section P10.24[**] may serve as a definition of sine and cosine. Use the statements in section P10.24[**] to prove that the functions $\mathrm{Re}\, e^{i\theta} = \frac{e^{i\theta}+e^{-i\theta}}{2}$ and $\mathrm{Im}e^{i\theta} = \frac{e^{i\theta}-e^{-i\theta}}{2i}$ coincide with $\cos\theta$ and $\sin\theta$.

[7] For this, readers should verify absolute (hence, commutative) convergences of those series, and the equalities of cosine and sine to the sums of their Taylor series, which might be done as follows. Prove that a function $f(t)$ infinitely differentiable for $|t - t_0| \le r$ has its Taylor series at t_0 absolutely convergent and is equal to the sum of this series at any point of that interval if the derivatives all together are uniformly bounded on any segment $|t - t_0| \le r_0 < r$: $\forall r_0 \in (0,r)$, that is, $\exists C = C(r_0)$ such that $|f^{[n]}(t)| \le C$ for $|t - t_0| \le r_0$ and $n = 0,1,\ldots$ [actually, if $\forall r_0 \in (0,r)$, then $f^{[n]}(t) \cdot {r_0}^n/n! = o(1)$ as $n \to \infty$, uniformly on $|t - t_0| \le r_0$].

This part of the problem is addressed to readers familiar with Cauchy-Riemann equations for complex holomorphic functions and ordinary differential equations (ODEs). `Show` that Re $e^{i\theta}$, Im $e^{i\theta}$ coincide with cos θ, sin θ, respectively, as being solutions of the same Cauchy problems as for those functions; `determine` the Cauchy problems for Re $e^{i\theta}$, Im $e^{i\theta}$ using the methods described previously for the exponential functions (section P10.4* above).

*P10.26**

An additional way of defining cosine and sine comes from the argument's addition formulas. `Prove` that $C = \cos t$, $S = \sin t$ are unique solutions of the system of functional equations

$$C(s+t) = C(s)\,C(t) - S(s)\,S(t), \qquad S(s+t) = S(s)\,C(t) + C(s)\,S(t),$$

defined on $(-\infty,\infty)$ and normalized by the conditions $C^2 + S^2 = 1$, $S(0) = 0$, $S(\pi/2) = 1$ and $S(t) > 0$ for $0 < t < \pi/2$ (Rabbot 1969).

*P10.27***

Previously, in section P10.20**, readers came across an example of e^A orthogonal to a skew-symmetric A. A similar example can be derived with Euler's formula from section P10.24**, resulting in $e^{\begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$.

1. `Prove`, for any skew-symmetric linear operator on a Euclidean space (or a skew-symmetric matrix) A, that e^A is always orthogonal: $A + {}^tA = 0 \Rightarrow e^A \circ e^{{}^tA} = E$ (E is the identity).
2. Regarding a conversion, let log A be a solution of the equation $e^X = A$; `show` that an orientation-preserving orthogonal linear operator has a skew-symmetric logarithm.
3. `Are` all of those logarithms skew-symmetric? `Describe` all of $\log\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\log R_\theta$, for a rotation $R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ $(0 < \theta < 2\pi)$.
4. For an orthogonal operator close to E, `establish` the unique existence of a logarithm close to 0 and its skew symmetry.
   The rest of the section is addressed to readers familiar with the language of differentiable manifolds. The last statement shows that the restriction of the exponential map to a

neighborhood of zero in the space of the skew-symmetric matrices $o(n) = \{A : {}^tA + A = 0\}$ is a homeomorphism onto a neighborhood of E in the group of the orthogonal matrices (the **orthogonal group**) $O(n) = \{B : {}^tB \circ B = E\}$.[8]

5. Using the preceding statement, `prove` that $O(n)$ is a smooth submanifold of the space of all linear operators on $\mathbb{R}^n$, of the same dimension as $o(n)$. (`What` is this dimension?)

For a basis of the space of skew-symmetric matrices $\{A = (a_{ij})_{i,j=1,\ldots,n} : a_{ij} + a_{ji} = 0\}$ formed by the elements $A = I^{(ij)}$ with the entries $a_{km}^{(ij)} = \begin{cases} 1 & \text{for } k = i,\ m = j \\ -1 & \text{for } k = j,\ m = i \\ 0 & \text{otherwise} \end{cases}$ $(1 \le j < i \le n)$, one-parameter linear groups $g^{(ij)} : t \mapsto e^{tI^{(ij)}}$ consist of rotations (by angles t mod 2π) in two-dimensional coordinate planes in $\mathbb{R}^n$ spanned on vectors e_i, e_j ($e_k = (0, \ldots \underset{k}{1} \ldots, 0)$), retaining the points of the corresponding orthogonal $n-2$-dimensional subspaces ("the axes") fixed. (A convenient description is provided by exterior algebra's concept discussed previously in the problem group "A Property of Orthogonal Matrices.". According to this concept, those planes correspond to the orthogonal basis of $\wedge^2\mathbb{R}^n$ consisting of $e_i \wedge e_j$.) Readers may verify by direct computations using the implicit function theorem that a map $(t_{21}, \ldots, t_{n1}, t_{32}, \ldots) \mapsto g^{(21)}(t_{21}) \circ \ldots \circ g^{(n1)}(t_{n1}) \circ g^{(32)}(t_{32}) \circ \ldots$ is a diffeomorphism of a neighborhood of the origin in $\mathbf{R}^{\binom{n}{2}}$ onto a neighborhood of E in $O(n)$. Applying finer methods of Lie group theory and accounting for the connectedness of $SO(n)$, we can show that this group is globally generated by its subgroups $g^{(ij)}(\mathbb{R})$ (but in no way being their direct product; note that matrices $I^{(ij)}$ do not commute). In solid body mechanics the parameters $\theta_{ij} = t_{ij}$ mod 2π of an element of $SO(n)$ (corresponding to a fixed coordinate system in the source space $\mathbb{R}^n$) are referred to as its **Euler angles**. In turn, $SO(n)$ is the configuration space for the **Euler** solid body's dynamic **equations** (in $\mathbb{R}^n$).

Let us generalize from the preceding example. Consider a nondegenerate bilinear form on a finite-dimensional space $\langle .,. \rangle : L \times L \to$ field of scalars. As readers may verify, the nondegeneracy may be formulated in three equivalent ways: (1) $\langle x, y \rangle = 0,\ \forall y \Rightarrow x = 0$; (2) $\langle x, y \rangle = 0,\ \forall x \Rightarrow y = 0$; (3) the Gram matrix $\left(\langle e_i, e_j \rangle\right)_{i,j=1,\ldots,\dim L}$ with respect to some (and thus, any) basis is nondegenerate.

6. `Prove` the equivalence of these conditions.[9]

[8] In fact, $\exp(o(n))$ lies in a connected component of E, which is a **special orthogonal group** $SO(n) = \{B \in O(n) : \det B = 1\}$ consisting of orientation-preserving matrices.

[9] The equivalence of (1) and (2) was discussed in section H8.11 related to the problem group "A Property of Orthogonal Matrices" above.

7. For a linear operator A, an adjoint operator A^* is defined via the identity $\langle Ax, y\rangle = \langle x, A^*y\rangle, \forall x,y$.[10] `Establish` the unique existence of A^* for a nondegenerate bilinear form. `Verify` that $A \mapsto A^*$ is a linear operator (in a vector space of the linear operators on L) and also that $E^* = E$, $(A \circ B)^* = B^* \circ A^*$, $(A^*)^{-1} = (A^{-1})^*$. Hence, $e^{(A^*)} = (e^A)^*$. (`Prove.`)
Warning: in general, $A^{**} \neq A$; readers may `prove` that $A^{**} = A$ for any linear operators if and only if the source bilinear form is either symmetric or skew-symmetric.

8. `Show` that a subset g: $= \{A : A^* + A = 0\}$ of the space gl(L) of all linear operators on L is a vector subspace, closed with respect to taking commutators: $A_1, A_2 \in \mathrm{g} \Rightarrow [A_1, A_2] = A_1 \circ A_2 - A_2 \circ A_1 \in \mathrm{g}$ (which means that g is a **Lie algebra**). `Show` that a subset $G := \{B : B^* \circ B = E\}$ of $GL(L)$ is a topologically closed subgroup. `Prove` the inclusion $\exp(\mathrm{g}) \subseteq G$.
G is called a group of (linear) automorphisms of the form $\langle\rangle$ since $\langle Bx, By\rangle = \langle x,y\rangle, \forall x,y \in L$ ($B \in G$). In other words, for a matrix of the form with respect to any fixed basis, F, and matrices of the elements of the group G with respect to the same basis, A, ${}^tA \circ F \circ A = F$. (`Verify.`)

9. One of the basic theorems of **Lie group** theory states that if a subgroup H of $GL(L)$ is closed, then it is a submanifold, a subset h: $= \{A : e^{tA} \in H, \forall t \in \mathbf{R}\}$ of gl(L) is a vector subspace in it, and exp(h) contains a neighborhood of E in H. {We cannot discuss a proof within the scope of the present book; interested readers may try to find one on their own or refer to other sources, for example, Chevalley (1946–1955), Serre (1965), Adams (1969), Warner (1983), Bourbaki (1968–1982).] Applying this theorem to $H = G$ and taking into account the aforementioned section P10.18**, `prove` that the restriction of the exponential map to a neighborhood of zero in g is a diffeomorphism onto a neighborhood of E in G. [Of course, in this case h as defined in the theorem equals g as defined previously in (8).] Use it to `show` that G is a smooth submanifold of the same dimension as g in the space of all linear operators.
Besides the orthogonal groups there are other special cases similarly constructed that have many applications. First are the **unitary groups** $U(n)$ and **symplectic groups** $SP(2n)$, consisting of the linear automorphisms of, respectively, sesquilinear **Hermitian forms** on complex spaces $\mathbb{C}^n$ and **symplectic forms** on even-dimensional spaces $\mathbb{R}^{2n}$. On $\mathbf{C} = \{z = x + iy\}$, consider a function $\langle z', z''\rangle := z' \cdot \bar{z''} = (x'x'' + y'y'') + i(x''y' - x'y'')$. It is bilinear with respect to scalar real factors and sesquilinear with respect to scalar complex ones, that is, $\langle \alpha z', z''\rangle = \alpha\langle z', z''\rangle$ and $\langle z', \alpha z''\rangle = \bar{\alpha}\langle z', z''\rangle$ for $\alpha \in \mathbb{C}$. The real and imaginary components of this form are, respectively, a symmetric positive definite and a nondegenerate skew-symmetric (second

[10] Alternatively, it may be defined via an identity $\langle x, Ay\rangle = \langle A^*x, y\rangle, \forall x,y$, which in general is not equivalent unless the bilinear form is either symmetric or skew-symmetric.

name: **symplectic**) forms on the realification $\mathbb{R}^2$ of $\mathbb{C}$.[11] Conversely, taking one of the forms $(z', z'') = x'x'' + y'y''$ or $[z', z''] = x''y' - x'y''$ on $\mathbf{R}^2_{x,y}$ and a complex structure agreeing with it (so that $(i \cdot ., i \cdot .) = (., .)$ or, respectively, $[i \cdot ., i \cdot .] = [., .]$), turning $\mathbb{R}^2$ into $\mathbb{C}$, there exists a unique sesquilinear form on $\mathbb{C}$ having the source form on $\mathbb{R}^2$ as its real (resp. imaginary) component. That is, for $z = xe_1 + ye_2$, setting $(a + ib)\cdot z := az + bIz$, where $I = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ with respect to an orthogonal well-oriented basis e_1, e_2 in $\mathbb{R}^2$, yields the sesquilinear form $\langle z', z'' \rangle = (z', z'') + i \cdot (z', iz'') = [iz', z''] + i \cdot [z', z'']$.

10. $U(1)$ is defined as a subgroup of $GL(1,\mathbb{C}) \cong \mathbb{C}^*$ consisting of elements preserving $\langle , \rangle$. `Prove` that $U(1)$ coincides with a scalar group of multiplications by complex numbers placed on a unit circle and, therefore, $U(1) \cong SO(2)$.

11. $SP(2)$ consists of the elements of $GL(2,\mathbb{R})$ preserving [,]. `Show` that $SP(2)$ coincides with the **unimodular** group $SL(2, \mathbf{R}) = \{B \in GL(2, \mathbf{R}) : \det B = 1\}$. `Prove` that topologically it is a solid torus: $SP(2) = SL(2,\mathbb{R}) \cong \mathbb{R}^2 \times \mathbf{S}^1$. `Is` this homeomorhism an algebraic isomorphism as well? Therefore, dim $SP(2) = 3$. In particular, $SP(2)$ contains, but does not coincide with, $U(1)$. The elements of $SP(2) \backslash U(1)$ preserve [,] and substitute a complex structure on $\mathbb{R}^2$ and the corresponding Hermitian form. Actually, $U(1)$ consists of the transformations preserving both real and imaginary components of the Hermitian form, so $U(1) = SP(2) \cap O(2) = SO(2)$. This statement can be generalized for any dimension; thus, $U(n) = SP(2n) \cap O(2n)$ [with regard to the orientation-preserving property of symplectic transformations, $U(n) = SP(2n) \cap SO(2n)$]. Also, $U(n) = SP(2n) \cap GL(n,\mathbb{C}) = GL(n,\mathbb{C}) \cap O(2n)$, which means that unitary transformations are those preserving the symplectic and complex structures and those preserving the complex and Euclidean structures.

$SL(2,\mathbb{R})$ [or $SP(2)$] is one of the most applicable groups in various branches of mathematics such as geometry, number theory, functional analysis, and mathematical physics. $SL(2,\mathbb{R})$ is the motion group for the Lobachevsky plane, and so the group of conformal transformations of open upper half-plane (hence, also of open unit disc and, generally, of any 1-connected domain on a Gaussian sphere $\bar{\mathbf{C}}$, of a boundary of more than one point). $SL(2,\mathbb{R})$ is the projective (linear-fractional) group of the projective space $\mathbb{R}P^1$. $SL(2,\mathbb{R})$ is the group of symmetries of the relativity theory for the "shortened" three-dimensional space-time. Readers will find further information and related discussions in the vast literature on those topics. For classic and modern ideas connected to representations of $SL(2,\mathbb{R})$ (mostly infinite-

[11] Geometrically, the value of the form [,] on a pair of vectors equals the oriented area of a parallelogram spanned on these vectors. Using the symbolism of skew-symmetry algebra (as discussed in the problem group "A Property of Orthogonal Matrices" above), $[,] = x \wedge y$ where linear functionals x, y are Cartesian coordinates.

dimensional) we encourage advanced readers to consult Lang (1975) and references therein.

12. `Prove` that nondegenerate skew-symmetric (symplectic) forms do not exist on $\mathbb{R}^{2n+1}$ and that on $\mathbb{R}^{2n}$ any such form is brought to $[z', z''] = \sum_{j=1}^{n} x''_j y'_j - x'_j y''_j$ with respect to an appropriate basis (called a **Darboux basis**). `Show` that any symplectic form [,] is the imaginary component of a sesquilinear Hermitian form on $\mathbb{C}^n$ obtained by equipping $\mathbb{R}^{2n}$ with a complex structure that agrees with [,].[12]
`Prove` that symplectomorphisms (symplectic transformation) preserve spatial orientation, in other words, $[.,.] = [B(.),B(.)] \Rightarrow \det B = 1$. Also, `prove` that the characteristic ($\chi_B(z) = \det(zE - B)$) polynomials of symplectomorphisms are reciprocal [$\chi_B(z) = z^{2n} \cdot \chi_B(z^{-1})$], in other words, that substituting z by z^{-1} and simultaneously multiplying by z^{2n} does not affect $\chi_B(z)$ or, equivalently, that the polynomial coefficients equidistant from the ends of the polynomial are equal. Therefore, symplectomorphisms have their spectra symmetric with respect to both the real axis and unit circle in $\mathbb{C}$.

13. `Calculate` dim $U(n)$ and dim $SP(2n)$ equating them to, respectively, dim u (n) and dim sp($2n$).
Warning: for $n > 1$, $U(n)$ and $SP(2n)$ differ from, respectively, $SO(2n)$ and $SL(2n,\mathbb{R})$.
A slight generalization of the foregoing construction yields **conformal** groups. Define, for a nondegenerate form,
$CG := \{B : B^* \circ B = \mu_B E, \text{ with a scalar } \mu_B\}$,
$\text{cg} := \{A : A^* + A = \nu_A E, \text{ with a scalar } \nu_A\}$.

14. `Show` that connections of ν_A to operator A and μ_B to operator B are invariant (do not depend on a spatial basis). `Verify` that CG is a closed subgroup of $GL(L)$, cg is a vector subspace (in fact, a Lie subalgebra) of gl(L), and exp (cg) $\subseteq CG$.

15. `Verify` the multiplicative dependence of μ_B on $B \in CG$. Thus, $B \mapsto \mu_B$ is a homomorphism from CG to a multiplicative group of the field of the scalars. (`What` is its kernel?) Hence, $\mu_{B^{-1}} = \mu_B^{-1}$. (Also `verify` it directly). `Show` that $B^* \circ B = B \circ B^*$, $B^{**} = B$, and $\mu_{B^*} = \mu_B$ if $B \in CG$.

16. `State` and `prove` the additive analogs of these claims regarding ν_A.

17. `Show` that for $B \in CG$ close to E, μ_B is close to 1 (specifically, for linear operators over the reals those μ_B are positive). `Prove` using the theorem from Lie group theory cited previously in **(9)** that the restriction of the exponential map to a neighborhood of zero in cg is a diffeomorphism onto a neighborhood of identity in CG. Use it to `show` that CG is a smooth submanifold of the space of all linear operators of the same dimension as cg.

[12] Geometrically, [,] is a sum of the oriented area elements on two-dimensional planes where planes in any pair are mutually orthogonal. These planes are invariant with respect to realification of the multiplication by i: $[,] = \sum x_i \wedge y_i$.

18. `Prove` that for any nondegenerate bilinear form on a vector space over the reals $\mu_B > 0$, $\forall B \in CG$. Also, `prove` that $\forall B \in CU(n)$ μ_B are real and positive.

19. Obviously, CG contains both G and the group of dilatations $D = (field\backslash\{0\})\cdot E$ and, thus, contains the group generated by them. `Show` that this inclusion is equality: $CG = D \cdot G = G \cdot D$.[13] [Therefore, the elements of $CO(n)$ are compositions of dilatations and orthogonal transformations.]

20. The elements of $CO(n)$ are referred to as **conformal transformations** because of their obvious "form-preserving," in other words, similarity-preserving or angle-preserving, property. `Prove` that for dimensions greater than one no other transformations, including nonlinear ones, possess this property. More exactly, `prove` that for $n > 1$ an orthogonality-preserving (right-angle-preserving) one-to-one map T: $\mathbb{R}^n \to \mathbb{R}^n$ ($y - x \perp z - x \Rightarrow Ty - Tx \perp Tz - Tx$, $\forall x,y,z$) is a composition of a parallel translation and a linear transformation from $CO(n)$.

P10.28[**]

In section P10.13 it was discussed that $\det e^A > 0$; in section P10.20[**] readers came across an example of $\det e^A = 1$ for $\operatorname{tr} A = 0$. `Prove` the generalizing statement $\det e^A = e^{\operatorname{tr} A}$. (Therefore, $\det e^A > 0$ and $\operatorname{tr} A = 0 \Leftrightarrow \det e^A = 1$.)[14]

This part of the problem is addressed to readers familiar with the language of differentiable manifolds. `Prove` that a special linear group $SL(n,\mathbb{R}) = \{B\colon \det B = 1\}$ is a smooth manifold and find its dimension.

P10.29[**]

`Prove` that

$\log B$ for $B \in GL(n,\mathbb{C})$ always exists in $\mathrm{gl}(n,\mathbb{C})$;

$\log B$ and $\sqrt{B}$ for $B \in GL(n,\mathbb{R})$ exist in $\mathrm{gl}(n,\mathbb{R})$ (with this, obviously, $\sqrt{B} \in GL(n,\mathbf{R})$) if and only if the Jordan boxes of B corresponding to the negative eigenvalues exist in equal pairs.

---

[13] This group is not the direct product $D \times G$, as, for example, $D \cap O(n) = \{E,-E\} \cong \mathbb{Z}(2)$ [for even n, $D \cap SO(n) = \{E,-E\}$] and $D \cap U(n) \cong \mathbf{S}^1$.

[14] Readers familiar with exact sequences and commutative diagrams may interpret this statement as follows: a diagram with rows

$$\begin{array}{ccccccccc} \{0\} & \to & sl(n,\mathbf{R}) & \to & gl(n,\mathbf{R}) & \overset{tr}{\to} & \mathbf{R} & \to & \{0\} \\ & & \downarrow \exp & & \downarrow \exp & & \downarrow \exp & & \\ \{E\} & \to & SL(n,\mathbf{R}) & \to & GL(n,\mathbf{R}) & \overset{\det}{\to} & \mathbf{R}^{+*} & \to & \{1\} \end{array}$$

is commutative. (`Draw` a similar diagram for the case of a complex field.) In Lie group theory this and similar diagrams express the functoriality of an exponential map.

[(According to section P10.18**, B has a logarithm if it can be included in a one-parameter linear group. Conversely, the existence of a logarithm implies the existence of the powers $B^t = e^{t \log B}$ $(t \in \mathbb{R})$,[15] which means that B can be included in a one-parameter linear group.]

`Find` necessary and sufficient conditions for the existence of $\sqrt{[p]}\,B$ (including the case where B is not invertible).

*P10.30***

How many one-parameter linear groups g exist given $g(t_0) = B \in GL(n,.)$ for a fixed $t_0 \neq 0$? `Prove` that there are infinitely many of these groups for g in $\mathbb{C}^n$. `Show` that the same holds for g in $\mathbb{R}^n$ when $B \in GL(n,\mathbb{R})$ and B satisfies the necessary and sufficient conditions from section P10.29**, with the exception of those B that satisfy the following two conditions: (A) all their eigenvalues are positive and (B) they have no more than one Jordan box of any dimension corresponding to the same eigenvalue.

*P10.31***

`Establish` the following version of **Gelfand's formula**:

$$\max_{\lambda \in Spec(A)} \mathrm{Re}\lambda = \lim_{t \to \infty} t^{-1} \ln \left\| e^{A t} \right\|$$

for any fixed matrix norm ||·||.

Hint

H10.0

Show that the image of a group homomorphism g: $A \to B$ is always a subgroup of B. Therefore, if im $g \subseteq GL(L)$, then g is a group homomorphism $\langle \mathbb{R},+\rangle \to \langle GL(L), \circ\rangle$, and so $g^0 = E$. Conversely, $g^0 = g^t \circ g^{-t} = E \Rightarrow g^t$ is invertible.

H10.1

$g(t)$ is a continuous function that has at least one positive value (at $t = 0$), and thus, having a negative value as well, it would also take a zero value! Advanced readers

[15] And the powers $B^z = e^{z \log B}$ $(z \in \mathrm{X})$, for complex matrices.

preferring the same arguments in a different package, using such terms as connectedness or arcwise connectedness, will find them in sections H6.8 and H12.1, which correspond to the aforementioned problem group "Polar and Singular Value Decomposition Theorems" and the subsequent "Least Squares and Chebyshev Systems" problem group.

H10.2

A value of $g(t_0)$ determines all of $g(nt_0)$: $n \in \mathbb{Z}$. This value determines also all of g (pt_0/q): $p/q \in \mathbb{Q}$ because of the existence of the unique positive roots of positive numbers. By virtue of continuity, it determines all of $g(t)$: $t \in \mathbb{R}$.

H10.3

Differentiate identity (*) with respect to s at $s = 0$ taking into account linearity and the continuity of multiplication.

H10.4

Inserting a formal power series with indeterminate coefficients $\sum a_n t^n$ into the initial condition and into the differential equation from section P10.3* yields $a_0 = 1$ and the formal equality

$$a_1 + 2a_2 t + 3a_3 t^2 + \ldots = k(a_0 + a_1 t + a_2 t^2 + \ldots),$$

from which, by induction, $a_1 = k$, $a_2 = k^2/2$, …, $a_n = k^n/n!$,….

H10.5

Provided that conditions (1)–(3) are satisfied, the sum of $g(t)$ will be a solution of the Cauchy problem in section P10.4* because of the continuity of the equation with respect to $g(t)$ and $\dot{g}(t)$ and the fact that taking a limit is a linear operation. Specifically, equalities of partial sums $k \sum_{i=1}^{n+1} [(i-1)!]^{-1} (kt)^{i-1} = k \cdot \sum_{i=0}^{n} (i!)^{-1} (kt)^i$ become the equality $\dot{g}(t) = kg(t)$ as $n \to \infty$.

In turn, (1)–(3) may be proved as follows: an exponential series has an infinite convergence radius; hence, it converges on the entire $\mathbb{R}$, and thus its sum is a so-called **analytic function** on $\mathbb{R}$. It is infinitely differentiable and satisfies (1)–(3). For the exponential series, condition (2) obviously follows from (1) since the series of

derivatives indexed using shift by one is term-by-term proportional to the source series. You can prove conditions (1)–(3) in the following three steps A–C:

A. A numerical series $\sum_{n=0}^{\infty} a_n$ is **converging absolutely** if $\sum_{n=0}^{\infty} |a_n|$ converges. (Absolute convergence implies convergence, but not the inverse; provide examples.) Derive the **d'Alembert test** for absolute convergence, or, which is the same thing, for the convergence of series with nonnegative elements a_n: *such a series converges if* $|a_{n+1}/a_n| < 1 - \varepsilon$ *for any large enough n and some* $\varepsilon > 0$. Use this test to obtain the (absolute) convergence of the exponential series $g(t)$ and the series $\dot{g}(t)$ for any $t \in \mathbb{R}$.

B. A series of functions $\sum_{n=0}^{\infty} a_n(t)$ is referred to as converging uniformly to a sum $s(t)$ when $\lim_{n\to\infty} \sum_{k=0}^{n} a_k = s$ uniformly on t, that is, $\forall \varepsilon > 0$, $\exists n$, $\forall m \geq n$, $\forall t$: $\left| s(t) - \sum_{k=0}^{m} a_k(t) \right| < \varepsilon$. Prove the famous

Abel's lemma *A power series* $\sum c_n t^n$ *convergent at a point* $t_0 \neq 0$ *converges, together with its series of the derivatives* $\sum (n+1)\, c_{n+1}\, t^n$, *absolutely and uniformly on* $[-t_1, t_1]$ *for any* $t_1 < |t_0|$.

(The same holds for complex series, replacing the segment by a closed disc of radius t_1 in $\mathbb{C}$, and in more general situations.) Therefore, the exponential series $g(t)$ and the series of the derivatives $\dot{g}(t)$ converge uniformly on compact subsets of $\mathbb{R}$ (and $\mathbb{C}$). (Why?)

C. Let us assume that all elements of a series $\sum a_n(t)$ are differentiable in a neighborhood of a segment $[t_1,t_2]$. Prove the following statement:

If the series of term-by-term derivatives $\sum \dot{a}_n(t)$ *converges uniformly on* $[t_1,t_2]$ *and* $\sum a_n(t)$ *itself converges at least at one point of this segment, it will converge uniformly on the whole segment and the sum will be differentiable, with the derivative of the sum equal to* $\sum \dot{a}_n(t)$.

(Perhaps more advanced readers are familiar with generalizations of this claim.) This will complete the proof of (1)–(3) for the exponential series.

H10.6

Let S, T be the sums of the exponential series in s, t, respectively. We must prove that ST equals the sum of $\sum (n!)^{-1}(s+t)^n$. For partial sums $S_n = \sum_{k=0}^{n} (k!)^{-1} s^k$, $T_n = \sum_{k=0}^{n} (k!)^{-1} t^k$ we will have $S_n T_n \to ST$. Show that $S_n T_n - \sum_{k=0}^{n} (k!)^{-1}(s+t)^k \;\to\; 0$.

To prove that sums of the exponential series are all continuous solutions of (*), we suggest proceeding by either of the following two methods:

Method 1. For small t, show the unique existence of small $K(t)$ such that $g(t) = e^{K(t)}$, where e^x is the sum of exponential series $\sum(n!)^{-1}x^n$. From satisfying (*) by the exponential series' sums derive a linear dependence $K(t) = kt$ for small t. Therefore, $g(t) = e^{kt}$, at least for small t. But a one-parameter group g is completely determined by its restriction on a neighborhood of zero (why?), so $g(t) = e^{kt}$ for all $t \in \mathbb{R}$.

Method 2. Deduce the desired result from the existence and uniqueness, for given $t_0 \in \mathbb{R}$ and $b > 0$, of k such that $e^{kt_0} = b$ (the existence is discussed in section P10.8* below and the uniqueness was discussed previously in section P10.2*).

H10.7

We suggest proceeding by either of the following two methods:

Method 1. Show that a difference $e^t - (1 + t/n)^n = \sum_{k=1}^{\infty} \left[1 - \frac{n(n-1)\ldots(n-k+1)}{n \cdot n \cdot \ldots \cdot n}\right] \cdot \frac{t^k}{k!}$ for fixed t [this series converges as the exponential series converges and $(1 + t/n)^n$ is a polynomial] tends to zero as $n \to \infty$.

Method 2. Use the properties of logarithms discussed in section P10.10* below: satisfying equation (**), differentiability (hence, continuity), and having a determinate value of the derivative at 1:

$$\ln \lim_{n\to\infty} (1 + t/n)^n = \lim_{n\to\infty} \ln (1 + t/n)^n = t \cdot \lim_{n\to\infty} \frac{\ln(1+t/n) - \ln 1}{t/n} = t \cdot \frac{d \ln x}{dx}(1) = t.$$

QED.

H10.8

(1) $e^{-t} = 1/e^t$ by virtue of equation (*) in section P10.0*

(2) $\lim_{t\to\infty} e^t = \infty$ since $e^t = \sum(n!)^{-1}t^n \geq 1 + t$ for $t \geq 0$

(3) $\lim_{t\to-\infty} e^t = 0$, as a result of (1) and (2)

(4) e^t is a monotonically growing function since $d(e^t)/dt = e^t > 0$ (use sections P10.3* and P10.1*)

H10.9

We have $(a^s)^t = e^{t \ln e^{s \ln a}} = e^{t(s \ln a)} = e^{(st) \ln a} = a^{st}$. The remaining verifications in section P10.9* are also straightforward.

H10.10

The logarithmic functions are inversions of the exponential functions. Therefore, they are solutions of (**). The implicit function theorem shows the differentiability of the logarithms and makes it possible to derive a differential equation for them from the equation in section P10.3* for the exponential functions: namely, for $s = f(t) = \log_a t$ and $t = g(s) = a^s$,

$$\dot{f}(t) = 1/\dot{g}(s) = 1/kg(s) = 1/kt \quad (k = \ln a).$$

(The same equation is satisfied with $\log_a | t |$.) Furthermore, a short computation $2f(-t) = f(t^2) = 2f(t)$ shows that the extension from $\mathbb{R}^{+*}$ to $\mathbb{R}^*$ of the solution $f(t)$ may be only $f(|t|)$; on the other hand, $f(|t|)$ will satisfy (**). Lastly, we must prove that all continuous nonzero solutions of (**) defined on $\mathbb{R}^{+*}$ are logarithms. For the continuous solutions, $f(e^t)$ are continuous homomorphisms $\langle \mathbb{R}, + \rangle \to \langle \mathbb{R}, + \rangle$, that is, linear functions over the reals: $f(e^t) = ct$ ($c \in \mathbb{R}$). (Why?) For $c \neq 0$ we have, denoting $k = c^{-1}$ and applying the statement from section P10.9*,

$$f(e^t) = t/k = \log_{e^k} e^t,$$

whence $f = \log_{e^k}$ (why?), which completes the proof.

H10.11

For a solution $g(t)$ of equations (*) in section P10.0*, obviously, $\varphi = \ln g(t)$ is a homomorphism $\langle \mathbb{R}, + \rangle \to \langle \mathbb{R}, + \rangle$ and $g(t) = e^{\varphi(t)}$. Conversely, for a homomorphism $\langle \mathbb{R}, + \rangle \xrightarrow{\varphi} \langle \mathbb{R}, + \rangle$, $g(t) = e^{\varphi(t)}$ satisfies (*). Therefore, $\varphi(t)$ is continuous if and only if $g(t)$ is continuous. Discontinuous homomorphisms $\langle \mathbb{R}, + \rangle \xrightarrow{\varphi} \langle \mathbb{R}, + \rangle$ were discussed previously in section H9.6 (problem group "Convexity and Related Classical Inequalities").

H10.13. A

The orientation-preserving property of the elements of one-parameter linear groups in $\mathbb{R}^n$ may be established for arbitrary n in the same manner as for $n = 1$ (discussed previously in section H10.1). The only difference is that the function $\det g(t)$ should be applied instead of $g(t)$.

H10.13. B

The kernel of the homomorphism g: $\langle \mathbb{R},+\rangle \to GL(L)$ is closed. Prove that $\langle \mathbb{R},+\rangle$ has closed subgroups of three kinds: $\{0\}$, **cyclic subgroups** (**discrete lattices**) $T\mathbb{Z} = \{Tn\colon n \in \mathbb{Z}\}$ ($T \neq 0$), and $\langle \mathbb{R},+\rangle$ itself. Or $g(\mathbb{R})$ will be isomorphic to $\langle \mathbb{R},+\rangle$, isomorphic to $\mathbb{R} \bmod 2\pi$, and equal to $\{E\}$.

H10.13. C

The existence of a common eigenspace for all of $g(t)$ can be established by an inductive procedure based on the lemma about commuting linear operators from section E6.10 (the problem group "Polar and Singular Value Decomposition Theorems" discussed previously) using the existence of eigenvalues in an algebraically closed field and considering that the dimension of the space cannot be reduced indefinitely. (Work out the details.) Next, on an eigenspace corresponding to an eigenvalue $\lambda_j(t) = \rho_j(t)\big(\cos\theta_j(t) + i\sin\theta_j(t)\big)$ the group's action is described as $g^t v = \lambda_j(t)v$, which imposes on $\lambda_j(t)$ equations (*) from section P10.0*, and thus it defines two continuous homomorphisms: a multiplicative, $\begin{cases} \langle \mathbb{R},+\rangle & \to & \langle \mathbb{R}^*, g\rangle \\ t & a & \rho_j(t) \end{cases}$, and an additive, $\begin{cases} \langle \mathbb{R},+\rangle & \to & \mathbb{R} \bmod 2\pi \\ t & a & \theta_j(t) \end{cases}$. The formula $\rho_j(t) = e^{k_j t}$ follows from section P10.6*. In turn, the formula $\theta_j(t) = (\omega_j t) \bmod 2\pi$ is derived from the following lemma:

Lemma *Any continuous homomorphism* $\mathbb{R} \to \mathbb{R} \bmod 2\pi$ *can be represented as* $t \; a\omega t \bmod 2\pi$ *for some "angular velocity"* $\omega \in \mathbb{R}$.

Beginning of a proof of the lemma. The kernel of the homomorphism Θ: $t \mapsto \theta_j(t)$ is a closed subgroup. If $\ker \Theta = \mathbb{R}$, then the statement that is being proved holds (with $\omega_j = 0$). If $0 \neq \theta \in \operatorname{im} \Theta$, then $\operatorname{im} \Theta \supseteq [0,\theta]$ (why?); hence, $\operatorname{im} \Theta$ contains $2\pi/n$ for large n, and so $\Theta(t) = 0$ for some $t \neq 0$. (Why?) Therefore, $\ker \Theta = T\mathbb{Z}$ for some $T > 0$. Complete this proof showing that $\theta_j(t) = (\omega_j t) \bmod 2\pi$, with the angular velocity $\omega_j = \pm 2\pi/T$.

H10.13. D

The necessity of conditions (a) and (b) can be established as follows. Matrix entries of the elements of the group $g(\mathbb{R}) = \mathbb{R} \bmod T$ with respect to any fixed basis are continuous functions of $t \in [0,T]$. Hence, they are uniformly bounded, so the **orbit** $\{g^t v\colon t \in \mathbb{R}\}$ is bounded for any $v \in L$. If some of g^t had an eigenvalue from $\mathbb{S}^1$, then the iterations such as $(g^t)^n = g^{nt}$, for n tending to $+\infty$ or $-\infty$, would extend the corresponding eigenvector to infinity, making the orbit unbounded. Next, if the absolute value of all eigenvalues was 1 but some of g^t had a Jordan box of dimension greater than one, the same thing would happen: $|g^{nt}v| \to \infty$ as $n \to \infty$ and some $v \in L$, as discussed in section E7.2 (the "2×2 Matrices That Are Roots of Unity" problem group discussed previously). **QED**. (We leave it to the reader to fill in all the details.)

In this proof, instead of using the boundedness of the orbits, one may refer to a weaker claim:

$$g(\mathbb{R}) = \mathbb{R} \bmod T \ \ (T \neq 0) \Rightarrow \forall t \in \mathbb{R},\ \exists n_1, n_2, \ldots \ \text{ such that } g^{n,t} \to E \text{ for } i \to \infty.$$

Now verify that claim. Do it separately for cases where t is commensurable and incommensurable with T.[16] In the first case we have $g^{n_i t} = E$ for an arithmetic progression of n_i. (Which one?) We suggest treating the second case by using either of the following two methods:

(1) Using Euclid's algorithm. At least as far back as the ancient Greek period [and probably much earlier, to the ancient East period (Arnol'd 2002)], mathematicians knew a remarkable fact that, given reals $\xi,\eta \neq 0$, the linear combinations $m\xi + n\eta$ $(m,n \in \mathbb{Z})$ with integral coefficients are either multiples of some $\zeta \neq 0$ [a greatest common divisor (GCD) of ξ, η] or everywhere dense in $\mathbb{R}$. They applied **Euclid's algorithm** based on the **Euclidean** (**division with remainder**) property

$$\forall \xi, \eta: \quad \xi \neq 0 \Rightarrow \exists k \in \mathbb{Z},\ \exists \zeta \in [0, |\xi|):\ \eta = k\xi + \zeta.$$

The remainder ζ is a linear combination of ξ, η with integral coefficients and $\zeta < |\xi|$. ζ is found in the first step of a procedure by Euclid's algorithm.[17]

[16] ξ and η are referred to as **commensurable** if they have a common divisor, otherwise, as **incommensurable**.

[17] In reality, ancient scientists kept to geometric images, and dealt with line segments instead of real numbers. Interested readers will find extensive discussions of this and related topics in Van der Waerden (1950).

In the next step, ξ, η are substituted by, respectively, ζ, ξ, and so on. Following the Greeks, prove that the computational process is finite under and only under the condition of commensurability of ξ, η, and then the last nonzero remainder is their GCD. Therefore, for incommensurable ξ, η (as a hypotenuse and a leg of isosceles or other non-Pythagorean right triangles) we have a **potentially infinite** computation procedure, with $\zeta_1 > \zeta_2 > \ldots > 0$ and with all ζ_n being linear combinations of ξ, η with integral coefficients![18]

(2) Using a technique based on the "Dirichlet boxe principle" (also known as the pigeonhole principle). This principle, in its simplest form, states that having put n balls into $k < n$ boxes, there is a box containing at least two balls. Here we suggest applying this principle in a form of the famous theorem:

Poincaré recurrence theorem. *Let H be an infinite transformation group of a variety K such that every element $h \in H$ preserves* the *volume*:

$$S \subseteq K \text{ is measurable} \Rightarrow \forall h \in H,\ h(S) \text{ is measurable and } \mathrm{vol}(h(S)) = \mathrm{vol}(S).$$

If $\mathrm{vol}(K) < \infty$*, then a positive-volume subset cannot have intersections of zero volume (say, empty) with any elements of its orbit:* $\mathrm{vol}(S)>0 \ \Rightarrow \ \exists h \in H :\ \mathrm{vol}(h(S) \cap S)>0$.

This theorem is applied below to prove that linear combinations $m\xi + n\eta$ with integral coefficients are everywhere dense in $\mathbb{R}$ for incommensurable ξ and η. A subgroup $H = (\eta\mathbb{Z}) \bmod \xi$ of rotations of a circle $\mathbb{S}^1 \cong \mathbb{R} \bmod \xi$ satisfies the conditions of this theorem. (Why?) Therefore, for an interval $I = (\theta_0 - \varepsilon, \theta_0 + \varepsilon) \subset \mathbb{S}^1$ and some $h \in H$ we have $|h\theta_0 - \theta_0| < 2\varepsilon$ (why?), which simply means that $|m\xi + n\eta| < 2\varepsilon$ for some integers m, n! (Why?)

Prove the Poincaré theorem.

Dirichlet introduced his box principle having proved in 1842 the famous theorem showing everywhere density under conditions described previously and allowing an effective density control as well:

For $Q > 1$ and $\mu \in \mathbb{R}$ there exist integers p, q such that $0 < q \le Q$ and $|p - \mu q| < Q^{-1}$, and also integers p, q such that $0 < q < Q$ and $|p - \mu q| \le Q^{-1}$.

Dirichlet's theorem and its multidimensional versions are widely applied to the geometry of numbers, Diophantine approximations, and ergodicity theories. One of these versions is formulated as follows:

For $Q > 0$ and linear operator L: $\mathbb{R}^m \to \mathbb{R}^n$ there exists a solution $(0,0) \ne (x,y) \in \mathbb{Z}^m \times \mathbb{Z}^n$ for the system of inequalities $\|x\| \le Q$, $\|Lx - y\| \le Q^{-m/n}$; here $\|\cdot\|$ denotes the maximum of absolute values of coordinates with respect to the usual coordinate bases.

This version follows straightforwardly (how?) from the famous **Minkowski theorem,** which is also based on the Dirichlet box principle:

[18] In fact, $\zeta_n \to 0$ as $n \to \infty$, but it is not necessary to use that because the incommensurability of ξ, η is already implied by the fact that the diminishing sequence of positive ζ_n is not a finite one.

Theorem 1. *A centrally (with respect to the origin of coordinates) symmetric compact (closed and bounded) convex body in* $\mathbb{R}^k$ *of volume equal to or greater than* 2^k *contains integer points besides the origin.*[19]

(The readers may depict several planar bodies satisfying these conditions and find those points.)

The foregoing multidimensional version of Dirichlet's theorem can be strengthened as follows:

For $Q > 0$ *and linear operator* L: $\mathbb{R}^m \to \mathbb{R}^n$ *there exists a solution* $(0,0) \neq (x,y) \in \mathbb{Z}^m \times \mathbb{Z}^n$ *for the system of inequalities* $\|x\| \leq Q$, $\|Lx - y\| < Q^{-m/n}$, *and a solution* $(0,0) \neq (x,y) \in \mathbb{Z}^m \times \mathbb{Z}^n$ *for the system of inequalities* $\|x\| < Q$, $\|Lx - y\| \leq Q^{-m/n}$; *here* $\|\cdot\|$ *denotes the maximum of absolute values of coordinates with respect to the usual coordinate bases.*

This claim is a special case $\left(\text{using } k = m + n \text{ and } M = \begin{pmatrix} E_m & 0 \\ -L & E_n \end{pmatrix}\right)$ of the following general theorem.

Theorem 2. *For* $c_1,\ldots, c_k > 0$ *and* $M \in GL(k,\mathbb{R})$ *there exists a solution* $(0,\ldots,0) \neq (x_1,\ldots, x_k) \in \mathbb{Z}^k$ *for the system of inequalities* $|(Mx)_1| \leq c_1$, $|(Mx)_i| < c_i$, $i = 2,\ldots, k$, *whenever* $c_1 \ldots c_k \geq |\det M|$.

This theorem follows (how?) from Theorem 3.

Theorem 3. *A solid closed or half-closed (containing at least one of the faces) parallelepiped in* $\mathbb{R}^k$ *with its center at the origin of coordinates contains nonzero integral points if its volume is greater than or equal to* 2^k.

(Evidently, a similar claim is false for open parallelepipeds.) In turn, Theorems 1 and 3 follow (how?) from Theorem 4.

Theorem 4. *A convex body in* $\mathbb{R}^k$ *of volume greater than* 2^k *that is centrally symmetric with respect to the origin of coordinates contains at least two integer points besides the origin of the coordinates.*[20]

More advanced readers familiar with multidimensional **integer lattices** and able to work with tools of the **geometry of numbers** are encouraged try to prove Theorem 4 on their own or learn a proof from sources such as Milnor and Husemoller (1973), Cassels (1957, 1959, 1978), or Schmidt (1980). (There are several nice proofs of Theorem 4 in these monographs. We will discuss those proofs and related materials in the "Fibonacci Numbers, Continued Fractions, and Pell Equation" problem group in volume 2 of this book.)

Equivalence of the two definitions of commensurability modulo a number. If $\frac{a_1+n_1 b}{k_1} = \frac{a_2+n_2 b}{k_2}$, then $k_2 a_1 - k_1 a_2$ is a multiple of b, so we may consider k_1, k_2 as relatively prime. Invert this implication using the previously discussed properties of GCD.

[19] Actually, the boundedness follows from the finiteness of the volume and the convexity (Cassels 1957).

[20] Theorems 2–4 also belong to Hermann Minkowski.

Necessity of condition (c). Let $g(\mathbb{R}) = \mathbb{R} \bmod T$. As discussed previously in section P10.13. C***, for the complex arguments of the eigenvalues of $g(t)$ we have $\omega_j t = \theta_j + 2\pi k_j$ [$k_j \in \mathbb{Z}$, $\theta_j = \theta_j(t)$]. For $t = T$ we obtain $\omega_j T = 2\pi m_j$, so $T = \frac{2\pi m_j}{\omega_j}$ $= \frac{2\pi m_j t}{\theta_j + 2\pi k_j}$ will be the same for all j, in other words, $\frac{\theta_1 + 2\pi k_1}{m_1} = \frac{\theta_2 + 2\pi k_2}{m_2} = \ldots$, which is equivalent to (c).

Sufficiency of condition (a)−(c). Working with a proper basis and taking into account conditions (a) and (b) reduces the problem to finding $k_j, m_j \in \mathbb{Z}$ for all j, so that $\frac{\theta_1 + 2\pi k_1}{m_1} = \frac{\theta_2 + 2\pi k_2}{m_2} = \ldots$. (Why?) By condition (c), for each pair (j,l) there exist k_{jl}, k_{lj}, m_{jl}, and m_{lj} such that the equalities $\frac{\theta_j + 2\pi k_{jl}}{m_{jl}} = \frac{\theta_l + 2\pi k_{lj}}{m_{lj}}$ are satisfied. There are two possibilities:

(1) If some θ_j are commensurable with 2π, then all of them are and k_j, m_j are found easily.
(2) θ_j are incommensurable with 2π. Readers may obtain k_j, m_j in steps **(i)–(v)** as follows.

(i) Show that for any given pair of indices j, l and all possible congruences $m_{jl}\theta_l \equiv m_{lj}\theta_j \bmod(2\pi)$ all vectors of integral coefficients (m_{jl},m_{lj}) are proportional to each other.
(ii) Given a series of equalities $\frac{\theta_1 + 2\pi k_{12}}{m_{12}} = \frac{\theta_2 + 2\pi k_{21}}{m_{21}}$, $\frac{\theta_1 + 2\pi k_{13}}{m_{13}} = \frac{\theta_3 + 2\pi k_{31}}{m_{31}}$, ... with relatively prime pairs (m_{12},m_{21}), (m_{13},m_{31}), ..., show that $k_{1j} - k_{1l}$ are divisible by $GCD(m_{1j},m_{1l})$ for any $j,l > 0$.
(iii) Show that if the equality $\frac{\theta_1 + 2\pi k_{1j}}{m_{1j}} = \frac{\theta_j + 2\pi k_{j1}}{m_{j1}}$ is true for some relatively prime m_{1j}, m_{j1} and for some integer k_{1j}, k_{j1}, then it will be true for all k_{1j}, k_{j1} that belong simultaneously to certain arithmetic progressions with the differences m_{1j}, m_{j1}, respectively.
(iv) Show that one can find a series of equalities in **(ii)** such that $k_{12} = k_{13} = \ldots$. Do this using the following well-known theorem from number theory.

Theorem. *For natural numbers* $m_1,\ldots, m_n \neq 0$ *and* $r_1,\ldots, r_n$, *a system of congruencies* $x \equiv r_j \bmod m_j$: $j = 1,\ldots, n$ *is solvable if and only if every* $r_j - r_l$ *is divisible by* $GCD(m_j,m_l)$ $(j,l = 1,\ldots, n)$.[21]

Prove also this nice theorem.

(v) Final step: take a series of equalities in **(ii)** with $k_{12} = k_{13} = \ldots$ and define $k_1 := k_{1j}$, then k_j for $j > 1$ from equations $\frac{\theta_1 + 2\pi k_{1j}}{m_{1j}} = \frac{\theta_j + 2\pi k_{j1}}{m_{j1}}$, then m_1 as the least common multiple (LCM) of m_{12}, m_{13}, ..., and, lastly, $m_j := m_{j1}m_1/m_{1j}$.

[21] A special case of this theorem, the **Chinese remainder theorem,** states that for any set of integer numbers $m_1,\ldots, m_n \neq 0$, where any pair is relatively prime, the indicated system of congruences is solvable for any $r_1,\ldots,r_n$.

H10.13. E

The equalities $\frac{\theta_1+2\pi k_1}{m_1} = \frac{\theta_2+2\pi k_2}{m_2} = \ldots$ are equivalent to $\frac{\omega_j}{\omega_m} = \frac{\theta_j+2\pi k_j}{\theta_m+2\pi k_m} \in \mathbb{Q}$, $\forall j,m$. Complete the proof by considering the eigenvalues equal to 1.

H10.13. F

For $g(\mathbb{R}) = \mathbb{R} \bmod 2\pi$ with $\ker g = T\mathbb{Z}$ and $g(t_0) = B$ the **least positive period** can be calculated as $T = \left|\frac{2\pi m_j t_0}{\theta_j+2\pi k_j}\right|$. The number of "degrees of freedom" in this formula corresponds to the number of the possible natural numbers k_j, which are found from the equations $\frac{\theta_1+2\pi k_1}{m_1} = \frac{\theta_2+2\pi k_2}{m_2} = \ldots$. All pairs of m_j are relatively prime, and so all k_j may simultaneously take values from certain arithmetic progressions with the differences m_j, as discussed in **(iii)**. (Specifically, for $\theta_1 = \theta_2 = \ldots$ we have $k_1 = k_2 = \ldots$, which may be any integer.) Thus, choosing any set of $\{k_1,k_2,\ldots\}$ that do not satisfy this condition yields $g(\mathbb{R}) \cong \langle\mathbb{R},+\rangle$. (Work out the details.)

H10.14

The eigenspaces corresponding to simple (nonmultiple) eigenvalues are one-dimensional, so the elements of a one-parameter linear group, in which this operator is included, are diagonalized in the same basis. Thus, the group defines the one-parameter scalar linear groups in the eigenspaces, which implies its uniqueness. (Work out the details using section P10.2*.) By contrast, the matrix $\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$ $= |\lambda| \begin{pmatrix} \cos(\arg\lambda) & -\sin(\arg\lambda) \\ \sin(\arg\lambda) & \cos(\arg\lambda) \end{pmatrix}$ is included in infinitely many one-parameter linear groups $\rho(t) \begin{pmatrix} \cos\theta(t) & -\sin\theta(t) \\ \sin\theta(t) & \cos\theta(t) \end{pmatrix}$ defined by $\rho(t) = e^{kt}$, $\theta(t) = (\omega t) \bmod 2\pi$, of $k = \frac{\ln|\lambda|}{t_0}$ and $\omega = \frac{\arg\lambda+2\pi m}{t_0}$, with arbitrary $m \in \mathbb{Z}$.

H10.15

We suggest proceeding by the same method as in section H10.3 by taking into account the linearity and continuity of matrix multiplication.[22]

H10.16

This is done in a manner similar to that in section H10.4 by substituting k with A.

H10.17

For a vector series, if all numerical series – of the first, second, . . ., last coordinates of its terms-with respect to some basis converge, then similar series of coordinates with respect to any basis also converge because coordinates with respect to two bases are bound by the same system of linear equations for any vector. (Work out the details.)

Claims (1)–(3) from section P10.5* imply that the fact that the matrix exponential series satisfies the Cauchy problem $\dot{g}(t) = Ag(t)$; $g(0) = E$ can be proved in the same way as for the numerical exponential series (see section H10.5).

We suggest proving claims (1)–(3) from section P10.5* for a matrix exponential series using **normal convergence**. This concept plays the same role for vectorial (particularly matrix) series as the absolute convergence, discussed previously in section H10.5, for numerical series. Introduce a **norm** of a matrix $A = (a_{ij})_{i=\ldots,\, j=\ldots}$ as $||A||: = \sum|a_{ij}|$. Verify that **(α)**: $||kA|| = |k|\cdot||A||$ (k is scalar), **(β)**: $||A + B|| \leq ||A|| + ||B||$, **(γ)**: $||A \,\tilde{}\, B|| \leq ||A||\cdot||B||$.

The functions on vector spaces satisfying both conditions **(α)** and **(β)** are referred to as **seminorms** or **convex functionals** [as discussed previously in the "Convexity and Related Classical Inequalities" problem group (section P9.29)]. Matrix seminorms satisfying **(γ)** are called multiplicative. A seminorm is called a **norm** if in addition to being a multiplicative seminorm it has the property **(δ)**: $||A|| = 0 \Rightarrow A = 0$. Obviously, $||A|| = \sum|a_{ij}|$ is a norm. In general, a matrix norm is not multiplicative, for example, $||A||: = \max|a_{ij}|$. A different, and the most applicable, example of a multiplicative norm is $\|A\|: = \sup_{x\neq 0}|Ax|/|x| = \sup_{|x|=1}|Ax|$, discussed previously in the "2 × 2 Matrices That Are Roots of Unity" problem group (section P7.6).

A vectorial (or matrix) series is referred to as **normally convergent**, *with respect to a fixed norm*, when a nonnegative series of the term-by-term norms of the source series converges. Show that the normal convergence of the exponential matrix

[22] Recall that linear operations over finite-dimensional matrices are continuous.

series with respect to norm $||A||:=\sum|a_{ij}|$ implies convergence.[23] Generalize, for the normal convergence, the claims in **A** and **B** of section H10.5 (originally formulated for absolute convergence). In that case, we will have claims (1)–(3) from section P10.5* holding for matrix exponential series.

Concerning the sums of exponential series for a diagonal matrix and for a nilpotent Jordan box, readers may compare the results of their own computations with the commonly known formulas: $e^{tE}=e^tE$,

$$e^{tN}=\begin{pmatrix} 1 & t & \cdots & t^{n-1}/(n-1)! \\ & \ddots & \ddots & \vdots \\ & & \ddots & t \\ & & & 1 \end{pmatrix} \quad (\dim N=n).$$

H10.18

Establish the binomial formula $(A+B)^n=\sum_{k=0}^{n}\binom{n}{k}A^kB^{n-k}$ $(n=0,1,\ldots)$ for commuting matrices A, B. Using this formula, show for partial sums of the exponential series $S_n=\sum_{k=0}^{n}(k!)^{-1}A^k$, $T_n=\sum_{k=0}^{n}(k!)^{-1}B^k$ that $S_nT_n-\sum_{k=0}^{n}(k!)^{-1}(A+B)^k \to 0$ as $n\to\infty$.

Taking into account the permutability of λE and N and using $e^{\lambda E}$, e^N found in section H10.17 we obtain $e^{\lambda E+N}=e^{\lambda E}\circ e^N=\begin{pmatrix} e^\lambda & e^\lambda & \cdots & e^\lambda/(m-1)! \\ & \ddots & \ddots & \vdots \\ & & \ddots & e^\lambda \\ & & & e^\lambda \end{pmatrix}$ $(\dim E=\dim N=m)$.

Next, we suggest proving that the exponential map realizes a diffeomorphism between neighborhoods of zero and the identity with the help of the implicit function theorem. Therefore, verify that this mapping satisfies the conditions of that theorem: equal dimensions of the preimage and the image, continuous differentiability of the map, and nondegeneracy of its differential at zero.

Draw the exponential representation of one-parametric linear groups using the local, near zero, diffeomorphness of the exponential map Aae^A and proceeding

[23] In fact, a series in a finite-dimensional space normally converges with respect to any norms if it does so with respect to one [because of the equivalence of the norms in a finite-dimensional space, as discussed in the "2 × 2 Matrices That Are Roots of Unity" problem group above (section H7.6)]. In turn, using the norm of maximum of absolute values of the coordinates with respect to a fixed basis shows that the normal convergence of any series in the finite-dimensional spaces implies convergence.

similarly to method 1 in section H10.6, taking into account the commutativity of sA and tA for any matrix A and any $s,t \in \mathbb{R}$.

Next, performing multiplication of series prove for a multiplicative seminorm that

$$e^A e^B e^{-A} e^{-B} = E + [A,B] + o\left(\|A\|^2 + \|B\|^2\right) = e^{[A,B]} + o\left(\|A\|^2 + \|B\|^2\right) \quad \text{as}$$
$$\|A\|, \|B\| \to 0,$$

which provides the desired formula.[24] (Work out the details.)

Thus, $\frac{e^{sA}e^{tB}e^{-sA}e^{-tB}-E}{s^2+t^2} = \frac{st}{s^2+t^2}\,[A,B] + o(1)$ as $s,t \to 0$. The first term on the right-hand side cannot be $o(1)$ unless $[A,B] = 0$. (Why? Answer assuming $s = t$.) **QED**.

Lastly, the commutativity of A^m and B^n for some integers m, n, or the commutation of e^A and e^B does not imply the commutativity of A and B themselves. Give counterexamples using 2×2 matrices.

H10.19

We suggest proceeding by one of the following two methods:

Method 1 (Arnol'd 1975). The series $e^A - (E + A/n)^n = \sum\limits_{k=1}^{\infty} \left[1 - \frac{n(n-1)\ldots(n-k+1)}{n\cdot n\cdot\ldots\cdot n}\right] \cdot \frac{A^k}{k!}$ is majorized by numerical series $e^{\|A\|} - (E + \|A\|/n)^n = \sum\limits_{k=1}^{\infty} \left[1 - \frac{n(n-1)\ldots(n-k+1)}{n\cdot n\cdot\ldots\cdot n}\right] \cdot \frac{\|A\|^k}{k!}$ (by the lemma in section E10.17). The later series converges (since it is a difference between a convergent series and a polynomial) and has its sum tending to zero as $n \to \infty$ (in accordance with section P10.7*). **QED**.

Method 2. Establish the following lemma.

Lemma. *An exponential function commutes with a similitude transformation:* $C^{-1} \circ e^A \circ C = e^{C^{-1} \circ A \circ C}$ ($\det C \neq 0$).

Obviously, $(E + A/n)^n$ also commutes with similitude. Therefore, $\lim\limits_{n\to\infty} (E + A/n)^n$ may be calculated for Jordan matrices only; for this, the Jordan boxes may be considered separately (by virtue of the invariance of root subspaces). Lastly, find these limits for those boxes, taking into account the statement in section P10.7* and using e^A for the same boxes found in section H10.18.

[24] Here, the desired formula follows from substituting A, B by sA, tB, respectively.

H10.20

If F is closed with respect to differentiation, then the finite Taylor sums are linear operators on F. Next, prove, using decomposition with respect to any fixed basis, the following lemma.

Lemma. *A sequence of linear operators on a finite-dimensional space L, convergent on any vector $v \in L$, converges in a space of the linear operators on L (and thus has a linear operator as its limit).*[25]

Therefore, the sum of the Taylor series is a linear operator on F. Now, the Taylor-expansion formula shows that $e^{s\frac{d}{dt}} = g^s$ and, in particular, that F is closed with respect to the shifts g^s.

A sharpened inversion of this statement – closure of the space with respect to the argument's shifts $\Rightarrow$ differentiability of the elements, and closure of the space with respect to the differentiation – can be proved for any finite-dimensional F consisting of continuous functions (as in the proposition from section P10.21*** below). Let F be closed with respect to the shifts of arguments g^s. Then these shifts form a one-parameter linear group in F. (Why? Answer using the continuity of the elements of F and the preceding lemma.) In finite-dimensional spaces those groups are differentiable (section P10.18**), so $\forall s \in \mathbb{R}$, $\exists\, \frac{dg^s}{ds} f(t) \in F$. But obviously,

$$\left(\frac{dg^s}{ds}|_{s=0} f\right)(t) = \frac{d}{ds}|_{s=0} f(t+s) = \lim_{s\to 0} \frac{f(t+s)-f(t)}{s} = \dot{f}(t),$$

which completes the proof.

H10.21

Proof of the proposition. *(i)* $\Rightarrow$ *(ii)* was accomplished previously in section H10.20.

(ii) $\Rightarrow$ *(i)* Show that the elements of F are analytic on $\mathbb{R}$, that is, they are equal to the sum of their Taylor series in a neighborhood of any point, and then complete the proof with the statement in section P10.20**.

(iii) $\Rightarrow$ *(i)* Closure of this space with respect to the shifts of the argument may be verified by direct computations. In addition, as advanced readers know, this space is a space of the solutions of an appropriate linear **autonomous** (i.e., with coefficients not dependent on time) **ODE** and, as these readers also know, closure with respect

[25] Proving a similar claim for infinite-dimensional space requires stipulating a kind of topological **completeness** of the space (the Banach-Steinhaus equicontinuity theorem). Interested readers may consult Dunford and Schwartz (1957), Hille and Phillips (1957), Yosida (1965), Riesz and Sz.-Nagy (1972), Reed and Simon (1972), Rudin (1973), and references therein.

to the shifts of the argument is equivalent to the autonomy of the equation! Interested readers will find further discussions in Arnol'd (1975, 1978, 1989).

(ii) ⇒ *(iii)* Obviously, if F_1, F_2 are finite-dimensional spaces of quasipolynomials (corresponding to the sets of $\lambda_{1j} + i\omega_{1j}$, N_{1j}, and $\lambda_{2j} + i\omega_{2j}$, N_{2j}, respectively), then the space $F_1 + F_2$ is also a space of quasipolynomials [corresponding to the union of the sets of $\lambda_{1j} + i\omega_{1j}$, $\lambda_{2j} + i\omega_{2j}$ and $N_j = \max(N_{1j}, N_{2j})$]. Now, let F satisfy (*ii*) and F_1 be its maximal quasipolynomial subspace. Assume that $F_1 \neq F$ and let $f \in F \setminus F_1$. f satisfies a linear autonomous ODE, so it is a quasipolynomial: $f = \sum_{j \in J_f} \left(p_{fj}(t)\cos\omega_{fj}t + q_{fj}(t)\sin\omega_{fj}t\right) e^{\lambda_{fj}t}$ (see section E10.21).

Use the linear independence of the functions $t^{k_j}e^{\lambda_j}\cos\omega_j t$, $t^{k_j}e^{\lambda_j}\sin\omega_j t$, corresponding to a family of distinct triplets of parameters (k_j,λ_j,ω_j) (why are they linearly independent?) to derive from (*ii*) that F must contain all these functions corresponding to $\lambda = \lambda_{fj}$, $\omega = \omega_{fj}$, and $k_j \in \{0, \ldots, \max(\deg p_{fj}, \deg q_{fj})\}$ $(j \in J_f)$. Let F_2 be a subspace spanned on all these functions. Therefore, $F_1 + F_2$ is a subspace of the quasipolynomials containing F_1 but distinct from it, by contradiction with the maximality of F_1.

An analogous proposition for functions on unit circle is formulated as follows:

Proposition 2. *The following conditions on a finite-dimensional space F of functions $\mathbb{S}^1 \to \mathbb{R}$ are equivalent:*

(i) *F consists of continuous functions and is closed to the shifts of the argument modulo 2π (in other words, rotations).*
(ii) *F consists of differentiable functions and is closed to the differentiation.*
(iii) *F is a space of the trigonometric polynomials $f = \sum_{k \in K_F} a_k \cos kt + b_k \sin kt$ corresponding to a finite set of natural numbers K_F.*[26]

The implications (*i*) ⇒ (*ii*), (*ii*) ⇒ (*i*), and (*iii*) ⇒ (*i*) can be proved using the same approach as in the proof of the foregoing proposition. Alternatively, they can be derived from these propositions (by considering functions on a circle as periodic functions on a straight line; complete the details).

The implication (*ii*) ⇒ (*iii*) is proved as follows. By the foregoing proposition, (*ii*) implies that F (as a space of periodic functions on $\mathbb{R}$) is the space of the quasipolynomials corresponding to a set of $\lambda_j + i\omega_j$ and N_j. (Provide the details.) Show that $\lambda_j = 0$, $N_j = 0$, and ω_j are integers for all j.

[26] Zero is considered a natural number.

H10.22

The determinant $\det(E + \varepsilon A)$ is a polynomial in ε, and a direct calculation yields the desired formula.[27] For an arbitrary matrix-valued function $W(\varepsilon)$ with $W(0) = E$ for which $\frac{dW}{d\varepsilon}(0)$ exists, $\det W(\varepsilon)$ is generally not a polynomial, but a similar direct calculation with the Taylor-series expansion $W(\varepsilon) = E + \varepsilon \cdot \frac{dW}{d\varepsilon}(0) + o(\varepsilon)$ yields the desired expansion for $\det W(\varepsilon)$. Lastly, for a differentiable function $W(t)$ with values being invertible matrices, we have

$$\begin{aligned}\left.\frac{d(\det W)}{dt}\right|_{t=t_0} &= \left.\frac{d\left(\det\left(W(t_0+\varepsilon)\circ W(t_0)^{-1}\right)\right)}{d\varepsilon}\right|_{\varepsilon=0} \cdot \det W(t_0)\\ &= tr\left(\dot{W}\big|_{t=t_0} \circ W(t_0)^{-1}\right) \cdot \det W(t_0),\end{aligned}$$

QED. (We leave it to the reader to fill in the details.)

H10.23

The correspondence $\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \leftrightarrow a + bi$ defines isomorphisms of both algebraic and topological structures of these fields (in common terms, this is an isomorphism of the topological fields). Thus, we will have $e^{\begin{pmatrix} a & -b \\ b & a \end{pmatrix}} \leftrightarrow e^{a+bi}$ as those series converge and $\sum_{j=0}^{n} (j!)^{-1} \begin{pmatrix} a & -b \\ b & a \end{pmatrix}^j \leftrightarrow \sum_{j=0}^{n} (j!)^{-1}(a+bi)^j, \forall n.$

H10.24

Taking complex vectors $\overrightarrow{OA} = 1 + 0i = 1, \overrightarrow{AB} = z/n, \overrightarrow{OB} = \overrightarrow{OA} + \overrightarrow{AB}, \overrightarrow{AC} = \mathrm{Re}\, z/n + 0i = \mathrm{Re} z/n$, and $\overrightarrow{CB} = i\ \mathrm{Im}\ z/n$ (create a figure) we will find from the triangle OBC:

[27] The coefficients in the remaining ε^k are also traces of certain linear operators on certain vector spaces (connected to the source operator A); can the reader specify them? (Advanced readers familiar with skew symmetry, as discussed, for example, in the problem group "A Property of Orthogonal Matrices" above, will be able to do this.)

$$\mathrm{Im}z/n + o(1/n) = \frac{\mathrm{Im}z/n}{1+\mathrm{Re}z/n} = tg\arg(1+z/n) = \arg(1+z/n) + o(1/n),$$

$$\begin{aligned}|1+z/n| - (1+\mathrm{Re}z/n) &= \sqrt{(1+\mathrm{Re}z/n)^2 + (\mathrm{Im}z/n)^2} - (1+\mathrm{Re}z/n)\\ &= \frac{(\mathrm{Im}z/n)^2}{\sqrt{(1+\mathrm{Re}z/n)^2 + (\mathrm{Im}z/n)^2} + (1+\mathrm{Re}z/n)}\\ &= O\left(1/n^2\right)\end{aligned}$$

(work out the details), which proves the desired asymptotic relations. From the first of them, $\mathrm{Im}z = n\cdot\arg(1+z/n) + o(1)$. From the second one, for any $\varepsilon > 0$ there exists an arbitrarily small $\delta > 0$ such that for large n $[n \geq n_0(\delta)]$ we have

$$e^{\mathrm{Re}z-\delta} - \varepsilon \leq \left(1 + \tfrac{\mathrm{Re}z-\delta}{n}\right)^n \leq |1+z/n|^n \leq \left(1+\tfrac{\mathrm{Re}z+\delta}{n}\right)^n \leq e^{\mathrm{Re}z+\delta} + \varepsilon$$

(fill in the details), which proves the required limit relationships. (Why?)

Using these relationships and the isomorphism $z = a + bi \mapsto \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$, the linear operator $\left[E + n^{-1}\begin{pmatrix} a & -b \\ b & a \end{pmatrix}\right]^n$ rotates the planar vectors by the angle $n\cdot\arg(1+z/n) = \mathrm{Im}z + o(1)$ and simultaneously expands them $|1+z/n|^n = e^{\mathrm{Re}z} + o(1)$ times. Applying section P10.19** and the inverse isomorphism $\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \mapsto a + bi$ completes the proof of Euler's formula. (Work out the details.)

A proof of Euler's formula based on section P10.13 (**C**) looks as follows. According to the isomorphism $a + bi \leftrightarrow \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ and the statement in section P10.13 (**C**) we will have, for $t \in \mathbb{R}$ and a fixed $z = a + bi$, $e^{z\cdot t} = e^{k_z t}(\cos\omega_z t + i\sin\omega_z t)$, with the appropriate $k_z, \omega_z \in \mathbb{R}$. (Work out the details.) Now, differentiating this equation with respect to t at $t = 0$, using the differentiability of the exponential functions, and taking into account the differential equation from section P10.15**, we find $z = k_z + i\omega_z$. **QED**.

A proof of Euler's formula using the exponential series and the Taylor-series expansions of cosine and sine looks as follows. Notice that exp is a homomorphism of groups $\langle\mathbb{C},+\rangle \to \langle\mathbb{C}^*,\bullet\rangle$. (Why? Use the statements in sections P10.18** and P10.23** for the answer.) Therefore, $e^z = e^{\mathrm{Re}z + i\mathrm{Im}z} = e^{\mathrm{Re}z}e^{i\mathrm{Im}z}$. Now, computing series $e^{it} = \sum_{n=0}^{\infty}(n!)^{-1}(it)^n$ taking into account the formulas for integer powers of i, $i^{4n\pm1} = \pm i$, $i^{4n+2} = -1$, $i^{4n} = 1$, and using commutative convergence of absolutely convergent series yields

$$e^{it} = \sum_{n=0}^{\infty}(-1)^n\left[\frac{t^{2n}}{(2n)!} + \frac{it^{2n+1}}{(2n+1)!}\right] = \cos t + i\sin t,$$

which completes the proof. (Work out the details; Taylor-series expansions for cos t and sin t at $t = 0$ can be found by direct differentiation.)

To describe a set *Log u*, readers may use the fact that exp is a homomorphism of groups $\langle \mathbb{C},+\rangle \to \langle \mathbb{C}^*,\bullet\rangle$. Find exp's kernel using Euler's formula as follows: $1 = e^{\mathrm{Re}z}(\cos \mathrm{Im}\ z + i \sin \mathrm{Im}\ z) \Rightarrow \sin \mathrm{Im}\ z = 0$, so Im z is either an integer or half-integer multiple of 2π, and lastly, from the equality $1 = e^{\mathrm{Re}z} \cos \mathrm{Im}z$ it follows that ker exp $= \{2\pi ni\colon n \in \mathbb{Z}\}$. (Work out the details.) Therefore, $Log\ u = \{z + 2\pi ni\colon n \in \mathbb{Z}\}$ for any fixed z such that $e^z = u$ (why?); thus, when you find one of these z, you will describe the whole set *Log u*. Applying Euler's formula, $|u|(\cos \arg u + i \sin \arg u) = e^{\mathrm{Re}z}(\cos \mathrm{Im}z + i \sin \mathrm{Im}z)$ and $z = \ln|u| + i \arg u$. Finally, $Log\, u = \{\ln|u| + i \arg u + 2\pi ni :\ \ n \in \not{C}\}$.

Also, for the description of *Log u*, readers may do without the preliminary finding of ker exp using

Euler's formula just once:

$$|u|(\cos \arg u + i \sin \arg u) = e^{\mathrm{Re}z}(\cos \mathrm{Im}z + i \sin \mathrm{Im}z)$$
$$\Leftrightarrow \quad \mathrm{Re}z = \ln|u| \quad \text{and} \quad \mathrm{Im}z \equiv (\arg u) \bmod 2\pi.$$

(Work out the details.)

H10.25

One may define trigonometric functions via the exponential one, using Euler's formula, as follows. Operator $\begin{pmatrix} \mathrm{Re}e^{it} & -\mathrm{Im}e^{it} \\ \mathrm{Im}e^{it} & Re\, e^{it} \end{pmatrix} = e^{\begin{pmatrix} 0 & -t \\ t & 0 \end{pmatrix}} = \lim_{n\to\infty} \begin{pmatrix} 1 & -t/n \\ t/n & 1 \end{pmatrix}^n$ rotates the planar vectors by angle t (as discussed in section H10.24; here, an *angle* is defined via its correspondence to an arc of unit circle; Work out the details!), which shows that Re e^{it}, Im e^{it} satisfy the usual definitions of cos t and sin t as, respectively, the abscissa and ordinate of the unit vector of a polar angle t. **QED**.

Also, cosine and sine might be defined through their Taylor-series expansions defined as the expansions for Re e^{it}, Im e^{it}, respectively; but we must warn readers that proceeding by this method suggests deriving the usual properties of cosine and sine from their Taylor series, which requires more experience, including familiarity with differential equations.

Next, advanced readers may derive Euler's formula or, conversely, define cosine and sine with the help of the Cauchy-Riemann equations proving the following lemma.

Lemma. *The **Cauchy (initial value) problem** $\dot{f} = f$, $f(0) = 1$ for $f = e^z$ is equivalent to a pair of Cauchy problems for the equation $\ddot{x} = -x$ corresponding to the initial data $x(0) = 1$ $\dot{x}(0) = 0$ and $x(0) = 0$, $\dot{x}(0) = 1$, respectively, for* Re e^z *and* Im e^z.

H10.26

Readers may proceed by the following steps.

A. Derive the identity $C(s) - C(t) = -2S\left(\frac{s+t}{2}\right) \cdot S\left(\frac{s-t}{2}\right)$.

B. Establish that $C(t)$ is decreasing [therefore, $S(t)$ is increasing] for $0 \leq t \leq \pi/2$.

The rest of the proof may be completed by different methods. One of the methods requires a modest amount of experience:

C. Prove that for $n = 1,2,\ldots$, $C(\pi/2^n) = \cos(\pi/2^n)$ and $S(\pi/2^n) = \sin(\pi/2^n)$.

D. Establish the continuity of $C(t)$ and $S(t)$.

E. Show that $C(t)$ is an even and $S(t)$ an odd function.

F. Show by induction that $C(t) = \cos t$ and $S(t) = \sin t$ for t being the sums of binary fractions, $t = \pi \sum_{i=0}^{n} k_i/2^i$ $(n = 0,1,\ldots, k_i \in \mathbb{Z})$ and hence, by virtue of continuity, for all t. It completes the proof.

Instead, readers may accomplish this as follows:

E′. Verify the identity

$$(C(s) + iS(s)) \cdot (C(t) + iS(t)) = C(s+t) + iS(s+t),$$

which shows, accounting for the monotonicity of $S(t)$ for $0 \leq t \leq \pi/2$ (see **B**), that the function

$$\theta : \begin{cases} [0, \pi/2] & \to & [0, \pi/2] \\ t & \mapsto & \arcsin S(t) = \arccos C(t) = \arg(C(t) + iS(t)) \end{cases}$$

is additive, $\theta(s + t) = \theta(s) + \theta(t)$ $(s, t, s + t \in [0, \pi/2])$. Therefore, it is linear over the field of rational numbers. Because of continuity, it is linear over the reals, and because of the end conditions $\theta(0) = 0$, $\theta(\pi/2) = \pi/2$, finally we have $\theta(t) = t$. Therefore, $C(t) = \cos t$ and $S(t) = \sin t$ on $[0, \pi/2]$, so, using the $C(t)$ and $S(t)$ definition (and **E**), $C(t) = \cos t$ and $S(t) = \sin t$ on $(-\infty, \infty)$. **QED**. (We leave it to the reader to fill in the details.)

The same reasoning can be presented in a different package:

E″. The identity in **E′** means that mapping $t \mapsto g(t) = \begin{pmatrix} C(t) & -S(t) \\ S(t) & C(t) \end{pmatrix}$ is a one-parameter linear group. [Work out the details accounting for the continuity of $C(t)$, $S(t)$ and the equality $g(0) = E$.] Taking into account that $\det g(t) = C^2 + S^2 = 1$ and using section P10.13 (**C**) yields $C(t) = \cos \omega t$ and $S(t) = \sin \omega t$. Finally, we have $\omega = 1$ since $\sin(\omega\pi/2) = 1$ and $\sin \omega t > 0$ for $0 < t \leq \pi/2$. (Work out the details.)

Also, a continuation after step **B** may be different:

C′. The function $\theta(t)$ from **E′** is defined and increasing for increasing (on $[0, \pi/2]$) $S(t)$. As in **E′**, $\theta(t)$ is linear over the field of rational numbers. In addition, it is bounded. Therefore, it is continuous, because otherwise it could not be bounded, following the claim in section P9.28 (discussed previous in the "Convexity and Related Classical Inequalities" problem group).[28] The remainder of the proof is the same as in **E′**. (Work out the details.)

The reasoning in **E″** shows that any continuous functions $C(t)$, $S(t)$ satisfying the conditions of section P10.26*, except $S(t) > 0$ for $0 < t < \pi/2$ and $S(\pi/2) = 1$, are $C = \cos \omega t$, $S = \sin \omega t$, with arbitrarily fixed ω, and satisfying the condition $S(\pi/2) = 1$ makes $\omega = 1 + 4n$ ($n \in \mathbb{Z}$).

On the other hand, there are discontinuous functions $C(t)$, $S(t)$ that also satisfy the conditions of section P10.26*, except $S(t) > 0$ for $0 < t < \pi/2$ and $S(\pi/2) = 1$; readers may verify that they all have the form $C = \cos \varphi(t)$, $S = \sin \varphi(t)$, with discontinuous homomorphisms φ: $\langle \mathbb{R}, + \rangle \to \langle \mathbb{R}, + \rangle$, as discussed previously in section P10.11*; satisfying condition $S(\pi/2) = 1$ makes φ map subgroup $\mathbb{Z}\pi/2$ onto $\mathbb{Z}(1 + 4n)\pi/2$ ($n \in \mathbb{Z}$).

H10.27

(1) Obviously, A commutes with ${}^tA$ if ${}^tA = -A$. Hence (by virtue of section P10.18**), $e^A \circ e^{{}^tA} = e^{A+{}^tA} = e^0 = E$.

(2) The orthogonal operators $R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ $(0 \le \theta < 2\pi)$ possess skew-symmetric logarithms, which are infinitesimal movements $(\theta + 2\pi n)I = \begin{pmatrix} 0 & -\theta - 2\pi n \\ \theta + 2\pi n & 0 \end{pmatrix}$, $\forall n \in \mathbb{Z}$ (by Euler's formula, using $I = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$). Using the spectral theorem, extend this claim to any orientation-preserving orthogonal linear operators [the spectral theorem was discussed previously in the "Polar and Singular Value Decomposition Theorems" problem group (section E6.8)].

(3) A description of $\log R_\theta$ is as follows:

For $0 < \theta < \pi$ or $\pi < \theta < 2\pi$, Log R_θ: = {all of $\log R_\theta$} is a discrete countable set of skew-symmetric matrices, more precisely, Log $R_\theta = \{(\theta + 2\pi n)I: n \in \mathbb{Z}\}$. In turn, Log E_2 (a 2x2 identity matrix) is a disjoint union of one single point corresponding to a zero matrix and a countable number of varieties, each geometrically a hyperboloid of two sheets. A set of hyperboloids is indexed by integers,

[28] Also, readers may do without a reference to the boundedness, arguing as follows. A monotonic function may have, at most, a countable number of discontinuities, which, if they exist, are finite jumps (why?); at the same time, by the claim in section P9.28, the function $\theta(t)$ would be discontinuous at all points if it were at some point. (Work out the details.)

with the points of the nth hyperboloid corresponding to the matrices similar to $2\pi nI$.[29] *The vertices of the nth hyperboloid correspond to* $\pm 2\pi nI$*; only these vertices correspond to skew-symmetric matrices.*[30] *Explicitly, this hyperboloid consists of the matrices* $\begin{pmatrix} \alpha & \beta \\ \gamma & -\alpha \end{pmatrix}$ *with* $\alpha^2 + \beta\gamma = -4\pi^2 n^2$.[31] *Log* $(-E_2)$ *is the union of similar hyperboloids described by the equation* $\alpha^2 + \beta\gamma = -(2n+1)^2\pi^2$*; the matrices* $\pm(2n+1)\pi I$ *correspond to the vertices of the hyperboloids' sheets.*

Start the proof by finding the eigenvalues of $\log R_\theta$. They are $\pm\, i\cdot(\theta + 2\pi n)$ for some integer n, which can be established using the description of $Log(\cos\theta + i\sin\theta)$ (discussed previously in section H10.24) and the following lemma.

Lemma. *For eigenvalue* λ *of linear operator* A, e^λ *is an eigenvalue of* e^A *of the same multiplicity. Moreover, the Jordan structure (the set of root subspaces in a complexified space) of* e^A *remains the same as that of* A*: if* J *is an* $n \times n$ *Jordan box with an eigenvalue* λ *[i.e.,* $(\lambda E - J)^k \neq 0$ *for* $k < n$ *and* $(\lambda E - J)^n = 0$*], then* e^J *is an* $n \times n$ *Jordan box with an eigenvalue* e^λ*:* $(e^\lambda E - e^J)^k \neq 0$ *for* $k < n$ *and* $(e^\lambda E - e^J)^n = 0$.

(Work out the details.) Therefore, any $\log R_\theta$ has the same Jordan structure as $(\theta + 2\pi n)I$ (which for $\theta \neq 0$ or $n \neq 0$ immediately follows from simplicity of the eigenvalues; why is this statement true for $\theta = 0$ and $n = 0$? Try to answer using this lemma once again). Hence, all $\log R_\theta$ matrices are similar, $\log R_\theta = C^{-1} \circ (\theta + 2\pi n)I \circ C$, and, as a result of the permutability of matrix similitude with the exponential function, $R_\theta = C^{-1} \circ R_\theta \circ C$. For $\theta = 0$ or $\theta = \pi$, C in this equation may be any invertible matrix (as $R_0 = E_2$ and $R_\pi = -E_2$), but for $\theta \neq 0,\pi$, it must have the form $C = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} = kR_\varphi$ $[k = |a + bi|,\ \varphi = \arg(a + bi)]$. (Prove this.) In this case, we have $log\, R_\theta = (\theta + 2\pi n)I$ (why?), which completes the proof.

(4) Unique existence of a logarithm close to 0 and its skew symmetry for an orthogonal operator close to E. An exponential mapping defines a diffeomorphism of a neighborhood of zero in the space of linear operators onto a neighborhood of the identity (by the statement in section P10.18**). Thus, we have the following uniqueness: small skew-symmetric operators A will be unique small logarithms for orthogonal operators e^A. Now establish the following existence: orthogonal operators that are close to the identity have small logarithms. By **(3)**, the existence

[29] Similitude transformations $A \mapsto C^{-1} \circ A \circ C$ with the orientation-preserving linear operators C transpose the elements inside the hyperboloids' sheets, whereas transformations with the orientation-reversing operators C transpose the sheets of every hyperboloid.

[30] Any point of a hyperboloid becomes the vertex (corresponding to one of $2\pi nI$, $-2\pi nI$) for a suitably defined scalar product in $\mathbb{R}^2$.

[31] This description is completely analogous to that in section P7.4 (the "2×2 Matrices That Are Roots of Unity" problem group above), with the difference that $\log E$ has a "logarithmic ramification," whereas $\sqrt[n]{E}$ has an "algebraic ramification."

holds for R_θ because $\theta \cdot I$ are the desired logarithms. Also, the existence holds for the identity operators (of any dimension) because zero operators (of the same dimension) are the desired logarithms. Consequently, it will hold for block operators with blocks R_θ and (optionally) $E_1 = 1$. Establish the existence of all orthogonal operators that are close to the identity with the help of the spectral theorem.

(5) $O(n)$ will be a smooth submanifold of dimension dim o(n) $= \binom{n}{2}$ in the space of linear operators on $\mathbb{R}^n$ if in a neighborhood of any fixed element B it is defined by a nondegenerate system of $n^2 - \binom{n}{2} = \binom{n+1}{2}$ smooth equations in the matrix elements X_{ij}. Using the implicit function theorem, prove that

$$Y_{ij}(X) + Y_{ji}(X) = 0: \quad 1 \le i \le j \le n, \qquad \text{for} \quad Y(X) = \log(B^{-1} \circ X)$$

is such a system, where log is the inversion of the exponential function in a neighborhood of the identity, as discussed in **(4)**. **QED**.

In the modern language of differential topology, a diffeomorphism $X \overset{\varphi_E}{\mapsto} \log X$ equips the orthogonal group with smooth coordinates in a neighborhood of the identity $\mathcal{U}(E)$, and, similarly, diffeomorphisms $X \overset{\varphi_B}{\mapsto} \log(B^{-1} \circ X)$ equip it with smooth coordinates in the neighborhoods $\mathcal{U}(B) = B \circ \mathcal{U}(E)$ of the elements B. The maps φ_B form a **smooth atlas**, which means that the "coordinate recalculation maps"

$$\varphi_{B_1} \circ \varphi_{B_2}^{-1}: \quad \varphi_{B_2}(\mathcal{U}(B_1) \cap \mathcal{U}(B_2)) \to \varphi_{B_1}(\mathcal{U}(B_1) \cap \mathcal{U}(B_2))$$

are diffeomorphisms.[32]

A different method for calculating dim $O(n)$ was discussed previously in the problem group "A Property of Orthogonal Matrices" (section E8.6).

(6) The Gram matrix of a bilinear form with respect to any fixed basis is degenerate if and only if some nontrivial linear combination of the corresponding basis elements is left-orthogonal to all these elements and, therefore, left-orthogonal to the entire space. A similar statement can be made for right orthogonality. (Work out the details.) Thus, nondegeneracy of the Gram matrix with respect to any basis is equivalent to the absence of left and right kernels of the form.

(7) As discussed previously in the problem group "A Property of Orthogonal Matrices" (section H8.11), for a nondegenerate form $\langle ., . \rangle$ on space L, any linear functional on L can be represented as $\langle ., z \rangle$, with a proper $z \in L$, and the correspondence $z \mapsto \langle ., z \rangle$ determines an isomorphism (duality) $L \overset{D}{\to} L^*$. Applying this to the functionals $f_y(x) = \langle Ax, y \rangle$ ($y \in L$) yields the unique existence and linearity of the adjoint operator and establishes the formula $A^* = D^{-1} \circ F_A$, where F_A denotes the linear map $y \mapsto f_y$. (Work out the details; also, find an explicit expression for the

[32] In Lie group theory this is referred to as a **left-invariant atlas**.

matrix A^* with respect to a fixed basis. A^* should be expressed using the matrix A and the Gram matrix. Describe how the matrix A^* changes if variables change linearly.)

Next, obviously, $F_E = D$, hence, $E^* = E$. In addition, we have $F_{AB} = F_B \circ A^*$, which gives

$$(A \circ B)^* = D^{-1} \circ F_{AB} = D^{-1} \circ F_B \circ A^* = D^{-1} \circ F_B \circ D^{-1} \circ F_A = B^* \circ A^*.$$

(Readers have probably noticed the implicit use of the uniqueness of the adjoint operator in those two deductions, which hide it in the formula $A^* = D^{-1} \circ F_A$. A simple exercise is to find arguments that use the uniqueness explicitly.) In particular, $E = (A \circ A^{-1})^* = (A^{-1})^* \circ A^*$, so $(A^{-1})^* = (A^*)^{-1}$.

Also, by induction, we have $(A \circ \ldots \circ A)^* = A^* \circ \ldots \circ A^*$, and, because of the continuity of the map $A \mapsto A^*$,[33] $e^{(A^*)} = (e^A)^*$. (Work out the details.)

Using the preceding expression for the matrix of operator A^* prove that the bilinear form is either symmetric or skew-symmetric if and only if $A^{**} = A$, $\forall A$.

(8) A verification of the fact that g is a Lie algebra can be made by straightforward computations; the fact that G is a group follows directly from the equalities $(A \circ B)^* = B^* \circ A^*$ and $(A^{-1})^* = (A^*)^{-1}$. The fact that G is closed is implied by the continuity of both the map $A \mapsto A^*$ and the matrix multiplication. Finally, the inclusion $\exp(\mathfrak{g}) \subseteq G$ is shown in exactly the same way as for the previously discussed skew-symmetric algebras and orthogonal groups in **(1)**. (Work out the details.)

(9) The inclusion $\{A\colon e^{tA} \in G,\ \forall t \in \mathbb{R}\} \subseteq \mathfrak{g}$ is proved using the asymptotic formula $e^{tX} \circ e^{tY} \circ e^{-tX} \circ e^{-tY} = E + t^2[X, Y] + o(t^2)$ as $t \to 0$, established in section H10.18, which implies the commutativity of X and Y if $e^{tX} \circ e^{tY} \circ e^{-tX} \circ e^{-tY} = E + o(t^2)$. The rest is done by applying the same method as in **(5)** discussed previously.

(10) Establish that $z' \cdot \bar{z''} = uz' \cdot \bar{u}\bar{z''} = |u|^2 \cdot z' \cdot \bar{z''}$ $(z', \bar{z''} \in \mathfrak{L})$ is equivalent to $|u|^2 = 1$.

(11) The matrix equation $\begin{pmatrix} x & z \\ y & t \end{pmatrix} \circ \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \circ \begin{pmatrix} x & y \\ z & t \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is equivalent to the scalar equation $xt - yz = 1$, so $SP(2) = SL(2,\mathbb{R})$, **QED**. (We leave it to the reader to fill in the details.)

To determine the structure of this group, consider its subgroups:

$$G_1 = \left\{ \begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix} \colon x > 0 \right\} \cong \langle \mathbb{R}^{+*}, g \rangle,$$

[33] Recall that all linear maps of finite-dimensional spaces are continuous. Interested readers will find a far-reaching discussion concerning the map $A \mapsto A^*$ in infinite-dimensional cases and related subjects in Riesz and Sz.-Nagy (1972) and references therein.

$$G_2 = \left\{ \begin{pmatrix} 1 & y \\ 0 & 1 \end{pmatrix} : \, y \in \mathbb{R} \right\} \cong \langle \mathbb{R}, + \rangle,$$

$$G_3 = \left\{ \begin{pmatrix} u & -v \\ v & u \end{pmatrix} : \; u, v \in \mathbb{R}, \;\; u^2 + v^2 = 1 \right\} \cong \mathbb{S}^1.$$

Verify that the intersection of each pair of the preceding three subgroups is the trivial subgroup $\{E\}$. Show that the set $G = G_1 G_2 := \{g_1 \circ g_2 : \; g_1 \in G_1, \; g_2 \in G_2\}$ is also a subgroup. It is not an Abelian group, so it is not a direct product of G_1, G_2 (in fact, G_2 is an invariant subgroup in G, but G_1 is not)[34]; but nonetheless, a map $\begin{cases} \mathbb{R}^2 \cong \mathbb{R}^{+*} \times \mathbb{R} & \cong & G \\ (t, y) \mapsto (e^t, y) & \mapsto & \begin{pmatrix} e^t & y \\ 0 & e^{-t} \end{pmatrix} \end{cases}$ is an obvious homeomorphism and, moreover, a diffeomorphism of smooth manifolds. (Work out the details.) Next, prove that the matrix equation

$$\begin{pmatrix} x & y \\ 0 & x^{-1} \end{pmatrix} \circ \begin{pmatrix} u & -v \\ v & u \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

with a, b, c, and d such that $ad - bc = 1$, is uniquely solvable for $x > 0$ and $u^2 + v^2 = 1$, and so $SL(2,\mathbb{R}) = GG_3$. (This is not a direct product since neither G nor G_3 is actually an invariant subgroup.) Prove that $SL(2,\mathbb{R})$ as a set is bijective to a Cartesian product $G \times G_3$ and, moreover, a map $\begin{cases} G \times G_3 & \to & SL(2, \mathbb{R}) \\ (g, g_3) & \mapsto & g \circ g_3 \end{cases}$ is a diffeomorphism of smooth manifolds. (Provide the details.)

(12) The Gram matrix of a bilinear skew-symmetric form with respect to any basis is skew-symmetric. Verify that skew-symmetric matrices of odd dimensions have zero determinants. Therefore, as follows from **(5)** discussed previously, there are no nondegenerate skew-symmetric forms on $\mathbb{R}^{2n+1}$.

Next, fix any symplectic (nondegenerate skew-symmetric) form [,] on $\mathbb{R}^{2n}$. Its Gram matrix with respect to a Darboux basis (if such exists) is $G_D = \begin{pmatrix} \boxed{\begin{matrix} 0 & -1 \\ 1 & 0 \end{matrix}} & & \\ & \ddots & \\ & & \boxed{\begin{matrix} 0 & -1 \\ 1 & 0 \end{matrix}} \end{pmatrix}$; the blocks correspond to two-dimensional orthogonal planes, with respect to [,]. (Usually, orthogonality with respect to a bilinear skew-symmetric form is referred to as **skew-orthogonality**.) Obviously, any two properly normed nonproportional vectors form a Darboux basis in $\mathbb{R}^2$.

[34] Such a product is called **semidirect**.

Prove by induction on n that for any even spatial dimension $2n$, a Darboux basis exists and that any nonzero vector can be included in such a basis.

Next, a *complex structure* is set by a linear operator I such that $I^2 = -E$. Using either the Jordan canonical form over the reals or using induction on the spatial dimension, verify that any such operator I has the same matrix as G_D with respect to an appropriate basis. (In turn, obviously, $G_D^2 = -E$.)

Finally, the operator $I = (-E)^{1/2}$ defined with respect to a Darboux basis by the same matrix as G_D, discussed previously, sets a complex structure agreeing with [,] (why?); consequently, $\langle , \rangle = [I., .] + i[,]$ is a Hermitian form. (Verify the details.) **QED**.

A subspace L of a space N equipped with a bilinear form [,] is referred to as **isotropic**, or **null space** when $[L,L] = 0$, that is, $[x,y] = 0$, $\forall x,y \in L$. For a nondegenerate form, show that $2 \cdot \dim L \leq \dim N$. For a symplectic form on $\mathbb{R}^{2n}$, a null space of the maximal dimension n is referred to as a **Lagrangian subspace** (or **Lagrangian plane**). For example, for $n = 1$ the Lagrangian subspaces are the straight lines passing through the origin. (Why?) Lagrangian planes exist for any dimension; indeed, for any given Darboux basis, list 2^n mutually transversal coordinate Lagrangian planes. Verify that arguments proving the existence of Darboux bases actually yield a strengthened claim:

Proposition. *In a symplectic space (a vector space with a fixed symplectic form), a null space is always contained in a Lagrangian plane:* $[K,K] = 0 \Rightarrow \exists L \supseteq K$: $[L,L] = 0$ & $\dim L = n$. *With this, there exists a Darboux basis* $\{\xi_j,\eta_j\}_{j\,=\,1,\ldots,n}$ *such that all* $\xi_j \in L$, $\xi_1,\ldots, \xi_{\dim\, K} \in K$ *and all* η_j *belong to a transversal to L Lagrangian plane. Hence, each Lagrangian plane has a transversal one.*[35]

Obviously, the operator $I = (-E)^{1/2}$, setting a complex structure agreeing with a symplectic form [,], preserves the property of being a Lagrangian plane; show that Lagrangian planes are actually mapped onto transversal Lagrangian planes: $[L,L] = 0$ & $\dim L = n \Rightarrow [I(L),I(L)] = 0$ & $L + I(L) = \mathbb{R}^{2n}$. In fact, this is not only a necessary but also a sufficient condition for the agreement between the complex and the symplectic structures! A complete statement is as follows:

Theorem. *The complex structure set by* $I = (-E)^{1/2}$ *agrees with a symplectic form if and only if I maps Lagrangian planes onto transversal ones. (Therefore, any complex and symplectic structures on* $\mathbb{R}^2$ *agree.) With this, I has the same matrix as* G_D *(discussed previously) with respect to an appropriate Darboux basis.*

Prove it.

Does I have the same matrices with respect to any Darboux bases?

Is the following formulation correct: "the complex and symplectic structures agree when I maps some Lagrangian plane onto a transversal one"?

The formalism of symplectic algebra and symplectic geometry based on it plays a great role in classical and modern branches of mathematical physics, such as Hamiltonian mechanics, geometrical optics, integration of evolutionary equations and others, and applied mathematical branches such as optimal control and decision making and others; we encourage more advanced readers to consult Arnol'd (1989, 1990), Arnol'd et al. (1985)and multiple references therein. Here is the illustration from **phenomenological thermodynamics**. The points of the four-dimensional positive orthant $(\mathbb{R}^{+*})^4 = \{p > 0, v > 0, S > 0, T > 0\}$ model the **thermodynamically equilibrium macrostates** of a system, where the coordinates correspond to, respectively, the **pressure** (p), the

[35] Also, readers can prove that for any two mutually transversal Lagrangian planes a third Lagrangian exists that is transversal to both of them.

volume of a unit mass (v), the **entropy** (S), and the **absolute** (**Kelvin**) **temperature** (T). The **one-component thermodynamic systems** are modeled by specific two-dimensional surfaces $\mathfrak{W}$ in this orthant, called the **Lagrangian surfaces** (or the **Lagrangian submanifolds**). This means that the tangent plane at any point of that surface is a Lagrangian plane with respect to a symplectic form $dT \wedge dS - dp \wedge dv$. An equivalent definition of $\mathfrak{W}$ stipulates that the change in local coordinates $(p,v) \mapsto (S,T)$ on $\mathfrak{W}$ must be area-preserving, $dT \wedge dS = dp \wedge dv$, that is, this change has its Jacobian equal to one: $\det \begin{pmatrix} \partial T/\partial p & \partial T/\partial v \\ \partial S/\partial p & \partial S/\partial v \end{pmatrix} = 1$.[36] Another equivalent definition [based on the "Stokes" (Newton-Leibnitz-Ostrogradskii-Gauss-Green-Stokes-Poincaré) formula from integral calculus] states the existence on $\mathfrak{W}$ of the following functions (generally multifunctions) referred to as **thermodynamic potentials**: **internal energy** $\mathcal{E}$ ($d\mathcal{E} = T\,dS - p\,dv$), **enthalpy** $\mathcal{W}$ ($d\mathcal{W} = T\,dS + v\,dp$), **Helmholtz potential** (or **free**) **energy** $\mathcal{F}$ ($d\mathcal{F} = -S\,dT - p\,dv$), and **Gibbs potential** (or **free**) **energy** Φ ($d\Phi = -S\,dT + v\,dp$). These potentials are (minus) **Legendre images** of each other, for example, $\mathcal{W} = \mathcal{E} + pv, \mathcal{F} = \mathcal{E} - TS, \Phi = \mathcal{E} - TS + pv = \mathcal{F} + pv = \mathcal{W} - TS$, or, using coordinate systems (v,S), (p,S), (v,T), and (p,T) on $\mathfrak{W}$ (when it is diffeomorphically projected onto the corresponding coordinate planes),

$$\begin{aligned}
\mathcal{W}(p,S) &= \min_v (pv + \mathcal{E}(v,S)),\\
\mathcal{F}(v,T) &= \min_S (-TS + \mathcal{E}(v,S)),\\
\Phi(p,T) &= \min_{v,S} (pv - TS + \mathcal{E}(v,S)) = \min_v (pv + \mathcal{F}(v,T)) = \min_S (-TS + \mathcal{W}(p,S)).
\end{aligned}$$

Surface $\mathfrak{W}$ coincides with the graphs of gradients of these potentials: $\mathfrak{W} = \{(-p,T) = \nabla \mathcal{E}\,(v,S)\}$ etc.[37] Readers may associate these formulations with the usual physical foundations of thermodynamics (Fermi 1937; Feynman et al. 1963; Rumer and Ryvkin 1972; Landau and Lifshitz 1976; Schutz 1982). Actually, methods of symplectic geometry and singularity theory provide a full phenomenological description including such important phenomena as phase transitions (following the **Maxwell phase equilibrium condition**, e.g., the famous **equal area rule** for the **Van der Waals equation**) and others. Interested readers will find a further development and discussions of related aspects in Rumer and Ryvkin (1972)), Landau and Lifshitz (1976), Poston and Stewart (1978), Gilmore (1981), and references therein. A far-reaching development involving Lagrange surfaces in $\mathbb{R}^{2n}$ for multidimensional thermodynamic system (e.g., considering electrodynamic parameters) appears in recently published works by V.P. Maslov and other authors.

Determinants of symplectomorphisms may be easily found using a technique of exterior algebra (discussed previously in the problem group "A Property of Orthogonal Matrices") (Arnol'd 1989). A symplectomorphism B remains the skew-symmetric 2-form $[,] = \sum x_j \wedge y_j$, and so its nth exterior degree $\Omega = \overset{n}{\wedge}[,] = n!\, x_1 \wedge y_1 \wedge \ldots \wedge x_n \wedge y_n$ remains invariant. Therefore, we have, using the lemma in section P8.2,

$$\Omega(v_1, \ldots, v_{2n}) = \Omega(Bv_1, \ldots, Bv_{2n}) = \det B \cdot \Omega(v_1, \ldots, v_{2n}),$$

[36] Actually, this is a restricted formulation because a coordinate system (p,v) [or (S,T)] may be defined only on open subsets of $\mathfrak{W}$ projected onto this coordinate plane without singularities.

[37] A definition and basic properties of the Legendre transform were discussed previously in the problem group "Convexity and Related Classical Inequalities."

which, for linearly independent $v_1,\ldots,v_{2n}$, implies $\det B = 1$. (Work out the details.)

Finally, the reciprocity of the symplectomorphism characteristic polynomials may be shown as follows (Arnol'd 1989). For the matrix of a symplectomorphism with respect to a Darboux basis we have $\mathrm{B} \circ G_D \circ {}^t\mathrm{B} = G_D$; thus, $\mathrm{B} = G_D^{-1} \circ {}^t\mathrm{B}^{-1} \circ G_D$, and so, taking into account that $\det {}^t = \det = 1$,

$$\begin{aligned}\det(z\mathrm{E} - \mathrm{B}) &= \det\left(z\mathrm{E} - G_D^{-1} \circ {}^t\mathrm{B}^{-1} \circ G_D\right) = \det\left(z\mathrm{E} - {}^t\mathrm{B}^{-1}\right)\\ &= \det\left(z\mathrm{E} - {}^t\mathrm{B}^{-1}\right) \cdot \det {}^t\mathrm{B} = \det(z\,{}^t\mathrm{B} - \mathrm{E}) = (-z)^{2n} \det\left(z^{-1}\mathrm{E} - {}^t\mathrm{B}\right)\\ &= z^{2n} \det\left(z^{-1}\mathrm{E} - \mathrm{B}\right).\end{aligned}$$

QED. (We leave it to the reader to fill in the details.)

(13) A commonly known spectral theorem for Hermitian forms states that such a form diagonalizes in appropriate bases of $\mathbb{C}^n$. [They are called Hermitian bases; the elements of $U(n)$ corresponding to a fixed Hermitian form just permute its Hermitian bases.] Verify that if a linear operator A has, with respect to a fixed Hermitian basis, a matrix ς, then A^* will have, with respect to the same basis, the matrix $\mathrm{A}^* = {}^t\bar{\mathrm{A}}$ (transposition plus complex conjugation). Derive from this that $\dim \mathrm{u}(n) = 2{\cdot}[1 + \ldots + (n-1)] + n = n^2$.

Next, show that if a linear operator A has, with respect to any fixed Darboux basis, a block matrix $A = \begin{pmatrix} \mathrm{A}_{11} & \mathrm{A}_{12} \\ \mathrm{A}_{21} & \mathrm{A}_{22} \end{pmatrix} \begin{matrix} \}x \\ \}y \end{matrix}$, then A^* will have, with respect to the same basis, the matrix $A^* = \begin{pmatrix} \mathrm{A}_{11}^* & \mathrm{A}_{12}^* \\ \mathrm{A}_{21}^* & \mathrm{A}_{22}^* \end{pmatrix}$ of the blocks $\mathrm{A}_{12}^* = -{}^t\mathrm{A}_{12}, \mathrm{A}_{21}^* = -{}^t\mathrm{A}_{21}$, $\mathrm{A}_{11}^* = {}^t\mathrm{A}_{22}$, $\mathrm{A}_{22}^* = {}^t\mathrm{A}_{11}$. Derive from this that $\dim \mathrm{sp}(2n) = n^2 + n(n+1) = n(2n+1)$.

(14) The invariance of v_A and μ_B with respect to linear changes in variables follows from the explicit expressions for matrices of X and X^* in section E10.27 **(7)**: $\mathrm{X} \mapsto C \circ \mathrm{X} \circ C^{-1}$, $\mathrm{X}^* \mapsto C \circ \mathrm{X}^* \circ C^{-1} \Rightarrow \mathrm{A} + \mathrm{A}^* \mapsto C \circ (\mathrm{A} + \mathrm{A}^*) \circ C^{-1}$, $\mathrm{B} \circ \mathrm{B}^* \mapsto C \circ \mathrm{B} \circ \mathrm{B}^* \circ C^{-1}$, so $v_A \mapsto v_A$ and $\mu_B \mapsto \mu_B$, **QED**. (We leave it to the reader to fill in the details.) The rest of the claim can be proved in the same way as in **(8)**, using **(7)**; in this way, readers will find that the symbol v_A is additive ($v_{A_1+A_2} = v_{A_1} + v_{A_2}$), the symbol μ_B is multiplicative ($\mu_{B_1 \circ B_2} = \mu_{B_1}\mu_{B_2}$), that both of these symbols are continuous ($v_{\lim A} = \lim v_A$, $\mu_{\lim B} = \lim \mu_B$), and also that $A_1, A_2 \in cg \Rightarrow [A_1, A_2] \in cg$ & $v_{[A_1,A_2]} = 0$ and $A \in cg \Rightarrow e^A \in CG$ & $\mu_{e^A} = e^{v_A}$. (Work out the details.)

(15) Because μ_B is multiplicative (obtained in **(14)**) and $\mu_E = 1$, $B \mapsto \mu_B$ is a homomorphism. (Work out the details using section H10.0 as a template.) Obviously, this homomorphism has group G as its kernel.

Next, the equality $\mu_{B^{-1}} = \mu_B^{-1}$, which is necessary for the homomorphism, may also be verified directly: $B^* = \mu_B B^{-1}$, so $(B^{-1})^* = (B^*)^{-1} = \mu_B^{-1} B$ and $B^{-1} \circ (B^{-1})^* = \mu_B^{-1}$. (Work out the details.)

Next, for $B \in CG$, B commutes with B^* since $B^* = \mu_B B^{-1}$. For the same reason, we have $B^{**} = \mu_B (B^{-1})^* = \mu_B (B^*)^{-1}$. In turn, this equality implies that $\mu_{B^*} = \mu_B$.

Verify that given a nondegenerate bilinear form, on a vector space, the linear operators such that $X^{**} = X$ form a subalgebra of the operator algebra in this space (which means invariance of this property with respect to the linear operations and compositions)[38]; hence, invertible operators having this property form a subgroup of the linear group. Thus, we have shown that this subgroup contains CG.

(16) $A \mapsto v_A$ is a homomorphism of the additive structure of the algebra cg onto the additive group of the field of scalars, having g as its kernel. (In particular, $v_{-A} = -v_A$.) This homomorphism is invariant with respect to the conjugation $v_A = v_{A^*}$. In addition, A commutes with A^* for $A \in \mathrm{cg}$.

(17) For $B \in CG$ close to E, μ_B is close to 1 as a result of the continuity of the map $B \mapsto \mu_B$. The rest is done by proceeding by the same method as in **(9)** discussed previously. [Work out the details; specifically, show that because $B \circ B^* = \mu_B E$ for B close to E {see section H10.27 **(15)**} then, for the logarithm $A = \ln B$ that is close to zero, the following two equalities are true: $[A,A^*] = 0$ and $A + A^* = (\ln \mu_B)\, E$.]

(18) Using the formula for matrices of adjoint operators from section E10.27 **(7)** discussed previously, show that for the bilinear nondegenerate forms $\det A = \det A^*$. Therefore, the numbers μ_B are squares in the field of scalars and, specifically, for the real field they are positive. These arguments are wrong for sesquilinear forms, but readers can establish that $\mu_B > 0$ for the unitary groups using the formula for the matrices of adjoint operators with respect to Hermitian bases from section H10.27 **(13)** discussed previously.

(19) For $B \in CG$, obviously, $\left(\mu_B^{-1/2} E\right) \circ B = B \circ \left(\mu_B^{-1/2} E\right) \in G$, which completes the proof. (Work out the details.)

(20) Show that preserving the orthogonality implies the affine property that consists of transposing affine subspaces of equal dimensions; in particular, two-dimensional affine planes are mapped onto analogous ones. Next, show that circles are mapped onto circles and diametrically opposite points are mapped onto diametrically opposite ones. Since straight lines are mapped onto straight lines too, this gives that the center of a circle is mapped onto the center of the circle image. (Why?) Therefore, T multiplies the lengths of the affine vectors by the same constant (why?), so a composition with a shift $x \mapsto x - T(0)$ and an appropriate dilatation makes T an isometry with respect to the Euclidean metric. The proof is completed with the following lemma.

Lemma. *An isometry of a Euclidean space that retains its origin as invariant is linear (and, thus, orthogonal).*

[38] As discussed previously in **(7)**, this subalgebra may be distinct from the whole algebra unless the source bilinear form is neither symmetric nor skew-symmetric.

Advanced readers will find arguments to prove that this lemma is also applicable to infinite-dimensional spaces with continuous bilinear positive definite forms (pre-Hilbert and Hilbert spaces).[39]

H10.28

Readers may establish the equality $\det e^A = e^{\mathrm{tr}\ A}$ using any of the four following methods:

Method 1. Use the formula for e^A from section P10.19** and the asymptotic formula for $\det(E + \varepsilon A)$ from section P10.22**.

Method 2. Use the lemma from section H10.27 concerning the eigenvalues of e^A.

Method 3. Using the permutability of the exponential function with the matrix similitude (see lemma from section H10.19) and using the determinant and the trace invariance with respect to the similitude, it is enough to establish the foregoing equality for a Jordan matrix. Since the determinant (trace) of a block matrix is the product of determinants (resp. the sum of traces) of the blocks, it is sufficient to consider a Jordan box. Finally, use the explicit formula for the exponent of a Jordan box from section H10.18.

Method 4. Apply the differential equation from section P10.22** to the matrix function $W(t) = e^{At}$.

To prove that $SL(n,\mathbb{R})$ is a smooth submanifold of codimension one in the matrix space, use the equality $\mathrm{sl}(n,\mathbb{R}) = \{\mathrm{tr}\ A = 0\}$. [Apply the theorem of Lie group theory from section P10.27** **(9)** and use section H10.27 **(5)** as a template.]

H10.29

Use the Jordan canonical form to show the equivalence of the existence of logarithms, the existence of square roots, and the conditions from section P10.29**.

Sufficient and necessary conditions of existence of $\sqrt[p]{\ }$ may be formulated as follows:

1. For $B \in GL(n,\mathbb{C})$, $\sqrt[p]{B}$ always exist in $GL(n,\mathbb{C})$.
2. For $B \in GL(n,\mathbb{R})$, $\sqrt[2p+1]{B}$ always exist in $GL(n,\mathbb{R})$.
3. For $B \in GL(n,\mathbb{R})$, $\sqrt[2p]{B}$ exist in $GL(n,\mathbb{R})$ if and only if the Jordan boxes corresponding to the negative eigenvalues come in equal pairs.
4. For B with complex (real) entries, $\sqrt[p]{B}$ with complex (resp. real) entries exist if and only if B satisfies conditions 1–3 regarding its nonzero eigenvalues, and the

[39] Origin-preserving isometries are always linear, at least in any normed spaces, as S. Mazur and S. Ulam proved; however, for non-Hilbert norms, this claim is not so easily obtained (Banach 1932 and references therein).

Jordan boxes corresponding to zero eigenvalue can be arranged in (distinct) groups $G^i_{k,r}$ ($i = 1,2,\ldots$, $k \in \{0,1,\ldots\}$, $r \in \{0,\ldots, p-1\}$), each consisting of $p - r$ boxes of dimension k and r boxes of dimension $k+1$.[40]

H10.30

We suggest that readers follow steps A–E below.

A. Prove that a given continuous homomorphism $\lambda\colon \langle \mathbb{R},+\rangle \to \langle \mathbb{C}^*,\bullet\rangle$ (resp. $\lambda\colon \langle \mathbb{R},+\rangle \to \langle \mathbb{R}^{+*},\bullet\rangle$) with $\lambda(t_0) = \lambda_0$, a Jordan box $J = \begin{pmatrix} \lambda_0 & 1 & \\ & \ddots & \ddots \\ & & \end{pmatrix}$ is included as element $g(t_0)$ in a one-parameter group $g(t)$, which acts on J's eigenspace by multiplying by the scalars $\lambda(t)$: $g(t)v = \lambda(t)v$, where $v = {}^t(v_1,0\ldots)$. Similarly, given a homomorphism $\lambda : \langle \mathbb{R},+\rangle \to \langle \pounds^*, g\rangle$ with $\lambda(t_0) = \lambda_0$, a Jordan matrix J consisting of two boxes of equal dimensions corresponding to $\lambda_0 < 0$ or complex conjugate pair $\{\lambda_0, \overline{\lambda_0}\}$ is included as the element $g(t_0)$ in a one-parameter group $g(t)$ in $GL(n,\mathbb{R})$ such that the complexification of $g(t)$ acts on J's eigenspace in the complexified space as the scalar operator $\lambda(t)$.

B. By section P10.2*, in the real-valued case, $\lambda(t)$ is uniquely determined as $\lambda(t) = e^{(\ln \lambda_0)\, t/t_0}$. By section P10.13 (**C**), in complex-valued cases, $\lambda(t)$ is determined in infinitely many ways, namely, by the equations $|\lambda(t)| = e^{(\ln|\lambda_0|)\, t/t_0}$ and $\arg \lambda(t) = (\arg \lambda_0 + 2\pi n)\, t/t_0$ for any fixed integer n. This completes the solution of Problem P10.30** for the complex-valued case and the case of two Jordan boxes of equal dimensions corresponding to a negative eigenvalue or a complex conjugate pair. (Work out the details.)

C. Prove that the group $g(t)$ in **A** is uniquely determined by the homomorphism $\lambda(t)$.

D. Let J be a Jordan matrix consisting of boxes $J_i = \begin{pmatrix} \lambda_0 & 1 & \\ & \ddots & \ddots \\ & & \end{pmatrix}$ ($i = 1,\ldots,k$) of any dimensions $m_1,\ldots, m_k$ corresponding to the same eigenvalue. According to **A**, given a continuous homomorphism $\lambda : \langle \mathbb{R},+\rangle \to \langle \pounds^*, g\rangle$ (resp. $\lambda : \langle \mathbb{R},+\rangle \to \langle \mathbb{R}^{+*}, g\rangle$) with $\lambda(t_0) = \lambda_0$, J can be included as the element $g(t_0)$ in a one-parameter group $g(t)$ that acts on J's eigenspace as scalar operator $\lambda(t)$. Prove that the group $g(t)$ is determined uniquely (which generalizes the similar statement in **C** for $k > 1$).

E. Consider a Jordan block J of boxes corresponding to a fixed $\lambda_0 > 0$ (as in **D**). Assume that the roots from J that can be included in one-parameter linear groups have only positive eigenvalues. Then these eigenvalues are equal to $\sqrt[p]{\lambda_0}$ ($p = 1,2,\ldots$), which uniquely determines $\lambda(t)$ (as $\lambda(t) = e^{(\ln \lambda_0)\, t/t_0}$; work out the

[40] A Jordan box of dimension zero is "no box."

details!) and, thus, the group $g(t)$ (by **D**). On the other hand, a root from J that can be included in a one-parameter linear group has a logarithm; hence, if it has negative or complex conjugate eigenvalues, then (see section P10.29**) the Jordan boxes of equal dimensions corresponding to the same eigenvalue must exist. Therefore, J will possess such Jordan boxes as well. (Furnish the details.) This proves that a linear operator from $GL(n,\mathbb{R})$ having all its eigenvalues positive, with no distinct Jordan boxes of equal dimensions corresponding to the same eigenvalue, can be included in the unique one-parameter linear group g in $GL(n,\mathbb{R})$ as $g(t_0)$, given $t_0 \neq 0$.

Conversely, verify that a Jordan matrix of two boxes of equal dimensions corresponding to the same eigenvalue $\lambda_0 > 0$ has p distinct roots of degree p for any natural p, and there exists a logarithm for each such root (and thus, the root can be included in a one-parameter linear group). This completes the solution of Problem P10.30**.

H10.31

Apply the lemma from section H10.27 **(3)**, Gelfand's formula from section P7.8 (the "2 × 2 Matrices That Are Roots of Unity" problem group discussed previously), and Euler's formula (in section P10.24**).

Explanation

E10.0

To show that the image of a group homomorphism $g: A \to B$ is always a subgroup of B, we must verify that

(1) $b_1, b_2 \in g(A) \Rightarrow b_1 b_2 \in g(A)$
(2) The identity element of B belongs to $g(A)$
(3) $b \in g(A) \Rightarrow b^{-1} \in g(A)$

The verification may look as follows:

(1) $b_1 = g(a_1), b_2 = g(a_2) \Rightarrow b_1 b_2 = g(a_1 a_2)$
(2) $g(A)$ is a group and has its own identity element e (why?), so we will have $ee^{-1} = e$, and hence, e is an identity element of B (by virtue of uniqueness)
(3) $b = g(a) \Rightarrow bg(a^{-1}) = e \Rightarrow b^{-1} = g(a^{-1})$ (by virtue of uniqueness)

E10.2

The existence of positive roots from positive numbers follows from the **completeness** of the real numbers' field with respect to its topology. (The completeness is in either case stipulated by defining this field.) The uniqueness of those roots follows from a monotonicity $x > y > 0 \Rightarrow x^q > y^q$ ($q = 1,2,\ldots$), which in turn follows from an elementary formula $x^q - y^q = (x - y)(x^{q-1} + x^{q-2}y + \ldots + y^{q-1})$. (Work out the details.)

E10.3

We have for a fixed t

$$\dot{g}(t) = \frac{d}{ds}\Big|_{s=0} g(s+t) = \frac{d}{ds}\Big|_{s=0} g(s)g(t) = \lim_{s\to 0} \frac{g(s)g(t) - g(0)g(t)}{s} =$$
$$\lim_{s\to 0} \left(\frac{g(s) - g(0)}{s} \cdot g(t) \right) = \left(\lim_{s\to 0} \frac{g(s) - g(0)}{s} \right) \cdot g(t) = \dot{g}(0)g(t).$$

QED.

E10.5

A. $\left| \sum_{k=m}^{n} a_k \right| \le \sum_{k=m}^{n} |a_k|$, so, $\left| \sum_{k=m}^{n} a_k \right| < \varepsilon$ for large m, n if $\sum_{k=m}^{n} |a_k| < \varepsilon$ for those m, n. Therefore, absolute convergence implies convergence.

The famous **Leibniz series** $1 - 1/2 + \ldots + (-1)^{k-1}/k + \ldots$ gives a simple and, probably, the most popular example of convergent but absolutely divergent series. Its convergence is confirmed by the **Leibniz test**: an alternating series $\sum a_n$ converges if $|a_n|$ monotonically decreases (to zero). The series of absolute values is the famous **harmonic series** $1 + 1/2 + \ldots + 1/k + \ldots$; produce an estimate $\sum_{k=2}^{2^n} k^{-1} \ge n/2$ or, equivalently, $\sum_{k=2}^{n} k^{-1} > \log_2 \sqrt{n}$, so the harmonic series diverges. Create a figure to derive a sharper estimate $\ln n > \sum_{k=2}^{n} k^{-1} > \ln(n+1) - 1$. Prove the existence of the limit $c = \lim_{n\to\infty} \left(\sum_{k=1}^{n} k^{-1} - \ln(n+1) \right)$ (which is the famous **Euler**

constant). For this, produce for $c_n = \sum_{k=1}^{n} k^{-1} - \ln(n+1)$ an estimate $0<c_{n+1} - c_n < \frac{1}{n+1} - \frac{1}{n+2}$ that yields $(0<)\ c_{n+1} - c_1 < \frac{1}{2} - \frac{1}{n+2}$, so that $c_{n+1} < c_1 + 0.5 = 1.5 - \ln 2 \approx 0.8069$; on the other hand, c_n grows monotonically.

Comparing the series to a geometric progression proves the d'Alembert test by the Cauchy criterion: $\left|\sum_{k=m+1}^{n} a_k\right| = \sum_{k=m+1}^{n} a_k < C \sum_{k=m+1}^{n} (1-\varepsilon)^k \to 0$ as $m, n \to \infty$ if $0 < a_{n+1}/a_n < 1 - \varepsilon$ for large n.

B. Comparing this to a geometric progression proves Abel's lemma too. That is, we have $\left|c_n t_0^n\right| \le C < \infty, \forall n$ (why?); therefore, for $|\, t\,| \le \mathrm{t}_1 < |\, t_0\,|, |\, c_n t^n\,| \le C\,|\, t_1/t_0|^n$, and so the series $\sum c_n t^n$ converges by the Cauchy criterion:

$$\left|\sum_{k=m+1}^{n} c_k t^k\right| \le \sum_{k=m+1}^{n} |c_k| t_1^k \le C \sum_{k=m+1}^{n} |t_1/t_0|^k \to 0 \quad \text{as} \quad m, n \to \infty.$$

A similar estimate is made for a series of the derivatives because the factors of a polynomial growth cannot disturb the convergence at the rate of a geometric progression: taking $t_1 < \mathrm{t}_2 < |\, t_0\,|$ we will have $n < (t_2/t_1)^n$ for large n, and so,

$$\left|\sum_{k=m+1}^{n} (k+1) c_{k+1} t^k\right| \le C t_1^{-1} \sum_{k=m+1}^{n} |t_2/t_1|^{k+1} |t_1/t_0|^{k+1} = C t_1^{-1} \sum_{k=m+1}^{n} |t_2/t_0|^{k+1} \to 0$$

as $m, n \to \infty$.

C. Prove that a uniformly convergent series is term-by-term integrable, which means that a series of integrals of the terms of a source series uniformly converges to an integral of the sum of this series. Thus, a series $\sum \int_{t_0}^{t} \dot{a}_n(t')\, dt'$ converges uniformly on t_0, $t \in [t_1, t_2]$ and

$$\int_{t_0}^{t} \sum \dot{a}_n(t')\, dt' = \sum \int_{t_0}^{t} \dot{a}_n(t')\, dt' = \sum (a_n(t) - a_n(t_0)),$$

which equals $\sum a_n(t) - \sum a_n(t_0)$ if a series $\sum a_n(t_0)$ converges, and in this case $\sum a_n(t)$ converges uniformly on $t \in [t_1, t_2]$. In addition, it will be differentiable and

$$\frac{d}{dt} \sum a_n(t) = \frac{d}{dt} \left(\sum a_n(t) - \sum a_n(t_0) \right) = \frac{d}{dt} \int_{t_0}^{t} \sum \dot{a}_n(t')\, dt' = \sum \dot{a}_n(t),$$

which completes the proof.

E10.6

A straightforward computation yields

$$\left|S_nT_n-\sum_{k=0}^{n}\frac{(s+t)^k}{k!}\right|=\left|\sum_{1\le k,m\le n,\ n+m\le 2n}\frac{s^kt^m}{k!m!}\right|\le(n!)^{-1}\sum_{n+m\le 2n}\binom{k+l}{k}|s|^k|t|^m$$
$$=(n!)^{-1}\sum_{k=n+1}^{2n}(|s|+|t|)^k\le(n!)^{-1}(|s|+|t|)^{2n+1}\to 0$$

for fixed s, t, and $n\to\infty$. (Work out the details.)

A proof, using Method 1, that exponential series sums are all continuous solutions of (*). Apply the implicit function theorem to establish for small t the existence of unique small $K(t)$ such that $g(t)=e^{K(t)}$. Use Abel's lemma discussed in section H10.5 to verify that the function e^t satisfies the conditions of this theorem. Next, using induction on natural p and taking into account the uniqueness of $K(t)$ and the fact that the exponential series' sums satisfy (*), we can prove that $K(pt)=pK(t)$ if both t and pt are small enough. Therefore, using uniqueness, $K(pt/q)=pK(t)/q$, provided that pt/q is small enough. Taking into account the continuity of $g(t)$ we will obtain the linearity of $K(t)$ for small t. (Work out all the details and complete the proof.)

E10.7

Method 1. Given $\varepsilon>0$, fix natural k_0 large enough so that $\sum_{k=k_0+1}^{m}(k!)^{-1}|t|^k<\varepsilon$ for large m. For $1\le k\le k_0$ and large n, $0\le 1-\prod_{i=0}^{k-1}\frac{n-i}{n}<\varepsilon$ (why?); hence, for a fixed large n we will have

$$\left|\left(\sum_{k=0}^{m}\frac{t^k}{k!}\right)-(1+t/n)^n\right|=\left|\left(\sum_{k=1}^{k_0}+\sum_{k=k_0+1}^{m}\right)\left(1-\prod_{i=0}^{k-1}\frac{n-i}{n}\right)\cdot\frac{t^k}{k!}\right|<\sum_{k=1}^{k_0}\frac{\varepsilon\mid t\mid^k}{k!}$$
$$+\sum_{k=k_0+1}^{m}\frac{\mid t\mid^k}{k!}<\varepsilon\cdot e^{|t|},$$

for any large m. Therefore, for large n,

$$|e^t-(1+t/n)^n|=\lim_{m\to\infty}\left|\sum_{k=0}^{m}(k!)^{-1}t^k-(1+t/n)^n\right|\le\varepsilon\cdot e^{|t|}.$$

QED.

E10.10

Any homomorphism $\langle \mathbb{R}, + \rangle \stackrel{\varphi}{\to} \langle \mathbb{R}, + \rangle$ is a linear function over rationals, $\varphi(pt/q) = p\varphi(t)/q$ (why?); thus, continuity implies linearity over the reals.

E10.13. B

Start from a property of the real field that was known to ancient mathematicians and later called, by David Hilbert (1930), "the Archimedes continuity axiom":

$$\forall \xi, \eta > 0, \quad \exists m \in \{1, 2, \ldots\} : \quad m\xi > \eta.$$

Taking $k = m - 1$ for the minimum of those natural m yields the Euclidean (**division with remainder**) property

$$\forall \xi, \eta : \quad \xi \neq 0 \Rightarrow \exists \mathrm{k} \in \mathbb{Z}, \ \exists \zeta \in [0, \ |\zeta|) : \quad \eta = \mathrm{k}\xi + \zeta.$$

The remainder ζ is a linear combination of ξ, η with integral coefficients, and $\zeta < |\xi|$. Therefore, for a subgroup $H \subset \langle \mathbb{R}, + \rangle$ an element $T \in H \setminus \{0\}$ of a minimal absolute value over H, divides any of $t \in H$. Because of this, $H = T\mathbb{Z}$. As a result, a subgroup that has no nonzero elements of a minimal absolute value is either $\{0\}$ or everywhere dense in $\mathbb{R}$. **QED**. (We leave it to the reader to fill in the details.)

E10.13. C

Completing the proof of the lemma. We must represent Θ, defined in section H10.13. C, as $t \mapsto (\pm 2\pi t/T) \bmod 2\pi$, that is, as a composition $\mathbb{R} \stackrel{\hat{\Theta}}{\cong} \mathbb{R} \stackrel{\Phi}{\to} \mathbb{R}$ mod2π with a linear map $t \stackrel{\hat{\Theta}}{\mapsto} (\pm 2\pi/T)t$ – so as to have a commutative diagram

$$\begin{array}{ccc} \mathbb{R} & \stackrel{\hat{\Theta}}{\to} & \mathbb{R} \\ \downarrow{\scriptstyle \Psi} & \searrow{\scriptstyle \Theta} & \downarrow{\scriptstyle \Phi} \\ \mathbb{R} \bmod T & \to & \mathbb{R} \bmod 2\pi \end{array}$$

(the lower horizontal arrow in the diagram is due to a fundamental homomorphism theorem). Consider a segment $I = [0, \varepsilon] \subset \mathbb{R}$, where $0 < \varepsilon < T$. $\Psi(I)$ is a segment on a circle $\mathbb{R}$ mod T; hence, $\Theta(I)$ is a segment on a circle $\mathbb{R}$ mod 2π. Let J be the inverse image of $\Theta(I)$ with respect to the map Φ lying in $\pm\ [0, \ 2\pi)$.[41] $J = \pm[0, \ \delta]$, where $0 < \delta < 2\pi$, and J is **homeomorphically**

[41] The sign is "+" ("−") when Θ preserves (resp. reverses) the usual orientation of the circle.

mapped onto $\Theta(I)$. (Why?) Thus, a composition of Θ with the foregoing local inversion Φ^{-1}: $\Theta(I) \to J$ gives a continuous additive map $I \to J$, which must be a restriction of a unique linear map $\mathbb{R} \to \mathbb{R}$ to I; since $\delta/\varepsilon = 2\pi/T$ (because of the commutativity of the lower triangle in the diagram), this map is $\hat{\Theta}$. Thus, $\Theta = \hat{\Theta} \circ \Phi$ on I, and, thus, on the entire $\mathbb{R}$. (Work out all details.)[42]

E10.13. D

Necessity of conditions (a) and (b).

(1) Method using Euclid's algorithm. Apply induction on the length of a finite computation procedure using Euclid's algorithm to verify that the last nonzero remainder divides ξ and η. On the other hand, this remainder is divisible by the common divisors of ξ and η because it is their linear combination with integral coefficients, and, thus, it is a GCD. Conversely, the existence of a GCD makes these linear combinations take their values discretely, with a "quantum" equal to the GCD. Hence, any computation procedures using Euclid's algorithm are finite. (Work out the details. Also, derive from these arguments a characterization of GCD as a linear combination of ξ, η with integral coefficients of a minimal absolute value.)[43]

(2) Method based on Dirichlet boxe principle; proof of Poincaré recurrence theorem. If this theorem were violated with some S, we would have $\mathrm{vol}(K) = \infty$, in contradiction to the conditions. Indeed, considering a sequence of subsets $H_n \subset H$ such that $\#H_n = n$ yields

$$\mathrm{vol}(K) \geq \sum_{h \in H_n} \mathrm{vol}\left(h(S) \setminus \bigcup_{f \in H_n, f \neq h} f(S) \right) = n \cdot \mathrm{vol}(S) \to \infty \qquad \text{as} \qquad n \to \infty.$$

[42] Readers familiar with homotopies and coverings will recognize in this argument a special case of the rising path theorem, which makes it possible, in particular, to rise topological group structures and homomorphisms to a covering (Massey 1967; Sze-Tsen Hu 1959).

[43] The absolute value on a cyclic group, generated by this divisor, serves as the so-called **Euclidean function**, which makes this group a **Euclidean module**. Let M be a module over a ring R and ε a function on M taking values in a partially ordered set with the minimum condition. M is referred to as a Euclidean module, with Euclidean function ε, if division with a remainder can be performed in M:

$$\forall \xi, \eta \in M: \quad \xi \neq 0 \quad \Rightarrow \quad \eta = k\xi + \zeta, \quad \textit{with} \quad k \in R, \quad \zeta = 0, \quad \textit{or} \quad \varepsilon(\zeta) < \varepsilon(\xi).$$

Any set of elements of the Euclidean module has a GCD, that is, any element of a submodule generated by this set corresponding to a minimal value of ε is a GCD. For the Euclidean module, computations using Euclid's algorithm are always finite and yield a GCD.

QED. (We leave it to the reader to fill in the details.)

Equivalence of two definitions of commensurability modulo a number. Let $m_1a_1 - m_2a_2 = nb$, with relatively prime m_1, m_2. Since the GCD of two numbers is a linear combination of those numbers with integer coefficients, we have $1 = k_2m_2 - k_1m_1$. Consequently, $\frac{a_1+(nk_1)\,b}{m_2} = \frac{a_2+(nk_2)\,b}{m_1}$, which completes the proof.

Sufficiency of conditions (a) – (c) (case where θ_j are incommensurable with 2π).

i. If for some j, l conditions (a)–(c) were not satisfied, then we would have a system of two independent linear equations on θ_j, θ_l, with integral coefficients and right-hand sides being integral multiples of 2π; thus, θ_j, θ_l would be rational multiples of 2π.

ii. Let GCD(m_{12}, m_{13}) = d, so that $m_{12} = ad$, $m_{13} = bd$, with relatively prime a, b. Eliminating θ_1 from the equations of indices (1,2) and (1,3) produces

$$\frac{\theta_2 + 2\pi k_{21}}{m_{21}b} = \frac{\theta_3 + 2\pi k_{31}}{m_{31}a} + \frac{2\pi(k_{12} - k_{13})}{abd}.$$

By (i), vector (m_{23}, m_{32}) is proportional to ($m_{21}b$, $m_{31}a$). Therefore, combining this equation with an equation of index (2,3) brings

$$\frac{k_{21} - k_{23}}{m_{21}b} = \frac{k_{31} - k_{32}}{m_{31}a} + \frac{k_{12} - k_{13}}{abd}.$$

(Work out the details.) By condition, (m_{12}, m_{21}) and (m_{13}, m_{31}) are pairs of relatively prime numbers, so m_{21} and m_{31} are relatively prime with d, and hence $k_{12} - k_{13}$ must be divisible by d. A similar proof can be produced for any indices j, l, and not just 2, 3.

iii. The covectors (k_{1j}, k_{j1}) are determined by a linear equation $m_{j1}k_{1j} - m_{1j}k_{j1} = n_j$. Therefore, all of them can be represented as sums of one of them and covectors orthogonal to (m_{j1}, $-m_{1j}$). The orthogonal complement is spanned on (m_{1j}, m_{j1}). Because of the relative primality of m_{1j} and m_{j1}, the integral points on this straight line are integral multiples of (m_{1j}, m_{j1}). (Apply the description of the closed subgroups of $\langle \mathbb{R}, + \rangle$ discussed previously in section P10.13. B** to work out the details.) **QED**.

iv. Proof of theorem. We suggest using ring-theoretic language [readers who prefer more elementary arguments may look at two proofs of the most applicable special case where any pair of m_i is relatively prime (Chinese remainder theorem); the proofs follow in this section]: a relation "*a divides b*" corresponds to the principal ideals' inclusion $B \subseteq A$ ($A = \langle a \rangle$, $B = \langle b \rangle$); furthermore, GCD(A,B) = $A + B$ [because GCD(a,b) generates $A + B$]; LCM(A,B) = $A \cap B$ (because LCM(a, b) generates $A \cap B$]; the relative primality of ideals A, B corresponds to equality of $A + B$ to the ring; and so on. The result of this proof can be reformulated as follows.

Theorem. *Given ideals* $A_1,\ldots, A_n \neq (0)$ *of ring* $\mathbb{Z}$ *and* $r_1,\ldots, r_n \in \mathbb{Z}$, *a system of congruences* $x - r_j \in A_j$: $j = 1,\ldots, n$ *is solvable if and only if* $r_j - r_l \in A_j + A_l$ *(j,l* $= 1,\ldots, n$*).*

The *necessity* is obvious: $\exists x, \forall j:\ x - r_j = a_j \in A_j \Rightarrow r_j - r_l = a_l - a_j \in A_l + A_j$. Consider the *sufficiency*. We must show that the intersection of the residue classes $r_j + A_j$ is nonempty. The case of $n = 1$ is trivial. Also, the case of $n = 2$ is easy: we must verify that $r - s \in A + B \Rightarrow (r + A) \cap (s + B) \neq \emptyset$; but since any residue class of $(A + B)$ mod A intersects with B, we have $(r - s + A) \cap B \neq \emptyset$, hence,

$$\emptyset \neq s + (r - s + A) \cap B = (r + A) \cap (s + B).$$

For $n > 2$, start from an obvious inclusion of ideals $(A + B) \cap C \supseteq (A \cap C) + (B \cap C)$. Use the unique factorization in the ring $\mathbb{Z}$ to prove that for $\mathbb{Z}$ this inclusion is an equality. A special case of this equality (for $C \supseteq A$) is $(A + B) \cap (A + C) = A + (B \cap C)$, and by induction, $\bigcap_{j=1}^{n-1} (A_n + A_j) = A_n + \bigcap_{j=1}^{n-1} A_j$.

Now, show that $\bigcap_{j=1}^{n} (r_j + A_j) \neq \emptyset$ by induction on the number of ideals. Suppose we have already done this for $n - 1$, and let $r \in \bigcap_{j=1}^{n-1} (r_j + A_j)$. We obtain

$$r + \bigcap_{i=1}^{n-1} A_i \subseteq \bigcap_{j=1}^{n-1} (r_j + A_j) + \bigcap_{i=1}^{n-1} A_i \subseteq \bigcap_{j=1}^{n-1} \left(r_j + A_j + \bigcap_{i=1}^{n-1} A_i \right) \subseteq \bigcap_{j=1}^{n-1} (r_j + A_j).$$

Therefore, it is sufficient to prove that $\left(r + \bigcap_{j=1}^{n-1} A_j \right) \cap (r_n + A_n) \neq \emptyset$. We have $r = r_j + a_j$, with $a_j \in A_j$ $(j = 1,\ldots, n - 1)$, and by the conditions of the theorem,

$$r_n - r = r_n - r_j - a_j \in (A_n + A_j) + A_j = A_n + A_j$$

Thus, we have $r_n - r \in \bigcap_{j=1}^{n-1} (A_n + A_j) = A_n + \bigcap_{j=1}^{n-1} A_j$, and so $r + \bigcap_{j=1}^{n-1} A_j$ intersects with $r_n + A_n$ (this is the case of $n = 2$ discussed previously). This completes the proof.

We leave it to the reader to make a decision on the applicability of the bounds of this proof. Readers have probably noticed that for $n > 2$ the theorem formally follows from the equality $(A + B) \cap (A + C) = A + (B \cap C)$, and so it will be applicable whenever such an equality arises. Readers may establish this equality for the ideals of a **Euclidean ring**, for the pairwise

relatively prime ideals of an associative unitary ring, for the vector subspaces of a vector space that are in a general position,[44] and for other cases; therefore, in all of these cases similar theorems are applicable!

Proof of Chinese remainder theorem using completely elementary arguments. Let us apply induction on n. For $n = 1$, the claim is trivial. Let $n = 2$. For integers r_1, r_2, there exist integers q_1, q_2 such that $r_1 - r_2 = q_2m_2 - q_1m_1$ (this is because the GCD of natural numbers is presentable by a linear combination of these numbers with integral coefficients; this fundamental fact of elementary number theory is implied by Euclid's algorithm, but it can also be proved directly), so an integer $q_1m_1 + r_1 = q_2m_2 + r_2$ has the residues r_1, r_2 modulo m_1, m_2, respectively. Assume that the claim has been proved for $n - 1$, so there exist integers having any residues $r_2,\ldots, r_n$ modulo $m_2,\ldots, m_n$, respectively. For fixed $r_2,\ldots, r_n$, these integers form arithmetic progressions with differences $m_2\ldots m_n$, or, in other words, there exists a unique residue modulo $m_2\ldots m_n$ such that any integer possessing this residue will also have residues $r_2,\ldots, r_n$ modulo $m_2,\ldots, m_n$, respectively. Thus, since m_1 is relative prime with $m_2\ldots m_n$, the theorem is reduced to its special case of $n = 2$, which completes the proof. (Work out the details.)

Proof of Chinese remainder theorem using arguments from elementary number theory. Denote by $\mathbb{Z}(m)$ the ring of all residues modulo integer m. For $m_1,\ldots, m_n$, a direct (or Cartesian) product $\prod\mathbb{Z}(m_i) = \{(r_1,\ldots, r_n): r_i \in \mathbb{Z}(m_i), \forall i = 1,\ldots, n\}$ furnished with by-component operations becomes a ring, so that the map $\rho: k \mapsto (k \bmod m_1,\ldots, k \bmod m_n)$ will be a homomorphism of rings $\mathbb{Z} \to \prod\mathbb{Z}(m_i)$. The following claim is evident.

Lemma. $\ker\rho = \bigcap \mathbb{Z}m_i$, *i.e.*, $\ker\rho$ *consists of all common multiples of* $m_1,\ldots, m_n$, *so that it is an ideal of the ring* $\mathbb{Z}$ *generated by the least common multiple*.

We will now prove the following theorem.

Factorization theorem. *Let* $m = \prod m_i$ *be the* ***primary factorization*** *of integer* $m(m_i = p_i^{\alpha_i}$, *where* p_i *are distinct integer primes). The map* $\sigma: k \bmod m \mapsto (k \bmod m_1,\ldots, k \bmod m_n)$ *is correctly defined and induces an isomorphism of rings:* $\mathbb{Z}(m) \cong \prod\mathbb{Z}(m_i)$.

Proof of factorization theorem. The map σ is correctly defined as $k_1 \bmod m = k_2 \bmod m \Leftrightarrow \exists q \in \mathbb{Z}: k_1 = k_2 + qm \Rightarrow k_1 \bmod m_i = k_2 \bmod m_i, \forall i = 1,\ldots, n$. σ is a homomorphism of rings since a difference and a product of residues are equal to the residues of the difference and the product, respectively. A composition $\mathbb{Z} \to \mathbb{Z}(m) \to \prod\mathbb{Z}(m_i)$, where the first map is a reduction modulo m and the second one is σ, equals ρ. Representatives of distinct elements of $\mathbb{Z}(m)$ in $\mathbb{Z}$ are not comparable modulo m, or, in other words, the difference between these representatives is not divisible by m, and, hence (by the foregoing lemma), this difference does not belong to $\ker\rho$. Therefore, σ maps distinct elements of $\mathbb{Z}(m)$ onto distinct elements of $\prod\mathbb{Z}(m_i)$, or, in other words, σ is a **monomorphism**. Hence, by the Dirichlet box principle (also known as the pigeonhole principle), it is an isomorphism because $\mathbb{Z}(m)$ and $\prod\mathbb{Z}(m_i)$ have the same finite number of elements. QED.

[44] The subspaces L_i are in a general position when $\operatorname{codim}\bigcap L_i = \sum\operatorname{codim} L_i$ for any finite collection of them. For example, the orthogonal complements to vectors of a Euclidean space that are linearly independent in their totality are in a general position. (Why?)

Arguments that are similar to those described previously show that if m is a common multiple of any fixed natural numbers $m_1,\ldots, m_n$, then a similar homomorphism σ: $\mathbb{Z}(m) \to \prod\mathbb{Z}(m_i)$ is correctly defined and a composition $\mathbb{Z} \to \mathbb{Z}(m) \to \prod\mathbb{Z}(m_i)$ is equal to ρ. Denote m' and σ' as the least common multiples of $m_1,\ldots, m_n$ and the respective homomorphism $\mathbb{Z}(m') \to \prod\mathbb{Z}(m_i)$. σ can be decomposed as $\mathbb{Z}(m) \overset{\text{mod}m'}{\to} \mathbb{Z}(m') \overset{\sigma'}{\to} \prod \mathbb{Z}(m_i)$, so im σ = im σ', and ρ can be decomposed as $\mathbb{Z} \to \mathbb{Z}(m) \to \mathbb{Z}(m') \to \prod\mathbb{Z}(m_i)$. The kernel of ρ consists of all common multiples of $m_1,\ldots, m_n$, so it is generated by m': ker $\rho = \bigcap \mathbb{Z}m_i, = \mathbb{Z}m'$, and hence, ker ρ is mapped onto a zero element of $\mathbb{Z}(m')$ and σ' is a monomorphism. Furthermore, $m' = m_1 \ldots m_n$ if any pair of m_i is relatively prime, and otherwise $m' < m_1 \ldots m_n$. Thus, we have established a more general version of the foregoing theorem:

Factorization theorem. *Let m be a common multiple of natural numbers* $m_1,\ldots, m_n$; *then the preceding homomorphism σ is always correctly defined, is a monomorphism if and only if m is the least common multiple, and is an epimorphism if and only if any pair of* m_i *is relatively prime.*

Thus we have proved the Chinese remainder theorem, as it states that *the preceding map ρ is an* ***epimorphism*** *[i.e., ρ covers all of* $\prod\mathbb{Z}(m_i)$*, or, in other words, for any residues modulo* $m_1,\ldots, m_n$*, respectively, there exists* $k \in \mathbb{Z}$ that *has these residues)*[45] *if any pair of* m_i *is relatively prime.*

The foregoing factorization theorem states, in particular, that the multiplicative group of the ring $\mathbb{Z}(m_1 \ldots m_n)$ is isomorphic to a direct (or Cartesian) product of multiplicative groups of $\mathbb{Z}(m_i)$, $\mathbb{Z}(m_1 \ldots m_n)^* \cong \prod\mathbb{Z}(m_i)^*$, if any pair of m_i is relatively prime. This gives a proof (one of multiple known proofs) of Euler's totient function's multiplicativity property: $\varphi(m_1 \ldots m_n) = \prod\varphi\ (m_i)$ *if any pair of* m_i *is relatively prime*. Indeed, we have

$$\varphi(m_1 \ldots m_n) = \#\mathbf{Z}(m_1 \ldots m_n)^* = \prod \#\mathbf{Z}(m_i)^* = \varphi(m_1) \ldots \varphi(m_n).$$

(Work out the details.) Note that $\mathbb{Z}(p^v)^*$ (p is prime, and v is natural) is a cyclic group for $p > 2$ or $v = 1,2$, and otherwise it is a Cartesian product of cyclic groups $\mathbb{Z}[2^{v-2}]$ and $\mathbb{Z}[2]$ having orders (the total numbers of elements) 2^{v-2} and 2, respectively[46] [$v = 2$ refers to the intermediate situation, as $\mathbb{Z}(2^v)^* = \mathbb{Z}(4)^* \cong \mathbb{Z}[2] = \{1\} \times \mathbb{Z}[2] = \mathbb{Z}[1] \times \mathbb{Z}[2] = \mathbb{Z}[2^{v-2}] \times \mathbb{Z}[2]$]. Note that in any case the group $\mathbb{Z}(p^v)^*$ has the order $p^{v-1}(p-1)$ since $\varphi(p^v) = p^v - p^{v-1} = p^{v-1}(p-1)$ (as there are p^{v-1} multiples of p in $\{1,\ldots, p^v\}$).

The foregoing description of $\mathbb{Z}(p^v)^*$ generalizes the fact that readers are probably familiar with the following statement: $\mathbb{Z}(p)^*$ is a cyclic group for any prime number p because the ring $\mathbb{Z}(p)$ is actually a field, and *a finite multiplicative group in a field is always cyclic*. (Readers who had not encountered this fact previously are encouraged to work out the details; use the structure theorem for finitely generated Abelian groups and the fact that a polynomial equation of degree n has at most n roots in a field.) Readers can try to prove the foregoing description of $\mathbb{Z}(p^v)^*$ on their own or look at a proof in number theory text such as, for example, Venkov (1937), Hardy and Wright (1960), Borevich and Shafarevich (1972), and Ireland and Rosen (1982).

[45] So, the **sequence** $0 \to \bigcap \mathbb{Z}m_i \to \mathbb{Z} \to \prod\mathbb{Z}(m_i) \to 0$ is **exact** (in the language of Homological algebra).

[46] In Abelian group theory, a cyclic group of order μ is usually denoted $\mathbb{Z}[\mu]$.

E10.17

First establish the following lemma.

Lemma. (Arnol'd 1975). *For* the *polynomial* $\mathcal{P}(x_1,\ldots,x_N)$, the *multiplicative matrix seminorm* $\|.\|$, *and* the *matrices* $A_1,\ldots,A_N$, $\|\mathcal{P}(A_1,\ldots,A_N)\| \le |\mathcal{P}|(\|A_1\|,\ldots,\|A_N\|)$, *where* $|\mathcal{P}|$ *is a polynomial with the coefficients being absolute values of the coefficients of* $\mathcal{P}$ *of the same indices*.

This lemma follows from (α)–(γ) by induction on the complexity of the polynomial $\mathcal{P}$. With this lemma's help, establish the normal convergence of the exponential matrix series finding for it a majorant convergent numerical series.[47] In turn, the normal convergence of a matrix series with respect to the norm $\|A\| = \sum|a_{ij}|$ obviously implies (by virtue of statements in section E10.5) convergences of the series of the entries corresponding to any fixed indices (why?) and, hence (by definition!), convergence of the source matrix series itself.

The claims in **A** and **B** of section H10.5 are deduced for the normal convergence word for word with the absolute convergence in section E10.5.

E10.18

By the binomial formula, $S_nT_n - \sum_{k=0}^{n}(k!)^{-1}(A+B)^k = \sum_{1\le k,m\le n,\ n+m\le 2n}\frac{A^kB^m}{k!m!}$; thus, using the lemma in section E10.17 and proceeding by the same method as in section E10.6 we have for any multiplicative seminorm

$$\left\|S_nT_n - \sum_{k=0}^{n}(k!)^{-1}(A+B)^k\right\| \le (n!)^{-1}(\|A\|+\|B\|)^{2n+1} \to 0,$$

and, hence (according to section H10.17), $S_nT_n - \sum_{k=0}^{n}(k!)^{-1}(A+B)^k \to 0$. **QED**.

In connection with the continuous differentiability of the exponential map, recall that a map f: $U \to M$, where U is an open set in a finite-dimensional vector space L and M also is a vector space, **is differentiable** at a point $x_0 \in U$ if its **differential** $df(x_0)$ exists and is **continuously differentiable** at x_0 if it is differentiable throughout a neighborhood of this point and $df(x) \to df(x_0)$ as $x \to x_0$. The differential at x_0 is a linear operator $L \to M$ [usually denoted $df(x_0)$, df_{x_0}, or $d_{x_0}f$] whose graph is parallel

[47] A nonnegative-element numerical series $\sum a_n$ is a majorant for $\sum A_n$ when $\|A_n\| \le a_n$, $\forall n$.

to a tangent plane for the graph of $f(x)$ at the point $(x_0, f(x_0)) \in L \times M$ [so $df(x_0)$ exists if and only if this tangent plane exists].[48] In equivalent analytical terms,

$$f(x) - f(x_0) = df(x_0)\,(x - x_0) + o(|x - x_0|).$$

(Therefore, differentiable functions are continuous.) The **directional**, in the direction of increment h, **derivative** of function $f(x)$ at x_0 is defined as

$$\frac{df}{dh}(x_0) := \frac{d}{dt}\Big|_{t=0} f(x_0 + th) = \lim_{t\to 0} \frac{f(x_0 + th) - f(x_0)}{t}.$$

If $df(x_0)$ exists, then $\frac{df}{dh}(x_0)$ equals the derivative of a composition $t \mapsto f(x(t))$ at $t = t_0$ for any differentiable curve $t \mapsto x(t)$ emerging from the point $x(t_0) = x_0$ with the "initial velocity" $\dot{x}(t_0) = h$. (Why?)

Given smooth coordinates $u^1,\ldots,u^n$ that are local in a neighborhood of x_0, the **partial derivatives** $\partial f/\partial u^i$ are defined as derivatives of the compositions $u^i \mapsto f\left(x\left(u_0^1, \ldots, u_0^{i-1}, u^i, u_0^{i+1}, \ldots, u_0^n\right)\right)$ ($u^j = const = u_0^j$, $\forall j \neq i$), respectively. Therefore, partial derivatives of a differentiable function f with respect to the coordinates of a fixed coordinate system are directional derivatives, in the directions of unit vectors of the coordinate lines $\partial/\partial u^i$ [**vector fields** tangent to these lines and normed and directed so that $du^i(\partial/\partial u^i) = 1$]. *Partial derivatives with respect to any coordinates can be defined only with the complete coordinate system.*

As readers know, the directional derivative of a univariable function is simply its derivative, and the existence of the derivative at some point implies differentiability at that point. However, in the multivariable case, the existence of the directional derivatives in any number of directions cannot ensure differentiability or even continuity. (Provide examples.) However, using the Lagrange intermediate value theorem, readers can prove that

$\frac{df}{dh}(x)$ exists for all $x \in U$ and any h and is continuous in x and linear in h if and only if $f(x)$ is continuously differentiable in U (and in this case, $\frac{df}{dh}(x) = df(x)(h)$).[49]

Given smooth coordinates $v^1,\ldots v^m$ that are local in a neighborhood of $f(x_0)$ (in addition to $u^1,\ldots,u^n$ in a neighborhood of x_0), if the partial derivatives $\partial v^i/\partial u^j$ exist throughout a neighborhood of x_0 and are continuous at that point, then $df(x_0)$ will exist, defined as

[48] Advanced readers are probably familiar with a generalization of this notion for manifolds and for infinite-dimensional spaces and manifolds. The differential of a function on an open set in an infinite-dimensional vector space is often referred to as the **Frechét differential** (or the **Frechét derivative**); for the manifolds, the term **tangent map** is commonly used. In the infinite-dimensional case, the definition of $df(x_0)$ includes the continuity requirement (which is automatically fulfilled for the finite-dimensional case).

[49] This claim allows infinite-dimensional generalizations. See Hörmander (1983) for details and connected results.

$$\begin{pmatrix} u^1-u_0^1 \\ \vdots \\ u^n-u_0^n \end{pmatrix} \mapsto \begin{pmatrix} \partial v^1/\partial u^1\left(u_0^1,\ldots,u_0^n\right) & \cdots & \partial v^1/\partial u^n\left(u_0^1,\ldots,u_0^n\right) \\ \vdots & & \vdots \\ \partial v^m/\partial u^1\left(u_0^1,\ldots,u_0^n\right) & \cdots & \partial v^m/\partial u^n\left(u_0^1,\ldots,u_0^n\right) \end{pmatrix} \begin{pmatrix} u^1-u_0^1 \\ \vdots \\ u^n-u_0^n \end{pmatrix}^{50}.$$

Establish the existence of the directional derivative $\frac{de^X}{dH}(X)$ for any X and H and its continuity in X and linearity in H.[50] For this, verify that term-by-term derivatives of the exponential series form a series $\sum\limits_{n=1}^{\infty}(n!)^{-1}\sum\limits_{k=1}^{n}\underbrace{X\circ\ldots\circ X}_{k-1}\circ H\circ\underbrace{X\circ\ldots\circ X}_{n-k}$, which converges normally with respect to the norm $\|A\|=\sum|a_{ij}|$[51] (as having a majorant numerical series $e^{\|X\|}\cdot\|H\|=\|H\|\cdot\sum\limits_{n=0}^{\infty}(n!)^{-1}\|X\|^n$). Next, establish a generalization of the statement in section H10.5 (**C**) for the series of univariable vector/matrix functions (which can be done by the same method as in section E10.5) and apply it to prove the existence of $\frac{de^X}{dH}(X)$.[52] Lastly, prove the continuity of the directional derivatives by producing the following estimate using induction:

$$\|A^n-B^n\|\le n\cdot\|A-B\|\cdot\max\left(\|A\|,\|B\|\right)^{n-1};$$

derive from this that

$$\left\|\underbrace{A\circ\ldots\circ A}_{k-1}\circ H\circ\underbrace{A\circ\ldots\circ A}_{n-k}-\underbrace{B\circ\ldots\circ B}_{k-1}\circ H\circ\underbrace{B\circ\ldots\circ B}_{n-k}\right\|$$
$$\le(n-1)\cdot\|H\|\cdot\|A-B\|\cdot\max\left(\|A\|,\|B\|\right)^{n-2},$$

and, finally, prove that $\left\|\frac{de^X}{dH}(A)-\frac{de^X}{dH}(B)\right\|\le\|H\|\cdot\|A-B\|\cdot e^{\max(\|A\|,\|B\|)}$. Thus, we will prove the continuous differentiability of the exponential map and the equality $de^X(0)=id$ [that is, $de^X(0)(H)=H$, $\forall H$], and so satisfying the conditions of the implicit function theorem by the exponential map. (Work out the details.[53])

---

[50] Actually, the existence of $n-1$ columns of this $m\times n$ Jacobi matrix in a neighborhood of $(x_0,f(x_0))$, their continuity in x_0, plus at least the existence (without continuity) of the last column in $(x_0,f(x_0))$ will ensure the existence of $df(x_0)$.

[51] Therefore, with respect to any matrix norm (here and everywhere subsequently).

[52] Also, readers may establish a similar generalization for the series of multivariable vector/matrix functions; to do this, instead of the integration technique in section E10.5, a multivariable (infinite-variable) version of the intermediate value theorem should be applied. Interested readers will find a common scheme of such a proof in Dieudonné (1960).

[53] Also, readers may generalize Abel's lemma (discussed previously in section H10.5) for the sum of a matrix power series $\sum a_nX^n$ as a function of X (replacing absolute convergence with normal convergence) and establish the infinite differentiability of this function for $\|X\|<\|X_0\|$ when a series $\sum a_nX_0^n$ converges.

Next, to prove the formula $e^A e^B e^{-A} e^{-B} = E + [A,B] + o(\|A\|^2 + \|B\|^2)$, we suggest proceeding by the following method step by step.

A. Prove that the sum of normally convergent matrix series with respect to the norm $\|A\| := \sum|a_{ij}|$ will not change after permutations of arbitrary finite or infinite number of series terms. This behavior is similar to the behavior of an absolutely convergent numerical series.

B. Prove that if $\sum\limits_n A_{1,n}, \ldots, \sum\limits_n A_{k,n}$ are normally convergent matrix series with respect to the norm $\|A\| := \sum|a_{ij}|$, then their term-by-term product $\sum\limits_{n_1,\ldots,n_k} A_{1,n_1} \circ \ldots \circ A_{k,n_k}$ is normally convergent,[54] with the sum equal to the product of sums of the series cofactors.[55] This is similar to the well-known result for absolutely convergent numerical series.

C. Repeating the arguments from section E10.5 for proving Abel's lemma, $|a_n| \le C\,|t_0|^{-n}$ for a power series $\sum a_n t^n$ convergent for $t = t_0$. Prove that if this series starts from a term of index n_0, that is, $\sum a_n t^n = t^{n_0} \sum a_{n_0+n} t^n$, then a series quotient $\sum a_{n_0+n} t^n = \sum a_n t^n / t^{n_0}$ converges absolutely and uniformly on $|t| \le t_1 < |t_0|$. Prove similar claims for the normally convergent matrix series, with respect to norm $\|A\| := \sum|a_{ij}|$.

D. With the help of a direct series term-by-term multiplication and using statements in **A**–**C**, derive that

$$\begin{aligned} e^A e^B e^{-A} e^{-B} &= E + [A,B] \\ &+ A^3 \circ \mathcal{R}_1(A,B) + A^2 \circ B \circ \mathcal{R}_2(A,B) + A \circ B^2 \circ \mathcal{R}_3(A,B) \quad + A \circ B \circ A \circ \mathcal{R}_4(A,B) \\ &+ B \circ A \circ B \circ \mathcal{R}_5(A,B) + B \circ A^2 \circ \mathcal{R}_6(A,B) \quad + B^2 \circ A \circ \mathcal{R}_7(A,B) + B^3 \circ \mathcal{R}_8(A,B), \end{aligned}$$

where $\mathbb{R}_i(A,B)$ are sums of power matrix series in A, B that converge normally and uniformly on $\|A\|$, $\|B\| \le r$ for any fixed $r > 0$. Denoting $\|A\| = a$, $\|B\| = b$, the ratio of a norm v of the sum of the terms in the last two lines to the value of $a^2 + b^2$ may be estimated as

$$\frac{v}{a^2+b^2} \le const \cdot \frac{a^3 + a^2 b + ab^2 + b^3}{a^2 + b^2} = const \cdot (a+b) = o(1) \quad \text{as} \quad a, b \to 0$$

(work out the details). **QED**.

Finally, establish the commutativity of the squares and exponents of noncommuting matrices $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, $B = \begin{pmatrix} 1 & -\pi \\ \pi & 1 \end{pmatrix}$. [To do this, readers may use claims in sections P10.23**, P10.24**, and P10.25** (without proving them here) to find e^B.

[54] By virtue of **A**, the terms' order is unimportant.

[55] This claim may also be applied to a different method of proving the above equality $e^{A+B} = e^A e^B$ for commuting matrices (in both one- and multi-dimensional cases). Give such a proof.

E10.19

Method 2. Proof of lemma. Similitude commutes by raising to a power and, on the other hand, is a continuous operator on matrix spaces. (So what?) **QED**.

Next, for a Jordan box $A = \lambda E + N = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}$ of dimension m we have

$$(E + A/n)^n = ((1 + \lambda/n)E + N/n)^n = \sum_{k=0}^{m-1} \binom{n}{k} (1 + \lambda/n)^{n-k} \left(\frac{N}{n}\right)^k$$

$$= (1 + \lambda/n)^n \sum_{k=0}^{m-1} \frac{n(n-1)\dots(n-k+1)}{k! n^k (1 + \lambda/n)^k}$$

$$\times N^k \to \begin{pmatrix} e^\lambda & e^\lambda & \cdots & e^\lambda/(m-1)! \\ & \ddots & \ddots & \vdots \\ & & \ddots & e^\lambda \\ & & & e^\lambda \end{pmatrix} = e^A \quad \text{as} \quad n \to \infty.$$

(Work out the details.)

E10.21

Proof of the proposition. (*ii*) ⇒ (*i*). Analyticity of elements of F. By induction, those elements are infinitely differentiable. Next, for $f \in F$ the functions $f, \dots, f^{[n]}$, where $n = \dim F$, cannot be linearly independent, and so f satisfies a **linear autonomous ODE** $f^{[m]} = \sum_{i=0}^{m-1} a_i f^{[i]}$ for some $0 \le m \le n$. Further differentiation yields $f^{[m+k]} = \sum_{i=0}^{m-1} a_i f^{[i+k]}$ $(k = 0,1,\dots)$. Using this equality, produce by induction the estimate $|f^{[k]}(0)| \le r^k$, with r large enough so that $\sum_{i=0}^{m-1} |a_i| r^i \le r^m$. Derive that the Taylor-series expansion of f at the origin converges absolutely and uniformly on any compact subset of $\mathbb{R}$ [as being majorized by the convergent positive-element

numerical series $e^{tr} = \sum(n!)^{-1}(tr)^n]$. An analytic function $\bar{f}(t) := \sum (n!)^{-1} f^{[n]}(0)\ t^n$ is infinitely differentiable and satisfies the same ODE as f and the initial conditions for $t = 0$: $f^{[i]}(0) = \bar{f}^{[i]}(0)$ $(i = 0,1,\ldots)$; therefore, by the **uniqueness theorem**, $f = \bar{f}$, which completes the proof.

Also, the equality $f(t) = \sum(n!)^{-1} f^{[n]}(0) t^n$ follows from the slow growth of the derivatives since $f^{[n]}(t) \cdot r_0{}^n/n! = o\ (1)$ as $n \to \infty$ uniformly on $| t - t_0 | \leq r_0$, $\forall r_0 > 0$ (as explained subsequently in the footnote to section P10.24**).

Proof of Proposition 2. *(ii)* ⇒ *(iii)*. $\lambda_j = 0$, because otherwise the corresponding elements of F would show exponential growth as $t \to \infty$ or $t \to -\infty$, which would contradict periodicity and continuity. This is quite evident when $\omega_j = 0$ and for $\omega_j \neq 0$ can be proved by various methods – for example, using an elementary identity $A \cos\theta + B \sin\theta = C \cos(\theta - \theta_0)$ $\left(C = \sqrt{A^2 + B^2},\ \cot\theta_0 = \frac{A}{B}\right)$, where, in our case, A and B are polynomial in t, $A^2 + B^2$ is a nonzero polynomial, and $\theta = \omega t$ with $\omega \neq 0$. θ_0 and C are functions of t taking values in $\mathbb{R}$ mod π and $[0,\infty)$, respectively, and having limits as $t \to \infty$, and lim C is either a positive number or infinity. (Why?) Therefore, for infinitely many t, such that $t \to \infty$, $C \cos(\theta - \theta_0)$ is bounded away from zero, $C \cos(\theta - \theta_0) \geq \varepsilon > 0$, and so, for these t, $C \cos(\theta - \theta_0)$ $e^{\lambda t}$ with $\lambda > 0$ grows like the exponential function $e^{\lambda t}$ (more exactly, not slower than $\varepsilon e^{\lambda t}$). We leave it to the reader to work out the details.

The same restriction (periodicity + continuity) also excludes polynomial growth, so all $N_j = 0$. Thus, $F = \left\{ \sum_{j \in J_F} a_j \cos\omega_j t + b_j \sin\omega_j t\ :\ a_j, b_j \in \mathbb{R} \right\}$, corresponding to a finite set J_F. Particularly, F contains sin $\omega_j t$. Verify that sin ωt is 2π-periodic if and only if $\omega \in \mathbb{Z}$. [For this use, for example, the argument-summation formula sin $(s + t) =$ sin s cos $t +$ sin t cos s.] This completes the proof.

E10.25

Definitions of cosine and sine via their Taylor-series expansions. We must show that the functions $C(t)$, $S(t)$, defined as the sums of series that converge absolutely and uniformly on any compact set

$$C = \sum_{n=0}^{\infty} \frac{(-1)^n t^{2n}}{(2n)!}, \qquad S = \sum_{n=0}^{\infty} \frac{(-1)^n t^{2n+1}}{(2n+1)!},$$

must be only cos t and sin t, respectively. Readers may proceed as follows. The functions $C(t)$, $S(t)$ are infinitely differentiable, and their derivatives may be obtained through term-by-term differentiation of the series (as discussed previously

in section H10.5). Thus, a direct verification shows that $\dot{C} = -S, \dot{S} = C$, and so these functions are solutions of the **Cauchy initial value problems** for the differential equation $\ddot{x} = -x$, corresponding to initial data $x(0) = 1$, $\dot{x}(0) = 0$ and $x(0) = 0$, $\dot{x}(0) = 1$, respectively. Because of the linearity of the equation, the solutions are extendable to the whole axis $-\infty < t < \infty$. Since cos t, sin t are solutions of the same Cauchy problems, we have $C(t) = \cos t$ and $S(t) = \sin t$, by virtue of their unique solvability. **QED**. (We leave it to the reader to fill in the details.)

Also, readers can arrive at the equalities $C(t) = \cos t$, $S(t) = \sin t$ without reference to the Cauchy problem but verifying directly that $C(t)$, $S(t)$ are, respectively, the abscissa and ordinate of the unit vector of a polar angle t. Consider the vector function $t \mapsto V(t) := (C(t), S(t))$. This function is defined by the foregoing Taylor series on the whole axis $-\infty < t < \infty$. Show that $|V(t)|^2 = C^2(t) + S^2(t) = 1$ [which is obtainable by a direct calculation using the defining series; alternatively, the reader can prove it using the fact that $\ddot{x} = -x$ is the **oscillation equation**, so its solutions obey the **energy conservation law**, $x^2 + \dot{x}^2 = const$; this *constant* equals 1 as follows from the defining series for $C(t)$ and $S(t)$]. For the derivative we will find $\dot{V}(t) = (-S(t), C(t))$, so $\dot{V}(t)$ is obtained from $V(t)$ by a 90° counterclockwise rotation. This means that $V(t)$ is a uniform rotation along the unit circle by angle t, starting from the initial position $V(0) = (1,0)$, which completes the proof. (Work out the details.)

E10.26. A+B

First obtain the identities $S(h) = 2 \cdot S(h/2) \cdot C(h/2)$ and $1 - C(h) = 2S^2(h/2)$. If we use these equalities and define $h = s - t$, then we obtain

$$\begin{aligned} C(t+h) - C(t) &= C(t) \cdot (C(h) - 1) - S(t) \cdot S(h) \\ &= -2S(h/2) \cdot [C(t) \cdot S(h/2) + S(t) \cdot C(h/2)] \\ &= -2 \cdot S(h/2) \cdot S(t + h/2), \end{aligned}$$

which is negative for $h/2$, $t + h/2 \in (0, \pi/2)$. **QED**.

Alternatively, readers may accomplish step **A** as follows. From

$$C(t) = C(0) \cdot C(t) + S(0) \cdot S(t) = C(0) \cdot C(t),$$

we have $C(0) = 1$. Then carry out step **E**. Next, using the oddness of $S(t)$, deduce the identity $C(s - t) = C(s) \cdot C(t) + S(s) \cdot S(t)$. Finally, calculate $C(s - t) - C(s + t)$.

C. Apply induction on n. For $n = 1$ the statement is true. Show with an elementary calculation that the system of quadratic equations

$$2CS = \sin\,\alpha, \qquad C^2 - S^2 = \cos\,\alpha$$

has two solutions: $(C, S) = \pm\,(\cos(\alpha/2), \sin(\alpha/2))$. In our case we must take the first of them ("+") since $S > 0$ and $\sin\,(\alpha/2) > 0$.

D. By virtue of **C** and monotonicity (**B**), we have $S(t) \to 0$ and $C(t) \to 1$ as $t \to 0$. (Work out the details.) Using this, derive continuity from the argument's addition formulas.

E. For $X = C(-t)$ and $Y = S(-t)$ prove that a system of linear equations

$$C(t) \cdot X - S(t) \cdot Y = 1, \quad C(t) \cdot X + S(t) \cdot Y = 0$$

has the unique solution $(X, Y) = (C(t), -\,S(t))$.

E10.27

(2) By the spectral theorem, an orientation-preserving orthogonal operator B is brought by an orthogonal transformation to a block-diagonal form that has two-dimensional blocks as R_θ and, optionally (when dim B is odd), a single one-dimensional block $E_1 = 1$:

$${}^tG \circ B \circ G = \begin{pmatrix} R_{\theta_1} & & & \\ & \ddots & & \\ & & R_{\theta_k} & \\ & & & 1 \end{pmatrix} \quad ({}^tG = G^{-1}).$$

To determine skew-symmetric logarithms of A, preserve the skew symmetry by orthogonal transformations, ${}^tA = -A \Rightarrow {}^t({}^tG \circ A \circ G) = -{}^tG \circ B \circ G$, and the permutability of the similitude transformation with the exponential function (lemma in section H10.19 above). (Work out the details.)

(3) Proof of lemma. Construct by induction a basis in which matrix

$$e^{\begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}} = \begin{pmatrix} e^\lambda & e^\lambda & \cdots & e^\lambda/(m-1)! \\ & \ddots & \ddots & \vdots \\ & & \ddots & e^\lambda \\ & & & e^\lambda \end{pmatrix}$$

has the form $\begin{pmatrix} e^{\lambda} & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & e^{\lambda} \end{pmatrix}$.

Proof that R_θ commutes only with matrices kR_φ when $\theta \neq 0,\pi$.
The matrix equation

$$\begin{pmatrix} c & -s \\ s & c \end{pmatrix} \circ \begin{pmatrix} x & y \\ z & t \end{pmatrix} \circ \begin{pmatrix} c & s \\ -s & c \end{pmatrix} = \begin{pmatrix} x & y \\ z & t \end{pmatrix} \quad (c^2 + s^2 = 1)$$

is equivalent to the system of two linear equations on unknowns $x - t$, $y + z$:

$$s^2(x - \mathrm{t}) + cs \cdot (y + z) = 0, \quad -cs \cdot (x - \mathrm{t}) + s^2 \cdot (y + z) = 0,$$

which has the determinant $D = s^4 + s^2c^2 = s^2$. For $s = 0$, the unknowns may take arbitrary values; for $s \neq 0$, the solution is $x = t$, $y = -z$. **QED**. (We leave it to the reader to fill in the details.)

(4) We will establish the existence of the logarithm with the help of the spectral theorem if we verify that closeness to the identity and closeness to zero are uniformly preserved by the similitude transformations using orthogonal matrices. (Why?) In terms of a fixed matrix norm ||.|| (see section H10.17 above) this means existence of $K > 0$ such that for any orthogonal operator G and $\varepsilon > 0$

$$\|X\| \leq \varepsilon \Rightarrow \|{}^tG \circ X \circ G\| \leq K\varepsilon, \qquad \|X - E\| \leq \varepsilon \Rightarrow \|{}^tG \circ X \circ G - E\| \leq K\varepsilon.$$

Verify that for a multiplicative norm the preceding statement is true for $K = \sup \|G\|^2$, where the least upper bound (sup) is considered over all orthogonal matrices G. Specifically, we may take $K = 1$ for $\|A\|_1 = \sup_{x \neq 0} |Ax|/|x| = \sup_{|x|=1} |Ax|$ (why?) and $K = n^3$ (the cubed dimension of the space) for $\|A\|_2 = \sum |a_{ij}|$. (Readers may derive the inequality $\|G\|_2 \leq n^{3/2}$ for orthogonal matrices G from the inequality between arithmetic and quadratic means discussed previously in the "Convexity and Related Classical Inequalities" problem group.) **QED**.

(5) The Jacobi matrix corresponding to the left-hand side of this system is the composition of nondegenerate matrices: the $n^2 \times n^2$ matrix $\partial[\log(B^{-1} \circ X)]/\partial X$ and some $\binom{n+1}{2} \times n^2$ matrix. (Work out the details.)

Similar arguments show that a diffeomorphism between manifolds maps submanifolds onto submanifolds.

(7) For matrix representations of linear operators and the bilinear form $\langle\rangle$ with respect to any fixed basis, show that the desired expression for the matrix A^* via the

matrix A and the Gram matrix G is $A^* = G^{-1} \circ {}^tA \circ G$ (using for the definition the dual identity $\langle x, Ay\rangle = \langle A^* x, y\rangle$ yields $A^* = {}^tG^{-1} \circ {}^tA \circ {}^tG$) and that a linear change of variables determined by matrix C substitutes matrices ς and ς^* by the same law $A \mapsto C \circ A \circ C^{-1}$, $A^* \mapsto C \circ A^* \circ C^{-1}$.)

Using this, identity $A^{**} = A$, $\forall A$ is equivalent to that $A = (G^{-1} \circ {}^tG) \circ A \circ (G^{-1} \circ {}^tG)^{-1}$, $\forall A$, which means that $G^{-1} \circ {}^tG$ commutes with any matrix and, thus, is a dilatation: $G^{-1} \circ {}^tG = \lambda E$. (Work out the details: show that being a dilatation follows from commuting with n^2 basis matrices $A^{km} = (\delta_{ij}^{km})_{i,j=1,\ldots,n}$ $(k, m = 1,\ldots, n)$; here $\delta_{ij}^{km} = \begin{cases} 1 & \text{if } k = i \text{ and } m = j \\ 0 & \text{otherwise} \end{cases}$).

From this, ${}^tG = \lambda G$, and hence, $G = \lambda^2 G$, so $\lambda = 1$ or $\lambda = -1$, which completes the proof.

(9) Proof of inclusion $\{A\colon e^{tA} \in G, \forall t \in \mathbb{R}\} \subseteq \mathrm{g}$. Denoting $B_t = e^{tA}$ we will have, according to **(7)**, $(B_t)^* = e^{tA^*}$ and $(B_{-t})^* = \left(B_t^{-1}\right)^* = (B_{-t})^*$ (thus, we may apply these symbols as B_{-t}^*, without brackets). We have $B_t \circ B_t^* \circ B_{-t} \circ B_{-t}^* = E$, which shows the commutativity of A and A^*. Therefore, $E = e^{tA} \circ e^{tA^*} = e^{t(A+A^*)}$, thus, for any small t we have $t(A + A^*) = 0$, and so (using $t \neq 0$) $A + A^* = 0$. **QED**. (We leave it to the reader to fill in the details.)

(11) First prove that a map $\{\mathbb{R}^2 \times \mathbb{C}^* \cong G \times G_3 \times G_4 \to GL^+(2,\mathbb{R}) = \{A : \det A > 0\}(g, g_3, r) \mapsto g \circ g_3 \circ r$., where G_4 is the group of dilatation by positive scalar factors ($G_4 = \mathbb{R}^{+*}E$), is a diffeomorphism of smooth manifolds. For this, show that the same matrix equation as in section H10.27 **(11)** with a, b, c, d, such that $ad - bc > 0$, is uniquely solvable for $x > 0$ and has the solution

$$x = r\Big/\sqrt{c^2 + d^2}, \quad y = (ac + bd)\Big/r\sqrt{c^2 + d^2}, \quad u = rd\Big/\sqrt{c^2 + d^2}, \quad v = rc\Big/\sqrt{c^2 + d^2},$$

where $r = (ad - bc)^{1/2}$ (x, y and u, v are homogeneous functions of a, b, c, d of degrees, respectively, zero and one. This means that multiplication of a, b, c, d by a common factor multiplies g_3 by this factor but does not affect g.) Thus, between two four-dimensional smooth manifolds, $\mathbb{R}^{+*} \times \mathbb{R} \times \mathbb{C}^{+*}$ and $GL^+(2,\mathbb{R})$ (the second one is a manifold, being an open subset of a vector space of matrices), with global coordinates (x,y,u,v) and (a,b,c,d), respectively, there is one-to-one correspondence, which is bidifferentiable at all points (work out the details using that $c^2 + d^2 \neq 0$), and, thus, is a diffeomorphism. **QED**.

Second, show that $SL(2,\mathbb{R})$ is a smooth submanifold of $GL^+(2,\mathbb{R})$ (of codimension one). For this, consider the smooth map $GL^+(2,\mathbb{R}) \xrightarrow{\det} \mathbb{R}^{+*}$, which is a group homomorphism as well, having $SL(2,\mathbb{R})$ as its kernel. Verify that det does

not degenerate (has a rank of one) at any point; derive from it, with the help of the implicit function theorem, that $SL(2,\mathbb{R})$, being the preimage of the single point (that is, of a smooth submanifold of zero dimension), itself is a smooth submanifold of codimension equal to 1:

$$\dim GL^+(2,\mathbb{R}) - \dim SL(2,\mathbb{R}) = \dim \mathbb{R}^{+*} - \dim\{1\} = 1 - 0 = 1.$$

This statement about the preimages of points corresponds to the so-called **transversality theorems**. Readers will find an extensive discussion in Arnol'd (1978), Arnol'd et al. (1982), and references therein.

Finally, the restriction of the preceding diffeomorphism $G \times G_3 \times G_4 \to GL^+(2,\mathbb{R})$ to the submanifold $G \times G_3 \times \{E\}$ obviously coincides with the map $\begin{cases} G \times G_3 & \to & SL(2,\mathbb{R}) \\ (g, g_3) & \mapsto & g \circ g_3 \end{cases}$; hence, the last one is a diffeomorphism, which completes the proof. (Work out the details.)

On the other hand, groups $SL(2,\mathbb{R})$ and $\mathbf{S}^1 \times \mathbb{R}^2$ cannot be algebraically isomorphic because the second group is Abelian, whereas the first is not.

(12) For skew-symmetric matrix A of dimension n we have

$$\det A = \det {}^t A = \det(-A) = (-1)^n \det A$$

(why?); hence, $\det A = 0$ for odd n.

Next, for a symplectic form [,] on $\mathbb{R}^{2n}$ and a one-dimensional subspace $Q_1 \subset \mathbb{R}^{2n}$, let M be a hyperplane that is its skew-orthogonal complement, $M = Q_1^{\perp\!\!\!\perp}$. (We have $Q_1 \subseteq M$; why?) Let P_1 be a one-dimensional subspace transversal to M (transversality means that $M + P_1 = \mathbb{R}^{2n}$). Since $P_1 \not\subset Q_1^{\perp\!\!\!\perp}$, [,] is nondegenerate on the two-dimensional subspace $P_1 \bigoplus Q_1$. Now, consider a $2n-2$-dimensional subspace $N = (P_1 \oplus Q_1)^{\perp\!\!\!\perp}$. We will find

$$N \cap (P_1 \oplus Q_1) \subseteq M \cap (P_1 \oplus Q_1) = Q_1$$

(why?), so a subspace on the left-hand side is equal to either zero space or Q_1. (Why?) The second option cannot be realized because it would imply $Q_1 \subseteq (P_1 \oplus Q_1)^{\perp\!\!\!\perp}$ and, hence, the skew orthogonality of Q_1 with P_1 while $P_1 \not\subset Q_1^{\perp\!\!\!\perp}$ (work out the details). Thus, $(P_1 \oplus Q_1)^{\perp\!\!\!\perp}$ has zero intersection with $P_1 \bigoplus Q_1$. (Make a diagram of all considered inclusions and intersections.) Therefore, [,] is a symplectic form on $(P_1 \oplus Q_1)^{\perp\!\!\!\perp}$, so the proof may be completed by induction on n. (Work out the details.)

Next, an operator $I = (-E)^{1/2}$ could have its Jordan boxes either as $\pm\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ or as $\begin{pmatrix} \pm\boxed{\begin{matrix} 0 & -1 \\ 1 & 0 \end{matrix}} & 1 & & \\ & & 1 & \\ & & \pm\boxed{\begin{matrix} 0 & -1 \\ 1 & 0 \end{matrix}} & \ddots \\ & & & \ddots \end{pmatrix}$ [$(-E)^{1/2}$ exist only for even dimensions], but the second option is topologically wrong (bringing $|I^n v| \to \infty$ as $n \to \infty$ and some $v \in \mathbb{R}^{2n}$, as discussed previously in section E7.2 from the "2 × 2 Matrices That Are Roots of Unity" problem group). Therefore, I can be represented by the matrix $\begin{pmatrix} \boxed{\begin{matrix} 0 & -1 \\ 1 & 0 \end{matrix}} & & \\ & \ddots & \\ & & \boxed{\begin{matrix} 0 & -1 \\ 1 & 0 \end{matrix}} \end{pmatrix}$ with respect to some basis.

The same statement can be proved also by induction as follows. Let L^{2k} be an I-invariant subspace of dimension $2k$ with $0 \le k < n$ and $v \notin L$. A two-dimensional plane Π spanned on v and Iv is I-invariant, and so its intersection with L is I-invariant; hence, it cannot be of dimension one, so $\Pi \cap L = \{0\}$ (why?), and the proof may be completed by induction.

Finally, let $I = (-E)^{1/2}$ have the same matrix as G_D with respect to some Darboux basis. Then, agreement of [,] with the complex structure is expressed by the equation ${}^t G_D \circ G_D \circ G_D = G_D$, which is satisfied as being equivalent to a complex-number equality $(-i)\cdot i \cdot i = i$. (Work out the details.) Thus, the agreement holds. **QED**.

For a null space L we have $L \subseteq L^\perp$, so $\dim L \le \dim L^\perp = \dim N - \dim L$ if the form is nondegenerate. Also, readers can establish the inequality $2 \dim L \le \dim N$ by applying the following lemma to the orthogonal complements to one-dimensional subspaces spanned on the elements of any fixed basis of L.

Lemma. *If a bilinear form is nondegenerate, then for linearly independent vectors* $v_1,\ldots,v_k$ *the hyperplanes* $H_i = (\mathbb{R}v_i)^\perp$ *are* ***in a general position*** *(which means that* $\operatorname{codim} \bigcap_j H_{i_j} = \sum_j \operatorname{codim} H_{i_j} = \sum_j 1$ *for any of those intersections).*

(Indeed, we have

$$\sum_j 1 = \dim \sum_j \mathbb{R}v_{i_j} = \operatorname{codim}\left(\sum_j \mathbb{R}v_{i_j}\right)^\perp \le \operatorname{codim} \bigcap_j H_{i_j} \le \sum_j \operatorname{codim} H_{i_j} = \sum_j 1.$$

QED. Work out the details.) Lastly, readers may establish the inequality $2 \dim L \le \dim N$ proceeding by a combinatorial method, for instance, by calculating the determinant of the Gram matrix via the **Laplace decomposition** (discussed previously in section E8.8 from the "A Property of Orthogonal Matrices" problem group) with respect to the rows (or columns) corresponding to L.

Proof of theorem from section H10.27 (**12**). Necessity. Under agreement of symplectic and complex structures, I maps Lagrangian planes onto transversal ones as these planes are orthogonal with respect to a positive-definite form $(,) = [I.,.]$: $(L, I(L)) = [I(L), I(L)] = [L, L] = 0$.

Sufficiency. Let the condition of the theorem be satisfied, L a Lagrangian plane, and Q_1 a one-dimensional subspace of L and $P_1 = I(Q_1)$. P_1 cannot be skew-orthogonal with Q_1 because otherwise $P_1 + Q_1$ would be contained in a Lagrangian plane [by the proposition from section H10.27 (12)], which is not transversal with its I-image. (Work out the details.) Therefore, $I(L) \not\subset Q_1^{\perp}$; thus, an intersection $M = I(L) \cap Q_1^{\perp}$ is a hyperplane in $I(L)$. (Why?) Next, we have $I(M) \subset L$ (create a diagram of all considered inclusions and intersections), so I (M) is skew-orthogonal to Q_1 (since L is a null space), and thus $M \bigoplus I(M)$ is skew-orthogonal to Q_1. Also, M is skew-orthogonal to P_1 (since $M + P_1$ is a null space). In addition, $Q_1 + M$ is a null space, and, by the conditions of the theorem, $I(Q_1 + M) = I(Q_1) + I(M)$ is a null space, so I (M) is skew-orthogonal to P_1. Finally, $Q_1 \bigoplus I(Q_1)$ and $M \bigoplus I(M)$ are skew-orthogonal complements of each other. Since $\mathbb{R}^{2n}$ is a sum of these subspaces, they have zero intersection (why?); hence, on each of them the symplectic form is nondegenerate. Therefore, we may complete this proof by induction, having established the theorem for $n = 1$. [Work out the details using **(11)** and the equality $\det I = 1$.] **QED**.

The operator $I = (-E)^{1/2}$ cannot have the same matrices with respect to any Darboux bases because the elements of such a basis may be nonorthogonal and have distinct lengths (with respect to the positive definite form $[I.,.]$). We suggest that more advanced readers carry out the following calculation for $n = 1$: the set of Darboux bases is parameterized by the elements of $SP(2)$ (how?); in turn, the set of bases with respect to which I has the matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is parameterized, according to section H10.27 **(3)** above, by the elements of $\mathbb{R}^{+*} \times SO(2) = \left\{ \begin{pmatrix} x & -y \\ y & x \end{pmatrix} : x^2 + y^2 \neq 0 \right\} \cong \mathbb{C}^*$ (how?); but these groups (considered as smooth submanifolds in the matrix space) cannot be diffeomorphic (even homeomorphic) with each other because $\dim \mathbb{C}^* < \dim SP(2)$. [Work out the details using **(11)**.]

For $n > 1$, mapping some Lagrangian plane onto a transversal one by $I = (-E)^{1/2}$ does not imply agreement of the complex and symplectic structures. Readers are encouraged to find counterexamples on their own.

(20) Preserving orthogonality implies an affine property since a subspace of codimension k that passes through a fixed point x_0 is a set of the points x such that vectors $x - x_0$ are orthogonal to a set of fixed mutually orthogonal other vectors $e_1,\ldots,e_k$ (work out the details.) Circles are mapped onto circles and diametrically opposite points are mapped onto diametrically opposite ones because planes are mapped onto planes and a circle is a set of the points in a plane from which any fixed diameter is observed at a right angle. The centers are mapped onto the centers as they are intersections of the diameters. (Work out the details.)

Proof of lemma. Using well-known elementary-geometric arguments, show that preserving the lengths of vectors implies preserving the angles between vectors. In turn, preserving the lengths and angles implies preserving the scalar product

$$\frac{\langle Tx, Ty\rangle}{|Tx| \cdot |Ty|} = \frac{\langle x, y\rangle}{|x| \cdot |y|}, \quad |Tx| = |x|, \quad |Ty| = |y| \Rightarrow \langle Tx, Ty\rangle = \langle x, y\rangle.$$

This implies the additive property of T:

$$\begin{aligned} \langle T(x+y) - Tx - Ty, Tz\rangle &= \langle T(x+y), Tz\rangle - \langle Tx, Tz\rangle - \langle Ty, Tz\rangle \\ &= \langle x+y, z\rangle - \langle x, z\rangle - \langle y, z\rangle = 0. \end{aligned}$$

Since the scalar product is a nondegenerate form, $\forall x,y,z$, we have $T(x + y) = Tx + Ty$. In turn, the additivity and preservation of the origin imply $\mathbb{Z}$-linearity, $T(nx) = nTx$ ($n \in \mathbb{Z}$), which is equivalent to $\mathbb{Q}$-linearity: $Tx = T(nx/n) = nT(x/n)$. (Work out the details.) Finally, this yields, by virtue of the continuity of T (why is it continuous?) and the density of $\mathbb{Q}$ everywhere in $\mathbb{R}$, the desired $\mathbb{R}$-linearity. **QED**. (We leave it to the reader to fill in the details.)

E10.28

Method 1 (Arnol'd 1975). By virtue of continuity of the determinant we have

$$\det e^A = \det \lim_{n\to\infty} (E + A/n)^n = \lim_{n\to\infty} \det (E + A/n)^n$$
$$= \lim_{n\to\infty} \left(1 + \operatorname{tr} A/n + O(n^{-2})\right)^n.$$

Finally, calculating the limit on the right-hand side in the same way as in section H10.24 yields $e^{\operatorname{tr} A}$. **QED**. (We leave it to the reader to fill in the details.)

Method 2. We have, taking into account the lemma in section H10.27,

$$\det e^A = \prod_{\lambda \in Spec(A)} e^\lambda = e^{\sum_{\lambda \in Spec(A)} \lambda} = e^{\operatorname{tr} A}.$$

QED. (We leave it to the reader to fill in the details.)

Method 3. We have for a Jordan box, using the explicit formula from section H10.18,

$$\det e^{\lambda E+N} = \left(e^\lambda\right)^m = e^{\lambda m} = e^{\operatorname{tr}(\lambda E+N)}.$$

QED. (We leave it to the reader to fill in the details.)

Method 4. Applying the formula (*) from section P10.22** with $W(t) = e^{At}$ and taking into account the differential equation for the exponential function from section P10.15** yields the differential equation $\dot{D} = D \cdot \operatorname{tr} A$ for $D(t) = \det e^{At}$. The obvious initial condition $D(0) = 1$ implies that the solution is $D = e^{t\cdot \operatorname{tr} A}$ and $D(1) = e^{\operatorname{tr} A}$. **QED**. (We leave it to the reader to fill in the details.)

E10.29

Existence of logarithms in space of matrices with complex entries. Since Jordan boxes corresponding to different root subspaces commute, it is enough to consider only one Jordan box. (Why? Answer using the results of section P10.18**.) For the

Jordan box $\begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}$ of dimension m determine by induction a basis with respect to which it has the matrix $\begin{pmatrix} \lambda & \lambda & \cdots & \lambda/(m-1)! \\ & \ddots & \ddots & \vdots \\ & & \ddots & \lambda \\ & & & \lambda \end{pmatrix} = e^{\begin{pmatrix} \mu & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \mu \end{pmatrix}}$ ($\mu = \ln\lambda = \ln|\lambda| + i\arg\lambda$). **QED**. Also, these arguments show the existence of a logarithm in the space of matrices with real entries when $\lambda > 0$. (Work out the details.)

The existence of logarithms implies the existence of square roots (and any rational powers) in the space of matrices (over any ring containing the rational field). Indeed, $\exists\ \ln B \Rightarrow (e^{\ln B/q})^q = \underbrace{e^{\ln B/q} \circ \ldots \circ e^{\ln B/q}}_{q} = e^{\ln B} = B$. (Work out the details.)

The existence of square roots in the space of matrices with real entries implies the condition in section P10.29**. If $\lambda < 0$ is an eigenvalue of a source matrix, then a square root has eigenvalues $\pm i\sqrt{-\lambda}$, so the corresponding boxes of the Jordan canonical form over the reals have the form $\sqrt{-\lambda} \cdot \begin{pmatrix} \boxed{\begin{matrix} -1 & \\ 1 & \end{matrix}} & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \end{pmatrix}$, which has the matrix $-\lambda \cdot \begin{pmatrix} \boxed{\begin{matrix} -1 & \\ & -1 \end{matrix}} & \boxed{\begin{matrix} -2 & \\ 2 & \end{matrix}} & 1 & \\ & \ddots & \ddots & \ddots \\ & & & \end{pmatrix}$ as its square. (Work out the details.) Convert the latter matrix to a matrix that consists of two equal diagonal blocks $\begin{pmatrix} \lambda & 1 & \\ & \ddots & \ddots \\ & & \end{pmatrix}$ of half-matrix dimension each by an appropriate change in the basis, which will complete the proof.

The condition in section P10.29** implies the existence of logarithms in the space of matrices with real entries.

(a) A logarithm of a matrix consisting of equal Jordan boxes corresponding to $\lambda < 0$ is found in a few successive steps as follows. First, move to a basis with respect to which the matrix has the form $|\lambda| \cdot \begin{pmatrix} \boxed{\begin{matrix} -1 & \\ & -1 \end{matrix}} & 1 & \\ & \ddots & \ddots \\ & & \end{pmatrix}$. Second,

find a basis providing the form $|\lambda| \cdot \begin{pmatrix} \boxed{\begin{matrix} -1 & \\ & -1 \end{matrix}} & \boxed{\begin{matrix} -1 & \\ & -1 \end{matrix}} & \boxed{\begin{matrix} -1/2! & \\ & -1/2! \end{matrix}} & \boxed{\begin{matrix} -1/3! & \\ & -1/3! \end{matrix}} & \cdots \\ & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$. Finally, verify that the last matrix has a logarithm $\ln|\lambda| \cdot E + \begin{pmatrix} \boxed{\begin{matrix} -\pi & \\ \pi & \end{matrix}} & 1 & \\ & \ddots & \ddots \end{pmatrix}$.

(b) A logarithm of a block matrix consisting of a pair of equal Jordan boxes corresponding to $\lambda = \rho e^{\pm i\theta}$ is produced by a generalization of the previous construction (which corresponds to the special case $\theta = \pi$). First, move to the form $\rho \cdot \begin{pmatrix} \boxed{\begin{matrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{matrix}} & 1 & \\ & \ddots & \ddots \end{pmatrix}$, then to the form $\rho \cdot \begin{pmatrix} \boxed{\begin{matrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{matrix}} & \boxed{\begin{matrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{matrix}} & \boxed{\begin{matrix} \cos\theta/2! & -\sin\theta/2! \\ \sin\theta/2! & \cos\theta/2! \end{matrix}} & \boxed{\begin{matrix} \cos\theta/3! & -\sin\theta/3! \\ \sin\theta/3! & \cos\theta/3! \end{matrix}} & \cdots \\ & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$ having $\ln\rho \cdot E + \begin{pmatrix} \boxed{\begin{matrix} -\theta & \\ \theta & \end{matrix}} & 1 & \\ & \ddots & \ddots \end{pmatrix}$ as one of its logarithms.

The existence of $\sqrt[p]{B}$ in $GL(n,\mathbb{C})$ and, for $B \in GL(n,\mathbb{R})$ which satisfies the condition from section P10.29**, in $GL(n,\mathbb{R})$ follows from the existence of the logarithms. Also, the existence of $\sqrt[p]{B}$ in $GL(n,\mathbb{C})$ for all B and p and, for $B \in GL(n,\mathbb{R})$ and odd p, in $GL(n,\mathbb{R})$ can be established with the help of the following calculation: for the Jordan box $\begin{pmatrix} \lambda & 1 & \\ & \ddots & \ddots \end{pmatrix}$ of dimension m determine the basis with respect to which it has the matrix $\begin{pmatrix} \mu^p & \mu^{p-1} & \cdots & \binom{p}{m-1}\mu^{p-m+1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \mu^{p-1} \\ & & & \mu^p \end{pmatrix}$ $(\mu = \lambda^{1/p})$, and verify its equality to $\begin{pmatrix} \mu & 1 & \\ & \ddots & \ddots \end{pmatrix}^p$.

Next, the conditions of the existence of roots of the matrices with zero eigenvalues can be obtained by raising the Jordan box $\begin{pmatrix} 0 & 1 & \\ & \ddots & \ddots \\ & & \ddots \end{pmatrix}$ of dimension m to power p because $\begin{pmatrix} 0 & 1 & \\ & \ddots & \ddots \\ & & \ddots \end{pmatrix}^p$ has $p - r$ Jordan boxes of dimension $[m/p]$ and r Jordan boxes of dimension $[m/p] + 1$, where $r = m - [m/p]\ p$ (the remainder from dividing m by p). (Work out the details.)

E10.30. A

Applying the same method as in section E10.29, find the basis with respect to which J has the matrix $\begin{pmatrix} \lambda_0 & \lambda_0 t_0 & \cdots & \lambda_0 t_0^{m-1}/(m-1)! \\ & \ddots & \ddots & \vdots \\ & & \ddots & \lambda_0 t_0 \\ & & & \lambda_0 \end{pmatrix} = \lambda(t_0) \cdot e^{t_0 N}$ ($N = \begin{pmatrix} 0 & 1 & \\ & \ddots & \ddots \\ & & \ddots \end{pmatrix}$) and define the necessary one-parameter group as $g(t) = \lambda(t) \cdot e^{tN}$. Work out the details.

C. By virtue of continuity, it is sufficient to verify the uniqueness of the roots, which means that

for any natural p, there exists a unique X such that $X^p = J$ *and on J's eigenspace* $X = \lambda(t_0/p) \cdot E$.

The uniqueness is established as follows. Since X commutes with J, it leaves the **root flag** of J invariant:

$$\boxed{L = L_0 \supset L_1 \supset \ldots \supset L_m = \{0\}}\ \&\ \boxed{\dim L_i = m - i}\ \&\ \boxed{(\lambda_0 E - J)^i (L_{m-i}) = \{0\}}$$
$$\Rightarrow \quad \boxed{X(L_i) = L_i,\ \forall i}.$$

(Why?) Therefore, for a Jordan basis $e_0, \ldots, e_{m-1}$ ($Je_0 = \lambda_0 e_0$, $Je_1 = e_0 + \lambda_0 e_1$, $Je_{m-1} = e_{m-2} + \lambda_0 e_{m-1}$) we have

$$\begin{aligned} Xe_0 &= \lambda(t_0/p) \cdot e_0, \\ Xe_1 &= \alpha_{10} e_0 + \alpha_{11} e_1, \\ &\ldots \end{aligned}$$

$$Xe_{m-1} = \alpha_{m-1,0}e_0 + \alpha_{m-1,m-1}e_{m-1}.$$

Prove that $e_0 + \lambda_0 e_1 = Je_1 = X^p e_1 = \left\{ \alpha_{10} \sum_{r=0}^{p-1} [\lambda(t_0/p)]^r \alpha_{11}^{p-1-r} \right\} e_0 + \alpha_{11}^p e_1$ and, hence, $\alpha_{11}^p = \lambda_0 = [\lambda(t_0/p)]^p$. Therefore, $\alpha_{11} = \lambda(t_0/p)$ (as otherwise we would have $1 = \alpha_{10} \sum_{r=0}^{p} [\lambda(t_0/p)]^r \alpha_{11}^{p-1-r} = \alpha_{10} \cdot \frac{\alpha_{11}^p - [\lambda(t_0/p)]^p}{\alpha_{11} - \lambda(t_0/p)} = 0$) and $\alpha_{10} = \{p \cdot [\lambda(t_0/p)]^{p-1}\}^{-1}$. Proceed by a similar method to establish the uniqueness of α_{ij} by induction [prove that $\alpha_{ii} = \lambda(t_0/p)$ and α_{ij} with $j < i$ are included in the corresponding equations with the same coefficients $p \cdot [\lambda(t_0/p)]^{p-1} \neq 0$]. **QED**.

D. Applying a similar method as in **C**, show that for any natural p there exists a unique X such that $X^p = J$ and on J's eigenspace $X = \lambda(t_0/p) \cdot E$. Consider the root flag of J. (This flag differs from that in **C** as follows: the dimensions of the subspaces L_i will decrease by any of 1,. . ., k as i grows.) Let $e_0^1, \ldots, e_{m_1-1}^1, \ldots, e_0^k, \ldots, e_{m_k-1}^k$ be a Jordan basis, so that L_i is spanned on $e_{\geq i}^1, \ldots, e_{\geq i}^k$. Similarly, as in **C**, we will have invariance of the subspaces L_i with respect to operator X. Suppose we have proved that

$$Xe_1^j = \sum_{i=1}^{k} \alpha_i e_0^i + \beta e_1^j \qquad \left(j \in \{1, \ldots, k\},\ \ m_j > 1\right). \tag{*}$$

Find that in this case, similarly to the case in **C**,

$$e_0^j + \lambda_0 e_1^j = Je_1^j = X^p e_1^j = \left\{ \sum_{r=0}^{p-1} [\lambda(t_0/p)]^r \beta^{p-1-r} \right\} \cdot \sum_{i=1}^{k} \alpha_i e_0^i + \beta^p e_1^j;$$

therefore, $\beta = \lambda(t_0/p)$, $\alpha_j = \{p \cdot [\lambda(t_0/p)]^{p-1}\}^{-1}$, and $\alpha_i = 0$ for $i \neq j$. More generally, if we have proved that

$$Xe_l^j = \sum_{i=1}^{k} \alpha_{li} e_0^i + \beta_{l1} e_1^j + \ldots + \beta_{ll} e_l^j, \tag{**}$$

then we could uniquely determine the coefficients by induction. Specifically, we will find that $\beta_{ll} = \lambda(t_0/p)$ and $\alpha_{li} = 0$ for $i \neq j$. (Verify!) Therefore, readers need to establish the relations (**).

To do so, verify that the set $*_M$ of linear operators on a space L, such that they map a fixed subspace M into itself, forms a subalgebra of the operator algebra gl (L)

(which means that $*_M$ is closed with respect to the linear operations over their elements and their compositions). Also, verify that the formula $\bar{A}\bar{x} := Ax+M$ $(\bar{x} = x + M \in \bar{L} = L \bmod M = L/M)$ provides a correct definition of a linear operator $\bar{A}$ on the quotient space L/M for $A \in *_M$ and the correspondence $A \mapsto \bar{A}$ defines a homomorphism of R_M into (in fact, onto) gl (L/M): $\overline{kA} = k\bar{A}$, $\overline{A+B} = \bar{A}+\bar{B}, \overline{A \circ B} = \bar{A} \circ \bar{B}, E_{\bar{L}} = \overline{E_L}$ (so $(\bar{A})^{-1} = \overline{A^{-1}}$ when A^{-1} exists). Applying this to L spanned on $e^1_{\leq 1}, \ldots, e^k_{\leq 1}$ and M spanned on $e^1_0, \ldots, e^k_0$ (which is J's eigenspace) yields $\bar{J} = \lambda_0 \cdot \bar{E}$.

Next, since commuting operators always have common eigenspaces, X has the same eigenspace as J, and so it has all its eigenvalues equal to $\lambda(t_0/p)$, and the characteristic polynomial is $\chi_X(\lambda) = (\lambda - \lambda(t_0/p))^k$. (Work out the details.) By the Cayley-Hamilton theorem, X satisfies the polynomial equation $(X - \lambda(t_0/p)\cdot E)^k = 0$, so $(\bar{X} - \lambda(t_0/p) \cdot \bar{E})^k = 0$ (why?); thus, $\bar{X}$ has all its eigenvalues equal to $\lambda(t_0/p)$. Since $\bar{X}$ is a root of the identity ($\bar{X}^p = \bar{J} = \bar{E}$) it must be a scalar operator: $\bar{X} = \lambda(t_0/p) \cdot \bar{E}$. (Why? Apply the same topological arguments as discussed previously in section E7.2 from the "2 × 2 Matrices That Are Roots of Unity" problem group to establish that $\bar{X}$ cannot have Jordan boxes of dimensions greater than one.) This proves relation (*) and yields $\beta = \lambda(t_0/p)$. (Why?)

To prove (**), apply the induction on the length of the root flag of J using equations (*). That is, assuming the relations $\bar{X}\bar{e}^j_l = \beta_{l1}\bar{e}^j_1 + \ldots + \beta_{ll}\bar{e}^j_l$ ($j = 1,\ldots, k$, $l = 1,\ldots, m_j - 1$) on the quotient space L_0/M, where, as described previously, L_0 is the whole space and M is J's eigenspace, yields the relations (**) on L_0; in turn, (**) $\Rightarrow \alpha_{li} = 0$ for $i \neq j$ (as discussed previously), so the assumed relations are inductively reproduced. **QED**. (We leave it to the reader to fill in the details.)

E. For a Jordan matrix J of two boxes of equal dimensions corresponding to the same eigenvalue $\lambda_0 > 0$, let $e^1_0, \ldots, e^1_{m-1}$, $e^2_0, \ldots, e^2_{m-1}$ be a Jordan basis as in **D**. Verify that the matrices $\begin{pmatrix} \sqrt[p]{\lambda_0}R_k & \alpha_{10}R_k & \cdots & \alpha_{m-1,0}R_k \\ & \ddots & \ddots & \vdots \\ & & & \alpha_{m-1,m-2}R_k \\ & & & \sqrt[p]{\lambda_0}R_k \end{pmatrix}$, where $R_k = \begin{pmatrix} \cos 2\pi k/p & -\sin 2\pi k/p \\ \sin 2\pi k/p & \cos 2\pi k/p \end{pmatrix}$ and α_{ij} are the same as in **D**, define linear operators that are roots of degree p from J (with respect to the rearranged basis $e^1_0, e^2_0, e^1_1, e^2_2, \ldots$). Verify that for $k = 0,\ldots, p-1$ these roots are distinct from each other and that a logarithm exists for each of them. (Use section P10.29** to work out the details.) **QED**.

Completing the Solution

S10.5. A

To prove the Leibnitz test for an alternating series, consider the sequences of its partial sums of odd and, respectively, of even number of elements. One of these sequences monotone grows, being bounded above (by the elements of the second one), while the second one monotone decreases, being bounded below (by the elements of the first one). Furthermore, the sequences have a common limit, since the differences between their members tend to zero. (Work out the details.)

The estimate $\sum_{k=2}^{2^n} k^{-1} > n/2$ is easily derived by induction, using an obvious inequality $\frac{1}{2^m+1} + \ldots + \frac{1}{2^{m+1}} > 2^m \cdot \frac{1}{2^{m+1}} = \frac{1}{2}$. To produce sharper estimates $\ln n > \sum_{k=2}^{n} k^{-1} > \ln(n+1) - 1$, consider the graph of a function $y = x^{-1}$ on $[1,\infty)$ and two series of rectangles based on segments [1,2], [2,3], [3,4], ... of the abscissas' axes: first, of heights 1, 1/2, 1/3,..., and the second, of heights 1/2, 1/3, 1/4, ... (Make a figure.) Derive (giving attention to all details!) from it estimates $\int_1^n \frac{dx}{x} > \sum_{k=2}^{n} k^{-1}$ and $\int_1^{n+1} \frac{dx}{x} < \sum_{k=1}^{n} k^{-1}$, which are equivalent to those being proved. Next, c_n are sums of areas of curvilinear triangles cut from the rectangles of the second series by the curve $y = x^{-1}$, which shows a monotone growth of c_n and brings all estimates in section E10.5 concerning them. (Work out the details.[56])

C. Let a series $\sum f_n(t)$ converge uniformly on $t \in [a,b]$, which means, using the Cauchy criterion, that $\forall \varepsilon > 0$, $\exists m_\varepsilon$: $\forall n > m \geq m_\varepsilon$, $\forall t \in [a,b]$, $\left| \sum_{i=m+1}^{n} f_i(t) \right| < \varepsilon$. Then for $a \leq t_1 \leq t_2 \leq b$,

$$\left| \sum_{i=m+1}^{n} \int_{t_1}^{t_2} f_i(t)\,dt \right| = \left| \int_{t_1}^{t_2} \sum_{i=m+1}^{n} f_i(t)\,dt \right| \leq \int_{t_1}^{t_2} \left| \sum_{i=m+1}^{n} f_i(t) \right| dt < \varepsilon(b-a),$$

which proves the term-by-term integrability. (We leave it to the reader to work out the details.)

[56] Direct using inequalities $\ln n > \sum_{k=2}^{n} k^{-1} > \ln(n+1) - 1$ brings rougher estimates $c_n \leq 1 + \ln n - \ln(n+1)$ (as $1 + \ln n - \ln(n+1) \geq 3/2 - 1/(n+1) - \ln 2$; provide more details).

S10.6

On the proof, using Method 1, that the exponential series' sums are all continuous solutions of (*). While solving the Problem P10.5*, we have proved continuous differentiability of the exponential function (the sum of the exponential series) and found that its derivative does not vanish (anywhere). For the readers' convenience, we briefly repeat those arguments here. By the d'Alembert test discussed in section H10.5, the exponential series converges at any point. From this, using the Abel's lemma (also discussed in section H10.5), it converges uniformly on any segment $[-a,a]$ (in the complex version, on any disk centered at the origin), so on any compact set, and the same is true for the series of its term-by-term derivatives of any order. Using section H10.5**C**, the sum of exponential series is infinitely differentiable (even infinitely continuously differentiable; why?).[57] Finally, $de^t/dt = e^t \neq 0$.

Due to these considerations, e^t satisfies the condition of the implicit function theorem. Using this theorem, the exponential map realizes a homeomorphism (in fact, an infinitely smooth, moreover, analytic diffeomorphism) of a neighborhood of zero onto a neighborhood of unit. (In fact, it is an analytic diffeomorphism of $\mathbb{R}$ onto $\mathbb{R}^{+*}$.) Therefore, the inverse (logarithmic) map exists in a neighborhood of unit and is continuous, so the function $t \mapsto \ln g(t) = K(t)$ is continuous. (In fact, the logarithmic map exists in $\mathbb{R}^{+*}$ and is analytic there, but its Taylor-series has a finite convergence radius with any center; can you find the relationship between the radius and the center?[58]) The rest of the proof is straightforward.

S10.13. D

Proof of the Dirichlet theorem (Cassels 1957, Schmidt 1980, and references therein). First consider integer Q. To prove the first statement (existence of integer p and $0 < q \leq Q$, such that $|p - \mu q| < Q^{-1}$), consider the following $Q + 1$ numbers from the half-interval [0,1): 0, $\{\mu\}$, ..., $\{Q\mu\}$

[57] Experienced readers probably are familiar with the **Weierstrass theorem** stating that *the sum of a uniformly convergent series* (in other terms, *the limit of a uniformly convergent sequence*) *of continuous functions itself is continuous*. It is not difficult to prove, but we do not refer to this theorem here.

[58] Readers familiar with elements of complex analysis know that the complex exponential function maps $\mathbb{C}$ onto $\mathbb{C}^*$, analytically, but not one-to-one. Indeed, this map is not one-sheeted but periodic (or infinite-sheeted), having $2\pi i\mathbb{Z}$ as its group of periods. [This is also discussed in this problem group (section H10.24).] Or the global complex logarithmic map is infinite-valued. The **ramification** (**branching**) relates to a nonsimple connectedness of the domain $\mathbb{C}^* = \mathbb{C}\backslash\{0\}$. This map cannot, due to its branching, be expanded into the Laurent series centered at the origin, although $\mathbb{C}^*$ is an annulus, and actually, a one-valued (or univalent) branch exists in a simply connected domain obtained from $\mathbb{C}$ by removing a ray of the vertex at the origin. In addition, the convergence radius of the Taylor series for (any branch of) the logarithmic map is finite with any center because it is equal to the distance from the center to the nearest singular point (which is the origin).

(where {.} denotes the fractional part of the corresponding number). If we partition [0,1) into Q half-intervals $[n/Q, (n+1)/Q)$: $n = 0,\ldots,Q-1$, at least two of these numbers will be contained in one interval, therefore, there exist integers p_1, p_2, q_1, q_2, such that $0 \le q_1 < q_2 \le Q$ and $|(\mu q_2 - p_2) - (\mu q_1 - p_1)| < Q^{-1}$. (Why?) Now we will complete the proof by setting $p = p_2 - p_1$ and $q = q_2 - q_1$. Similarly may be proved, the second statement (existence of integer p and $0 < q < Q$, such that $| p - \mu q | \le Q^{-1}$). (Provide the details.) In turn, having proved that statement for integer Q, existence of integer p and $0 < q < Q$, such that $| p - \mu q | < Q^{-1}$, will be proved for noninteger Q (by considering $[Q] + 1$, where [.] is the integral part; work out the details), which will complete the proof of the theorem.

Deduction of multidimensional generalization using Minkowski Theorem 1. The system of inequalities appearing in the theorem formulation defines, in the space $\mathbb{R}^{m+n}_{x,y}$, a closed ($m+n$-dimensional) parallelepiped, having the (m-dimensional) area of its (m-dimensional) base $(2Q)^m$ and the (n-dimensional) length of its (n-dimensional) height $(2Q^{-m/n})^n$ (provide a figure), so having the ($m+n$-dimensional) volume $(2Q)^m \cdot (2Q^{-m/n})^n = 2^{m+n}$. Thus, a straightforward application of the Minkowski theorem yields the required result.

Deduction of Minkowski Theorem 2 from Theorem 3. With respect to the coordinates $y = Mx$, the system of inequalities as in Theorem 2 determines a parallelepiped of volume $2^k c_1 \ldots c_k$ with the center at the origin of the coordinates, and returning to coordinates x retains the property of being a parallelepiped and the center and multiplies the volume by $|\det M|^{-1}$. (Work out the details.)

Deduction of Minkowski Theorem 1 from Theorem 4 (Cassels 1957). Let K denote a convex body as in Theorem 1. Application of Theorem 4 to a body $(1+\varepsilon)K$ ($\varepsilon > 0$) brings a nonempty set of nonzero integral points of it, which is bounded and discrete and, thus, finite. Hence, since $(1+\varepsilon_1)K \supseteq (1+\varepsilon_2)K$ when $\varepsilon_1 > \varepsilon_2$, there exists a nonzero integral point belonging to all of $(1+\varepsilon)K$ or, which is the same thing, belonging to $\bigcap\{(1+\varepsilon)K\colon \varepsilon > 0\} = \{$*the closure of* $K\} = K$. (Work out the details.)

Deduction of Minkowski Theorem 3 from Theorem 4 (Cassels 1957). Do not widen the preceding K in all possible directions producing $(1+\varepsilon)K$; do so only in the direction of edges coming in one of the contained faces. (We leave it to the reader to work out the details.)

(iii) *For relatively prime integers* $m_1,\ldots,m_n$*, all integer-valued vectors (the vectors of integer components) proportional to* $(m_1,\ldots,m_n)$ *are integral multiples of* $(m_1,\ldots,m_n)$. Indeed, consider a subgroup $G \cong \langle \mathbb{R}, + \rangle$ of additive group $\langle \mathbb{R}^n, + \rangle$, formed by the vectors proportional to $(m_1,\ldots,m_n)$. $H = G \cap \mathbb{Z}^n$ is its closed subgroup. Using the previous description in section P10.13. B**, H is a discrete subgroup: $H \cong \mathbb{Z}$. (Why?)[59] Let $(q_1,\ldots,q_n)$ be a generator of H; so we have $(m_1,\ldots,m_n) = k\cdot(q_1,\ldots,q_n)$ for some $k \in \mathbb{Z}$, which implies $k = \pm 1$ (by virtue of the relative primality of $m_1,\ldots,m_n$), completing the proof.

(iv) On the proof of the theorem. If for the ideals of a ring the equality $(A + B) \cap C = (A \cap C) + (B \cap C)$ holds, putting $A + C$ in place of C yields

[59] There is a generalization stating that *a subgroup of a direct sum of infinite cyclic groups is a similar direct sum itself*. (The number of summands, called **rank**, is invariant, similar to the dimension of a vector space.) For finite ranks, this can be proved geometrically, generalizing the approach developed in section S10.13. D (iii). Also, it may be proved with purely algebraic methods using the following formulation: *a subgroup of a free Abelian group is free itself*. Interested readers will find such a proof in Lang (1965). A complete proof, considering infinite ranks, is in Fuchs (1970). Connected results, concerning free objects in categories of groups, modules, and others, are discussed in group theory and homological algebra.

$$(A+B)\cap(A+C) = A\cap(A+C)+B\cap(A+C) = A+(B\cap A)+(B\cap C)$$
$$= A+(B\cap C).$$

In turn, the equality $(A + B) \cap C = (A \cap C) + (B \cap C)$ holds for the principal ideals domain. Indeed, let a, b, and c be generators for A, B, and C, respectively. Consider factorizations by primes $a = \prod p_\alpha^{k_\alpha}, b = \prod p_\alpha^{l_\alpha}, c = \prod p_\alpha^{m_\alpha}$ (which formally are taken over all equivalence classes of primes, but actually each contains a finite number of factors with nonzero exponents). The ideals $(A + B) \cap C$ and $(A \cap C) + (B \cap C)$ in the ring $\mathbb{Z}$ (more generally, in any **Euclidean ring**[60]) are generated by $\prod p_\alpha^{\max(m_\alpha,\min(k_\alpha,l_\alpha))}$ and $\prod p_\alpha^{\min(\max(k_\alpha,m_\alpha),\max(l_\alpha,m_\alpha))}$, respectively, so establishing the equality $\max(m,\min(k,l)) = \min(\max(k,m),\max(l,m))$ for elements k, l, m of a linearly ordered set (say, of the natural numbers) will complete the proof; we leave the details to the reader.

The preceding deduction of the implication $(A + B) \cap C = (A \cap C) + (B \cap C) \Rightarrow (A + B) \cap (A + C) = A + (B \cap C)$ shows that the implication holds for any sets of ideals $\mathbb{K} = \{A,B,\ldots\}$ of a ring R, closed with respect to sums [that is, $(A + B) \in \mathbb{K}$ if $A \in \mathbb{K}$ and $B \in \mathbb{K}$]. And so the result will hold for the ideals from $\mathbb{K}$, as long as the antecedent holds. That implication holds for a set of mutually prime ideals of an associative unitary ring R, including R itself. Also, for such K, the antecedent holds; indeed, since R has a unit, we have

$$(A+B)\cap C = R\cap C = C = R\cdot C = (A+B)\cdot C = A\cdot C + B\cdot C \quad \subseteq \quad (A\cap C)+(B\cap C)$$
$$\subseteq \quad C,$$

and so for these $\mathbb{K}$ the result holds.

On the other hand, one may easily give examples of rings and ideals such that both antecedent and consequent are strict inclusions, but not equalities. Indeed, in the ring of polynomials in x, y, with the coefficients from an infinite field, consider the ideals A, B, C generated by $x + y$, x, y, respectively. (We leave it to the reader to work out the details.)

S10.13. F

Let θ_1, θ_2 be complex arguments of the eigenvalues of B, incommensurable with 2π and commensurable modulo 2π: $m_2\theta_1 \equiv m_1\theta_2 \bmod (2\pi)$ for some relatively prime m_1, m_2. Choosing integers k_1, k_2 such that $\frac{\theta_1+2\pi k_1}{m_1} \neq \frac{\theta_2+2\pi k_2}{m_2}$ and the remaining k_j

[60] This is a ring for which any computation using Euclid's algorithm is finite. More formally, it is defined as an integral domain considered as a Euclidean module (as defined in section E10.13. D) over itself whose Euclidean function has the following extra property: $ab = c \Rightarrow \varepsilon(a), \varepsilon(b) \leq \varepsilon(c)$. The Euclidean ring is a unitary one because the existence of the identity element is implied by the property of being a Euclidean module; indeed, if $\varepsilon(a) = \min(\varepsilon)$ and $ae = a$, then e is that element. (Fill in the details. ***Warning***: the ring of even integers is a Euclidean module over $\mathbb{Z}$, but it is not a Euclidean ring!) Also, the Euclidean ring is a principal ideal domain, but the unique factorization theorem (holding for any such domains) may be proved for Euclidean rings without use of an axiom of choice (using an induction on the ordering in the range of the Euclidean function). Interested readers will find the details in Van der Waerden (1971, 1967).

arbitrarily, and defining the corresponding angular velocities as $\omega_j = (\theta_j + 2\pi k_j)/t_0$ for any fixed $t_0 \neq 0$ yields the one-parameter group g, with $g(t_0) = B$, which on the eigenspaces acts as multiplications by scalars $\cos \omega_j t + i \sin \omega_j t$, respectively, but $\omega_j t$ will never (for any t) become integral multiples of 2π simultaneously for all j. (We leave it to the reader to work out the details.)

S10.17

By the lemma in section P10.17**, $e^{\|A\|}$ is a majorant convergent numerical series for the matrix series e^A.

A generalization of the statement in section H10.5**C** for the series of univariable vector or matrix functions may be established by the same method as in section E10.5, employing vector- valued (resp. matrix- valued) integrals instead of scalar ones.

S10.18

A simple example of a nondifferentiable function, but with directional derivatives, is produced by $f(x,y) = \begin{cases} x \cdot \sin\left(2\,\mathrm{arctg}\frac{y}{x}\right) = \frac{2x^2y}{x^2+y^2}, & \text{for} \quad x^2+y^2 \neq 0 \\ 0, & \text{for} \quad x^2+y^2 = 0 \end{cases}$. Indeed, if it had the tangent plane to its graph at $(0,0,0) \in \mathbb{R}^3_{x,y,f}$, it would be $f = 0$ (because the graph contains straight lines $x = 0$ and $y = 0$ of this plane), so the differential would be zero at the origin. Readers may verify that the directional derivative at the origin in the direction of any increment $v = (\cos\theta, \sin\theta)$ exists, equaling $\frac{\partial f}{\partial v}(0,0) = \cos\theta \cdot \sin 2\theta$, which in general is distinct from zero, thereby making the function nondifferentiable. The directional derivative depends on the increment nonlinearly (since $\cos\theta \cdot \sin 2\theta$ is identical with $a \cdot \cos\theta + b \cdot \sin\theta$ for no constants a, b). Also this proves nondifferentiability at the origin. $f(x,y) = \begin{cases} \frac{xye^{-1/y^2}}{x^2+e^{-2/y^2}}, & \textit{for } x,y \neq 0, \\ 0, & \textit{for } x = 0 \textit{ or } y = 0 \end{cases}$ yields a finer example of a nondifferentiable function with a directional derivative at the origin, linear with respect to the increment (in fact, the derivative equals zero). A limit relation $f(x,y)/(x^2+y^2)^{1/2} \to 1/2$ for $y = (-\ln|x|)^{-1/2}$, as $x \to 0$, shows nondifferentiability at the origin. (Fill in all the details.)

The functions of the preceding examples were continuous. Readers may find discontinuous functions having directional derivatives one their own; also, examples may be found in Gelbaum and Olmsted (1964).

Next, readers may obtain the estimates in section E10.18 as follows. First,

$$\begin{aligned} \|A^n - B^n\| = \|A^n - A \circ B^{n-1} + A \circ B^{n-1} - B^n\| \quad &\leq \quad \|\mathrm{A}\| \cdot \|A^{n-1} - B^{n-1}\| \\ &+ \|\mathrm{A}\|^{n-1} \cdot \|A - B\|, \end{aligned}$$

which yields the desired result by using the inductive hypothesis. (Work out the details.) Using it and adding $\pm \underbrace{B \circ \ldots \circ B}_{k-1} \circ H \circ \underbrace{A \circ \ldots \circ A}_{n-k}$ to $\underbrace{A \circ \ldots \circ A}_{k-1} \circ H \circ \underbrace{A \circ \ldots \circ A}_{n-k} - \underbrace{B \circ \ldots \circ B}_{k-1} \circ H \circ \underbrace{B \circ \ldots \circ B}_{n-k}$ yields

$$\begin{aligned}
&\left\| \underbrace{A \circ \ldots \circ A}_{k-1} \circ H \circ \underbrace{A \circ \ldots \circ A}_{n-k} - \underbrace{B \circ \ldots \circ B}_{k-1} \circ H \circ \underbrace{B \circ \ldots \circ B}_{n-k} \right\| \\
&\quad \leq \|H\| \cdot \|A - B\| \cdot \Big[(k-1)\|A\|^{n-k} \max\left(\|A\|, \|B\|\right)^{k-2} \\
&\quad + (n-k)\|B\|^{k-1} \max\left(\|A\|, \|B\|\right)^{n-k-1} \Big],
\end{aligned}$$

which produces the desired result. (Work out the details.) From this, in turn,

$$\begin{aligned}
&\left\| \frac{de^X}{dH}(A) - \frac{de^X}{dH}(B) \right\| \leq \sum_{n=1}^{\infty} (n!)^{-1} \sum_{k=1}^{n} \| \underbrace{A \circ \ldots \circ A}_{k-1} \circ H \circ \underbrace{A \circ \ldots \circ A}_{n-k} \\
&\quad - \underbrace{B \circ \ldots \circ B}_{k-1} \circ H \circ \underbrace{B \circ \ldots \circ B}_{n-k} \| \\
&\quad \leq \|H\| \cdot \|A - B\| \cdot \sum_{n=1}^{\infty} (n!)^{-1} n(n-1) \cdot \max\left(\|A\|, \|B\|\right)^{n-2} \\
&\quad = \|H\| \cdot \|A - B\| \cdot e^{\max(\|A\|, \|B\|)}.
\end{aligned}$$

QED. (We leave it to the reader to fill in the details.)

We do not want the previous calculations to give the reader the impression that the exponential series have some special properties of differentiability. This is wrong, and, in fact, similar estimates are valid for any matrix series with nonzero convergence radii; indeed, the coefficients' decreasing at the rate of a geometric progression will remain after multiplying them by factors of a polynomial growth.

A. The sum of a convergent numerical series $\sum a_n$ of nonnegative elements, S, is the least upper bound for the partial sums: $S = \sup_n \sum_{i=1}^{n} a_i$. Therefore, $\forall \varepsilon > 0$, $\exists n = n(\varepsilon)$ such that $\sum_{i \in I} a_i < \varepsilon$ for any finite index set I of elements greater than n (ε). This implies commutative convergence, which means that the sum remains invariant after any permutations of the terms. Indeed, with the notation $n_\sigma(\varepsilon) = \max_{1 \leq i \leq n(\varepsilon)} (\sigma^{-1}(i))$, relative to a permutation σ on the natural number series, we have $\left| \sum_{i=1}^{n} a_i - \sum_{j=1}^{k} a_{\sigma(j)} \right| < \varepsilon$ for any $n \geq n(\varepsilon)$ and $k \geq n_\sigma(\varepsilon)$. (Work out the details.)

Similarly, for a normally convergent series $\sum A_n$, for any $\varepsilon > 0$ there exists a natural $n(\varepsilon)$ such that $\sum_{i\in I} \|A_i\| < \varepsilon$ for any finite index set I consisting of elements greater than $n(\varepsilon)$, which implies the commutative convergence of $\sum A_n$, keeping the sum invariant. Indeed, we have $\left\| \sum_{i=1}^{n} A_i - \sum_{j=1}^{k} A_{\sigma(j)} \right\| < \varepsilon$ for any $n \geq n(\varepsilon)$ and $k \geq n_\sigma(\varepsilon)$. (We leave it to the reader to work out the details.)

A famous classic **theorem** established by **Bernhard Riemann** states that the commutative convergence of a numerical series implies absolute one (and so is equivalent to it),[61] which may be generalized for series in finite-dimensional vector spaces (replacing absolute convergence by normal one).

In the foregoing considerations, the permutations were assumed finite, that is, there was always a finite number of elements of the rearranged series $\{\sigma(0), \sigma(1), \ldots\}$ between any two elements $\sigma(n)$ and $\sigma(n + 1)$. May we expect an equivalence of commutative and absolute (resp. normal) convergences using transfinite rearrangements that break the natural-type ordering[62]? Our intuition votes for the positive answer; but first we must define commutative and normal convergences of series of such exotically ordered index sets!

Such definitions might comprise a transfinite induction on the ordering; however, in the end we will have to deal with the supremum on sums of positive numbers taken over finite index subsets, so we do not need any unjustified complications. Following Schwartz and Huet (1961), let us call a countable family $\{v_i\}_{i\in I}$ in a vector space that is complete with respect to a fixed norm $\|.\|$ (e.g., in a finite-dimensional space it may be the maximum, or the sum of absolute values of coordinates with respect to a fixed basis) **summable** to a sum $S = \sum_{i\in I} v_i$ if, given $\forall \varepsilon > 0$, there exists a finite subset $J_\varepsilon \subseteq I$ such that $\left\| S - \sum_{i\in J} v_i \right\| < \varepsilon$ for any finite set $J_\varepsilon \subseteq J \subseteq I$. As readers may easily verify, summability is equivalent to the existence, $\forall \varepsilon > 0$, of a finite subset $J_\varepsilon \subseteq I$ such that $\left\| \sum_{i\in K} v_i \right\| < \varepsilon$ for any finite set $K \subseteq \not\!\!I_\varepsilon$ (the **Cauchy criterion**), which in turn implies the uniqueness of the sum. A **normal summability** is defined in a similar way, using the inequality $\sum_{i\in K} ||v_i|| < \varepsilon$. (In other words, the normal summability of a family is the summability of the family of its term-by-term norms in $\mathbb{R}^+$.) Note that we do not consider the index set I as having any order, so that the summability is commutative by definition.

As discussed previously in the "2 × 2 Matrices That Are Roots of Unity" problem group (see section H7.6), the norms in a finite-dimensional space are equivalent, $c_1 \cdot \|v\|_1 \leq \|v\|_2 \leq c_2 \cdot \|v\|_1$, $\forall v$, with appropriate (constant) $c_1, c_2 > 0$, and so we may do without the norm's specification when dealing with finite-dimensional spaces.

---

[61] A proof close to the original can be found in Fichtengoltz (1970).

[62] For example, defining a new order "$\prec$" as $2\,k \prec 2n$, $2\,k + 1 \prec 2n + 1$ for $k < n$, $2\,k + 1 \prec 2n$, $\forall k,n$, or even in a more sophisticated way, as $\prod p_i^{\alpha_i} \prec \prod p_i^{\beta_i}$, if, for some n, $\alpha_i \leq \beta_i$, $\forall i > n$ and $\alpha_n < \beta_n$. (In the last formula, the products are formally taken over all primes $p_1 = 2$, $p_2 = 3, \ldots$, but actually each contains a finite number of factors with nonzero exponents.)

The reader should verify that normally convergent series (single or multiple) are normally summable families, and *a convergent, but not absolutely convergent, numerical series is not a summable family*. (Use Riemann's theorem to find inadmissible sets $J \supseteq J_\varepsilon$.)

Evidently, normal summability implies summability. The analog of Riemann's theorem also holds.

Theorem. *A summable family in a finite-dimensional space is normally summable.*

(Specifically, this theorem concerns real-, complex-, quaternion-, matrix-, and similar valued families.) A proof is made using the following lemma.

Lemma. *A family is summable if, for some finite partition of the index set* $I = \bigcup I_\alpha$, *the corresponding subfamilies are summable, and in this case* $S = \sum S_\alpha$. *Conversely, for any finite partition of the index set, the subfamilies are summable if the whole family is summable.*

This claim is directly implied by the Cauchy criterion. (Work out the details.) Now, a common proof of the theorem for scalar (real-valued) families uses partitions into nonnegative and negative subfamilies, as the summability of these subfamilies obviously is equivalent to absolute summability. (Work out the details.) Actually, all finite-dimensional families allow similar partitions! Finding them will require a bit more consideration. Cover the unit sphere by a finite number of convex bodies (say, balls) of diameters less than one-half. (Why is this possible?) Using for the vectors the spherical coordinates $v \leftrightarrow (r,\sigma)$ ($r = |v|$, $\sigma = v/|v|$), partition the source family into subfamilies of σ belonging to those bodies, respectively. We must establish the normal summability of these subfamilies, so consider any of them as a separate family. Let J_ε be a finite index subset such that $\left\|\sum_{i\in K} v_i\right\| < \varepsilon$ for any finite set $K \subseteq \bigwedge J_\varepsilon$ (in accordance with the Cauchy criterion). We obtain

$$\varepsilon > \left\|\sum_{i\in K} r_i\sigma_i\right\| = \left(\sum_{i\in K} r_i\right)\cdot\left\|\sum_{i\in K}\frac{r_i}{\sum\limits_{j\in K} r_j}\sigma_i\right\|.$$

Our convex body contains the vector $\sum_{i\in K}\frac{r_i}{\sum_{j\in K} r_j}\sigma_i$ (why?), so its modulus cannot be less than one-half, and, finally, $2\varepsilon > \sum_{i\in K} r_i$, which, using the Euclidean norm (since any norm may be employed), shows normal summability. **QED**. (We leave it to the reader to fill in the details.)

B. (A special case of this theorem concerning absolutely convergent numerical series was established by A.L. Cauchy.) Consider, for simplicity, the product of two normally convergent series. (The general case follows from it by induction.) We have, using the notations $a_i = \|A_i\|$, $b_i = \|B_i\|$, $\sum\limits_{0\le i\le k,\,0\le j\le n} a_i b_j = \left(\sum\limits_{i=0}^{k} a_i\right)\cdot\left(\sum\limits_{j=0}^{n} b_j\right)$, which, accounting for the lemma in section E10.17, shows the (normal) convergence of the double series $\sum\limits_{k,n} A_k \circ B_n$. To find its sum, we may, due to the normal convergence, use partial sums of $k = n$

only. (Why? Readers may fill in the details drawing some quite elementary arguments or using a more advanced technique discussed previously in paragraph **A**.) And the sequence of those partial sums of series $\sum\limits_{k,n} A_k \circ B_n$ is the same as that of products of partial sums of series $\sum A_n$ and $\sum B_n$, which completes the proof.

Applying the proof to the exponential series $e^{A+B} = \sum\limits_{k,n} (k!n!)^{-1} A^k \circ B^n$ (for commuting A, B) yields the equality of this series to the product of the sums of series e^A and e^B. (We leave it to the reader to work out the details.)

C. We have

$$\left| \sum_{k=m+1}^{n} a_{k+n_0} t^k \right| \le t_1^{-n_0} \sum_{k=m+n_0+1}^{n+n_0} |a_k| t_1^k \le C t_1^{-n_0} \sum_{k=m+n_0+1}^{n+n_0} |t_1/t_0|^k \to 0 \quad \text{as}$$

$$m, n \to \infty,$$

which yields the required result using the Cauchy criterion.

D. Denoting $\max(\|A\|,\|B\|) = t$ and replacing coefficients of the series $R_i(A,B)$ by their absolute values yields majorizing numerical series, having uniformly bounded sums for bounded t. (Apply paragraph **C** to work out the details.)

S10.20

On the proof of the lemma in section H10.20. Consider, fixing any basis, matrices for the linear operators of a sequence, convergent with application to any vector. Show that the sequence of these matrices converges in the matrix space. (Consider the convergences of the sequences of columns corresponding to any fixed indices.)

S10.21

On the step "(*ii*) ⇒ (*iii*)" in the proof (in section H10.21) of the proposition from section P10.21***. All quasimonomials, including $(k_j!)^{-1}t^{k_j}e^{\lambda_j t}$, and $(k_j!)^{-1}t^{k_j}e^{\lambda_j t}\cos\omega_j t$, $(k_j!)^{-1}t^{k_j}e^{\lambda_j t}\sin\omega_j t$ for $\omega_j \neq 0$, of $k_j = 0,1,\ldots,N_j$, for any fixed set of nonnegative integers N_j and any fixed set of distinct complex numbers $z_j = \lambda_j + i\omega_j$, are intheir totality linearly independent because they form a fundamental base (a basis of the vector space) of solutions of a linear autonomous ODE of order $\sum_{\omega_j=0}(N_j+1) + 2\sum_{\omega_j\neq 0}(N_j+1)$ [63] (namely, "$\prod_{\omega_j=0}\left(\frac{d}{dt}-\lambda_j E\right)^{N_j+1}\prod_{\omega_j\neq 0}\left[\left(\frac{d}{dt}-\lambda_j E\right)^2+\omega_j^2 E\right]^{N_j+1}x=0$"). A straightforward verification shows that, actually, for the operator d/dt on the space of solutions S, z_j (taken with their multiplicities) are the eigenvalues, and those quasimonomials form a Jordan basis (over the reals). (The reader is encouraged to perform all the necessary

[63] Readers not experienced enough in ODE are encouraged to use a different method of establishing the linear independence of those quasimonomials (as functions on $\mathbb{R}$). Let some linear combination of these monomilas be identical to zero: $f := \sum\left(p_j(t)\cos\omega_j t + q_j(t)\sin\omega_j t\right)e^{\lambda_j t} \equiv 0$. The same arguments as were used in section E10.21 to prove Proposition 2 show that f cannot include nontrivial exponential or polynomial terms (because otherwise f would be unbounded; complete the details), so f has the form $f = \sum A_j\cos\omega_j t + B_j\sin\omega_j t$, with constant coefficients A_j, B_j. We may assume that all $\omega_j \geq 0$. (Why?) We must verify that the coefficients vanish. (The summand $B_j\sin 0$ corresponding to $\omega_j = 0$ may be excluded from consideration.) The summands of f may be expanded into power series, convergent for any t, so f may be expanded into power series itself. Such a series is unique (and thus is a Taylor series) because *a function that is the sum of a convergent power series cannot possess a nondiscrete zero set*. (Fill in the details.) Hence, Taylor expansions of the summands of f are summed in a zero series, which, using commutative convergence (as discussed in section S10.18 above), yields the equations

$$\sum A_j = 0,\quad \sum A_j\omega_j^2 = 0,\quad \sum A_j\omega_j^4 = 0,\ldots \quad\text{and}\quad \sum B_j\omega_j = 0,\quad \sum B_j\omega_j^3 = 0,\quad \sum B_j\omega_j^5 = 0,\ldots.$$

The generalized Vandermonde determinants $\det\begin{pmatrix}\omega_0^{k_0} & \cdots & \omega_n^{k_0}\\ \vdots & & \vdots\\ \omega_0^{k_n} & \cdots & \omega_n^{k_n}\end{pmatrix}$ are positive for $0 < \omega_0 < \ldots < \omega_n$ and $k_0 < \ldots < k_n$. [Readers should try to prove these relations; also, we give a proof subsequently in the "Least Squares and Chebyshev Systems" problem group: a quite elementary proof, for a special case of natural k_j, is discussed in section H12.4, and in the general case, a proof follows from the results of section P12.10; in addition, a different proof can be found in Polya et al. (1964).] Therefore, we will find that A_j and B_j vanish. **QED**. (We leave it to the reader to fill in the details.)

computations! The linear independence of the preceding solutions may be proved using the proposition discussed later in this section.)

We must prove that a vector space of differentiable functions, closed with respect to the derivatives of its elements, will contain all of * if it contains a quasipolynomial $f = \sum \left(p_j(t)\cos\omega_j t + q_j(t)\sin\omega_j t\right)e^{\lambda_j t}$ with max (deg p_j, deg q_j) $= N_j$, $\forall j$. Arrange the quasimonomial basis as a sequence $u_0^1, \ldots, u_{N_1}^1, \ldots$, $v_0^m, w_0^m, \ldots, v_{N_m}^m, w_{N_m}^m, \ldots$ (vectors $u_{k_j}^j$ and $v_{k_j}^j, w_{k_j}^j$ correspond, respectively, to real and nonreal z_j) and consider the related Jordan flag $\ddot{\imath} = \ddot{\imath}_0^1 \supset \ldots \supset \ddot{\imath}_{N_1}^1 \supset \ldots \supset \ddot{\imath}_0^m \supset \ldots \supset \ddot{\imath}_{N_m}^m \supset \ldots \supset \{0\}$ (of the dimensions decreasing, respectively, by one and two in its first and second parts). By construction, the flag is invariant with respect to d/dt (which means that d/dt maps all subspaces of the flag into themselves). Verify that for real eigenvalues, $d/dt - \lambda_j E$ maps $\ddot{\imath}_{k_j}^j$ onto $\ddot{\imath}_{k_j+1}^j$ (where $\ddot{\imath}_{N_j+1}^j = \ddot{\imath}_0^{j+1}$) and is an isomorphism on a root subspace corresponding to a distinct z_k, but for nonreal eigenvalues, $d/dt - \lambda_j E$ is an isomorphism on a root subspace corresponding to both z_j and any other nonreal eigenvalue (because dx/dt and x are not real-proportional, regardless of the x chosen in such a subspace), and along with it, $(d/dt - \lambda_j E)^2 + \omega_j{}^2 E$ maps $\ddot{\imath}_{k_j}^j$ onto $\ddot{\imath}_{k_j+1}^j$ (where $\ddot{\imath}_{N_j+1}^j = \ddot{\imath}_0^{j+1}$). Derive from it the linear independence of the vectors

$$f,\ \left(\frac{d}{dt} - \lambda_1 E\right)f, \ldots,\ \left(\frac{d}{dt} - \lambda_1 E\right)^{N_1+1} f,\quad \left(\frac{d}{dt} - \lambda_1 E\right)^{N_1+1}\left(\frac{d}{dt} - \lambda_2 E\right)f, \ldots,$$
$$\prod_j \left(\frac{d}{dt} - \lambda_j E\right)^{N_j+1} f, \prod_j \left(\frac{d}{dt} - \lambda_j E\right)^{N_j+1}\left(\frac{d}{dt} - \lambda_m E\right)f,$$
$$\prod_j \left(\frac{d}{dt} - \lambda_j E\right)^{N_j+1}\left[\left(\frac{d}{dt} - \lambda_m E\right)^2 + \omega_m^2 E\right]f, \ldots$$

(what is the last term?), completing the proof.

Note that similar results may be proved concerning a Jordan flag for any linear operator on a finite-dimensional vector space that possesses a single Jordan box (in a complexified space) corresponding to any of its eigenvalues! In particular, the following proposition is true.

Proposition. *If, for nonzero elements* $e_{11}, \ldots, e_{1,n_1}, \ldots, e_{k1}, \ldots, e_{k,n_k}$ *of some vector space, a linear operator A on it, and distinct scalars* $\lambda_1, \ldots, \lambda_k$, $(A - \lambda_i E)e_{ir} = e_{i,r+1}$ *for* $r = 1, \ldots, n_i - 1$ *and* $(A - \lambda_i E)e_{i,n_i} = 0$, $\forall i = 1, \ldots, k$, *then these vectors are linearly independent.*

Proof. If, for nonzero elements $v_1, \ldots, v_n$ of some vector space and a linear operator B on it, $Bv_1 = v_2$, $\ldots$, $Bv_{n-1} = v_n$ and $Bv_n = 0$, then these vectors are

linearly independent: a linear combination $l = \sum \alpha_i v_i$ equaling zero does not contain the term $\alpha_1 v_1$ with $\alpha_1 \neq 0$, because that would imply that $0 \neq \alpha_1 v_n = B^{n-1} l = B^{n-1} 0 = 0$, and so on. Hence, $e_{i1}, \ldots, e_{i,n_i}$ are linearly independent for any i. The subspaces L_i spanned on these vectors are invariant with respect to all of $A - \lambda_j E$. For $j \neq i$, the restriction of $A - \lambda_j E$ to L_i is an isomorphism because it does not map a nonzero vector onto zero: that is, a linear combination $\sum \alpha_r e_{ir}$ that has its first nonzero coefficient at $r = r_0$ is mapped onto a vector of the form $(\lambda_i - \lambda_j)\alpha_{r_0} e_{i,r_0}$ + *terms proportional to* e_{ir} *with* $r > r_0$, which is distinct from zero. Thus, L_i intersects with a kernel of $\prod_{j \neq i} (A - \lambda_j E)^{n_j+1}$, which contains the sum of the rest of L_j, by $\{0\}$.

We leave it to the reader to complete the proof.

As readers familiar with linear differential equations know, an ODE of nth order $f^{[n]} + a_1 f^{[n-1]} + \ldots + a_n f = 0$ may be considered a linear system $\dot{x} = Ax$: $x = (x_0, \ldots, x_{n-1})$, $A = \begin{pmatrix} 0 & 1 & & \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & 1 \\ -a_n & \cdots & \cdots & -a_1 \end{pmatrix}$. Such a matrix possesses a single Jordan box (in a complexified space) corresponding to any of its eigenvalues (and, of course, conversely, a matrix with this property is similar to A, with certain $a_1, \ldots, a_n$), and has the characteristic polynomial $\chi_A(\lambda) = \lambda^n + a_1 \lambda^{n-1} + \ldots + a_n$. Indeed, a direct computation yields for an eigenvector x corresponding to an eigenvalue λ equations

$$x_1 = \lambda x_0, \ldots, \quad x_{n-1} = \lambda x_{n-2}, \quad \lambda x_{n-1} + a_1 x_{n-1} + \ldots + a_n x_0 = 0,$$

from which, taking into account that $x \neq 0$, so that x_0 may be set to 1, we obtain, in succession, $x_k = \lambda^k$, and from the final equation, $\lambda^n + a_1 \lambda^{n-1} + \ldots + a_n = 0$. **QED**. (We leave it to the reader to fill in the details.) Another version of the proof using linear differential equations is in Arnol'd (1975).

S10.24

More advanced readers familiar with differential forms and the complex (Cauchy) integration may determine ker exp by arguing as follows. It is a discrete group, and so the (many-valued) complex logarithm satisfies the equations (**) from section P10.10* for s close to 1 and any t. (Why? Work out the details.) Derive from it the differential equation for a logarithmic function. (Also, this equation may be obtained from the differential equation for the exponential function from section P10.3*, using the implicit function theorem as discussed in section H10.10.) We have the differential form $d \ln z = dz/z$, which is one-valued on the punctured complex straight line $\mathbb{C}^* = \mathbb{C} \backslash \{0\}$. It is a holomorphic, then (by the Cauchy theorem; work out the details) a closed form, and therefore, it is an exact form (a differential) on a simply connected domain obtained from $\mathbb{C}$ by removing a ray with its vertex at the origin. Thus, a one-valued logarithm's branch exists in such a domain, that is, defined as $\ln z = \int_1^z \frac{du}{u}$, with the integral on the right-hand side taken along any **rectifiable** curve in this domain between 1 and z. However, dz/z is not a differential

on $\mathbb{C}^*$ because values of the preceding integral taken along different curves in $\mathbb{C}^*$ may differ by a multiple of a certain number because it is just a generator of the group ker exp. (In turn, the multiplicity equals the **index** of the curve, that is, the number of complete counterclockwise minus clockwise rotations around the origin.)[64] Thus, the generator may be found by contour integration $\oint \frac{dz}{z}$ along any (rectifiable) closed loop, one time going around the origin (any such loop will do because of the closure of dz/z, which implies equality to zero of the integral along a boundary; fill in the details). Finally, integration along a circle centered at the origin gives $2\pi i$ for the integral, resulting in $2\pi i\mathbb{Z}$ for ker exp. **QED**. (We leave it to the reader to fill in the details.)

S10.27

(5) The Jacobi matrix can be decomposed into the following three factors: (1) an $n^2 \times n^2$ matrix, corresponding to an invertible, thus nondegenerate, linear map $X \mapsto B^{-1} \circ X$; (2) $n^2 \times n^2$ matrix $\partial \log X/\partial X$, nondegenerate in a neighborhood of E, by section P10.27** **(4)**; and (3) an $\binom{n+1}{2} \times n^2$ matrix, corresponding to the linear system $Y_{km} + Y_{mk} = 0$: $k,m = 1,\ldots,n$, nondegenerate because it has $\binom{n}{2}$ linearly independent solutions $\left\{Y_{km}^{(ij)}\right\}_{k,m=1,\ldots,n}$ $(1 \le j < i \le n)$ of $Y_{km}^{(ij)} = \begin{cases} 1 & \text{for } k = i,\ m = j \\ -1 & \text{for } k = j,\ m = i \\ 0 & \text{otherwise} \end{cases}$. (We leave it to the reader to complete the proof.)

(7) The desired formula $\mathrm{A}^* \mapsto C \circ \mathrm{A}^* \circ C^{-1}$ formally follows from the formulas $\mathrm{A}^* \mapsto G^{-1} \circ {}^t\mathrm{A} \circ G$, $\varsigma \mapsto C \circ \varsigma \circ C^{-1}$ and $G \mapsto {}^tC^{-1} \circ G \circ C^{-1}$. (We leave it to the reader to work out the details.)

(9) Alternatively, readers may prove the implication $e^{tA} \circ e^{tA^*} = E,\ \forall t \ \Rightarrow A + A^* = 0$ via differentiation. Differentiating the left-hand side of the identity $e^{tA} \circ e^{tA^*} = E$ at $t = 0$ by the differentiation formula for composite functions (or maps)

$$\frac{d}{dt}\Big|_{t=t_0} f(u(t), v(t), \ldots) = \frac{\partial f}{\partial u}(u(t_0), v(t_0), \ldots) \circ \frac{du}{dt}(t_0) + \frac{\partial f}{\partial v}(u(t_0), v(t_0), \ldots) \circ \frac{dv}{dt}(t_0) + \ldots,$$

[64] Readers familiar as well with coverings know that, topologically, an exponential map realizes an infinite-sheeted covering $\begin{cases} \mathbb{C} & \cong & \mathbb{R} \times \mathbb{R} & \to & \mathbb{R}^{+*} \times \mathbb{S}^1 \cong \mathbb{C}^* \\ z & = & x + yi & a & e^x \cdot e^{yi} \end{cases}$. The raising of a closed loop $[0,2\pi] \to \mathbb{C}^*$, rotating n times around the origin, into the total space of the covering is an open loop connecting points u, v, of the kth and $(k + n)$th sheets of this covering, respectively, such that $e^u = e^v$. (Produce a figure.) Pay attention to the realization of the raising, as it is made by the logarithmic map (inverse to the exponential), defined by integrating the differential form dz/z along the source loop in $\mathbb{C}^*$!

or, more exactly, by its special case commonly known as the Leibniz differentiation rule for multilinear functions (or maps) f, yields $A + A^*$, and so $A + A^* = 0$. **QED**. (We leave it to the reader to fill in the details.)

(11) Obviously, the map $(x,y,u,v) \mapsto (a,b,c,d)$ is (infinitely) differentiable (actually, rational, with no poles if $x \neq 0$). It is also invertible since the denominators in the explicit inversion formulas from section E10.27 **(11)** do not vanish (so the definition of the inverse map by these formulas is correct for any a, b, c, and d with $ad - bc > 0$). The map defined by the inversion formulas is (infinitely) differentiable since the nonvanishing of the denominators makes the partial derivatives continuous everywhere. (Use the results discussed previously in section E10.18.) Therefore, the (source and inverse) map is (infinitely) bidifferentiable.

Actually this map is bianalytic (an analytic diffeomorphism), which may be verified by direct expansion into a Taylor series. (Work out the details.) Also, it may be found using the implicit function theorem because the map is analytic and has a nonvanishing Jacobian [equal to $-2x^2\cdot(u^2 + v^2)$ (compute it!), which does not vanish if $x \neq 0$ and $u^2 + v^2 \neq 0$].

(12) A general example of disagreement of complex and symplectic structures while some Lagrangian plane is mapped by $I = (-E)^{1/2}$ onto a transversal one may be given as follows. Consider a symplectic space $\mathbb{R}^4$ with a Darboux basis ξ_1, η_1, ξ_2, η_2 (so that for $v' = x'_1\xi_1 + y'_1\eta_1 + x'_2\xi_2 + y'_2\eta_2$ and $v'' = x''_1\xi_1 + y''_1\eta_1 + x''_2\xi_2 + y''_2\eta_2$, $[v', v''] = \sum_{i=1,2} x'_i y''_i - x''_i y'_i$). Readers may verify that a linear operator mapping the basis vectors onto, respectively, η_2, ξ_2, $-\eta_1$, $-\xi_1$ is a square root from $-E$ and disagrees with [.,.]. (Work out the details; for instance, verify that I maps the Lagrangian space spanned on $\xi_1 + \eta_1$ and $\xi_2 + \eta_2$ onto itself, and not onto a transversal plane.)

(13) A complex $n \times n$ matrix ς satisfying the equation $A + {}^t\bar{A} = 0$ is uniquely defined, given any n complex entries above its diagonal and any n real diagonal entries, so the dimension, over the reals, is dim u(n) $= n^2$. In turn, a real $2n \times 2n$ matrix $\begin{pmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{pmatrix}$ satisfying the equations $A_{12} + {}^tA_{12} = 0$, $A_{21} + {}^tA_{21} = 0$, and $A_{11} - {}^tA_{22} = 0$ is uniquely determined by $n(n + 1)$ entries of A_{12} and A_{21} placed above and on their diagonals and n^2 entries of ς_{11}, so dim sp(n) $= n(n + 1) + n^2 = n(2n + 1)$.

(20) To complete the proof, we must show that if the map T preserves orthogonality, it will preserve parallelism and coefficients of proportionality between affine vectors. Preserving orthogonality implies preserving parallelism, so the vertices of a parallelogram are mapped onto the vertices of a parallelogram; therefore, establishing proportional lengthening of the vectors of the same starting point is enough. With this, since vectors of equal lengths and of the same starting point are lengthened equally, it is enough to consider collinear vectors. (Work out the details.) Therefore, restrict T to a straight line. Since for three sequential points x, y, z on this line y is always a basis of the height of a rectangular triangle with a hypotenuse xz, T preserves or inverts the order of the points, in other words, is monotone. (Why?) Since preserving the equality of lengths implies proportionally lengthening the vectors of commensurable lengths (why?), T has no gaps of

nonempty interior in the range of its images and so, taking the monotonicity into account, is continuous on this straight line. (Work out the details.) As a result, T lengthens all vectors proportionally. **QED**.

Some Problems in Combinatorics and Analysis That Can Be Explored Using Generating Functions

Problems

P11.0

Preliminaries. For a *finite* (or countably infinite) sequence a_n, the generating function is defined as the polynomial (resp. formal series) $\sum a_n\, t^n$. For a multi-sequence $a_{m,n,\ldots}$, the generating function is a polynomial or series in several variables $\sum a_{m,n,\ldots} t^m u^n \ldots$.[1]

**Readers familiar with formal power series and partial fractions probably know that generating functions can be effectively used for the explicit determination of general terms of linear recursive sequences.[2] Let us consider the sequence $\{f_n\}_{n\,=\,0,1,\ldots}$ with $f_0 = 1$ satisfying the recursion relation $f_{n+1} = \lambda f_n$. These conditions impose on the generating function $F(t) = \sum f_n t^n$ a linear equation $F - 1 = \lambda t F$, or $(1 - \lambda t)\cdot F = 1$, having the solution $F = (1 - \lambda t)^{-1} = \sum (\lambda t)^n$, so that $f_n = \lambda^n$. Generalizing, a sequence obeys a linear k-term recursion relation $f_{n+k} = m_1 f_{n+k-1} + \ldots + m_k f_n$ ($m_k \neq 0$) if and only if its generating function $F(t)$ satisfies the equation $F - f_{k-1} t^{k-1} - \ldots - f_0 = m_1 t\,(F - f_{k-2} t^{k-2} - \ldots - f_0) + \ldots + m_k t^k F$, or $[t^k \chi(t^{-1})]\cdot F = p(t)$, where $\chi(t) = t^k - m_1 t^{k-1} - \ldots - m_k$ and p is a polynomial of degree smaller than or equal to $k - 1$ defined via $f_0, \ldots, f_{k-1}$ and $m_1, \ldots, m_k$; this polynomial is linear on tuples $(f_0, \ldots, f_{k-1})$ and has the constant term f_0.[3] If $\lambda_1, \ldots, \lambda_l$ are the distinct roots of $\chi(t)$ and have multiplicities $k_1, \ldots, k_l$ ($k_1 + \ldots + k_l = k$), then the partial fraction expansion of $F = p(t)/[t^k \chi(t^{-1})]$ over the field $\mathbb{C}$ gives a linear combination with numerical coefficients of $(1 - \lambda_i t)^{-s}$: $i = 0, \ldots, l$, $s = 0, \ldots, k_i$. In turn, there are formal expansions such as $(1 - \lambda_i t)^{-s} = \sum (c_0 n^{s-1} + \ldots + c_{s-1})(\lambda_i t)^n$, with numerical coefficients c_j depending only on s and λ_i and such that $c_0 \neq 0$, which follows directly from the foregoing expansion of $(1 - \lambda t)^{-1}$ and the identity $d^{s-1}[(1 - \lambda t)^{-1}]/dt^{s-1} = (s - 1)!\ \lambda^{s-1}(1 - \lambda t)^{-s}$. This makes it possible to define a

[1] Readers experienced in analysis will probably notice that the generating function resembles a Laplace transform.

[2] The algorithm described subsequently is important for the design of recursive digital filters (Williams 1986).

[3] *How* would the equation for F change if sequences were assumed to obey the recursion relation for $n \geq n_f$, depending on the sequence?

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_11, © Springer Science+Business Media, LLC 2013

canonical basis in the vector space of all sequences of the preceding recursion relation[4] in terms of the distinct roots $\lambda_1,\ldots, \lambda_l$ of χ and their multiplicities $k_1,\ldots, k_l$. Over the field $\mathbb{C}$, this basis consists of $\{\lambda_1^n\},\ldots, \{n^{k_1-1}\lambda_1^n\},\ \ldots\ \ldots, \{\lambda_l^n\},\ldots, \{n^{k_l-1}\lambda_l^n\}$, and over $\mathbb{R}$ (when $m_1,\ldots, m_k$ are real), it includes $\{\lambda_i^n\},\ldots, \{n^{k_i-1}\lambda_i^n\}$ for the real roots λ_i and $\{\mathrm{Re}\lambda_i^n\},\ldots, \{n^{k_i-1}\mathrm{Re}\lambda_i^n\}$, $\{\mathrm{Im}\lambda_i^n\},\ldots, \{n^{k_i-1}\mathrm{Im}\lambda_i^n\}$ for the complex conjugate pairs $\lambda_i,\ \bar{\lambda}_i$. (*Why* are these sequences linearly independent?) Thus, *a sequence* $\{f_n\}$ *is a linear combination with numerical coefficients of elements of a canonical basis, related to fixed nonzero* $\lambda_1,\ldots, \lambda_l$ *and positive integers* $k_1,\ldots, k_l$ *if and only if it obeys the linear k-term recursion relation* $f_{n+k} = m_1 f_{n+k-1} + \ldots + m_k f_n$ *such that* $k = k_1 + \ldots + k_l$ *and* $\chi(t) = t^k - m_1 t^{k-1} - \ldots - m_k$ *is a polynomial of the distinct roots* $\lambda_1,\ldots, \lambda_l$, *with the multiplicities* $k_1,\ldots, k_l$, *respectively*. Also, using the fact that the aforementioned polynomial p has the constant term f_0, we can conclude that *a finite-dimensional vector space of sequences consists of all sequences obeying the same linear recursion relation if and only if a sequence remains in the space after the substitution of the generating function as* $F \mapsto (F - f_0)/t$, *which is equivalent to removing the first element,* f_0, *from the sequence*.[5]

For example, using $\lambda = 1$ of multiplicity k results in $\chi(t) = (t-1)^k$, and therefore, every polynomial sequence of degree smaller than or equal to $k-1$, $f_n = a_0 n^{k-1} + \ldots + a_{k-1}$, satisfies the linear k-term recursion relation with $m_q = (-1)^{q-1}\binom{k}{q}$, $q = 1,\ldots, k$.[6] Hence, $\sum_{l=0}^{k}(-1)^l\binom{k}{l}(n-l)^j = \begin{cases} 0 & \text{if} \quad k, \\ k! & \text{if} \quad k = j \end{cases}$ for integers $n \geq k \geq j \geq 0$.[7]

For a sequence of a linear k-term recursion relation, the coefficients with respect to the canonical basis can be obtained by decomposing the first k terms of the sequence with respect to the corresponding terms of elements of the basis. Consider, for example, a generalized Fibonacci sequence $f_{-1}, f_0, f_1, f_2,\ldots$ defined by the initial conditions $f_{-1} = \xi$, $f_0 = \eta$ and the recursion relation $f_n = m\, f_{n-1} + f_{n-2}$ for $n > 0$ [usually, m, ξ, and η are integers and $m > 0$, so that

[4] It is a k-dimensional space isomorphic to the space of all k-tuples $(f_0,\ldots, f_{k-1})$.

[5] It can also be proved without using generating functions. (*How*?) A "continuous" analog, related to functions and ordinary linear differential equations instead of discrete sequences and linear recursion relations, was discussed previously in sections P10.20, P10.21 and P10.22 ("One-Parameter Groups of Linear Transformations" problem group). To learn more about the close ties between linear recursion relations and autonomous ordinary linear differential equations see Arnol'd (1975).

[6] m_q can also be found by a different method – from a system of linear equations with a Vandermonde determinant as $h_k^{\,j} = m_1 h_{k-1}^{\,j} + \ldots m_k h_0^{\,j}$: $j = 0,\ldots, k-1$, using $h_l = l$, $\forall l = 0, \ldots, k$; the solution of the system, $m_{k-i} = \prod_{s,\, s\neq i}(h_k - h_s)/(h_i - h_s) = (-1)^{k-1-i}\binom{k}{i}$, $i = 0,\ldots, k-1$, is obtained by Cramer's rule taking into account the explicit expression for the Vandermonde determinant (see, e.g., section E1.1, the "Jacobi Identities and Related Combinatorial Formulas" problem group above), by means of the Jacobi identities (section P1.1), or using the fact that m_{k-i}, as a function of h_k, is a **Lagrange interpolating polynomial** of degree $k-1$ in h_k becoming δ_{ij} for $h_k = h_j$, $j = 0,\ldots, k-1$. See also a proof in Zalenko (2006).

[7] For $k = j$ the result is implied by the related Jacobi identity from section P1.1 using $x_l = n - l$, $l = 0,\ldots, k$. (Similarly, the result can be deduced from the identities of section P1.1 for any $k \geq j$; readers can also formulate the results for $k < j$ and for negative j, which can be derived from the identities of sections P1.5 and P1.7, respectively. *Fill* in the details.) In addition, readers may prove these identities differently using combinatorial arguments.

$(m^2 + 4)^{1/2}$ is irrational; $m = 1$, $\xi = 0$, and $\eta = 1$ are used most frequently]. The generating function $F(t) = \sum f_{n-1}t^n$ is sought from the equation $F - \eta t - \xi = mt\,(F - \xi) + t^2F$, or $(1 - mt - t^2)\cdot F = (\eta - m\xi)\,t + \xi$, which gives $f_n = \frac{\xi+\eta\lambda_+}{\lambda_+-\lambda_-}\lambda_+^n - \frac{\xi+\eta\lambda_-}{\lambda_+-\lambda_-}\lambda_-^n$, where $\lambda_\pm = ½\,[m \pm (m^2 + 4)^{1/2}]$. Using $\xi = 0$ and $\eta = 1$, $f_n = ({\lambda_+}^{n+1} - {\lambda_-}^{n+1})/(\lambda_+ - \lambda_-) = ({\lambda_+}^{n+1} - {\lambda_-}^{n+1})/(m^2 + 4)^{1/2}$.

Various problems of combinatorial analysis and probability theory are successfully explored with such powerful tools as generating functions. In this problem group readers will encounter several problems related to the combinatorics of binomial coefficients (sections P11.1, P11. 2, P11. 3, P11. 4 and P11. 5), the theory of partitions (sections P11.6*, P11.7*, P11.8*, P11.9*, P11.10* and P11.11*), and renewal processes in probability theory (sections P11.12**, P11.13**, P11.14** and P11.15**), which can be explored using generating functions. We would like to emphasize that the problems in sections P11.1, P11.2, P11.3, P11.4, P11.5, P11.6*, P11.7*, P11.9*, P11.10* and P11.11* does not require an ability to deal with power series (likewise, the problems in sections P11.6*, P11.7*, P11.8*, P11.9*, P11.10* and P11.11* assume no familiarity with the theory of partitions) and can be solved by readers with limited experience.[8]

P11.1

Establish identities

$$\sum_{i=l}^{m} (-1)^{i-l} \binom{m+1}{i+1} \cdot \binom{i}{l} = 1, \text{ for natural numbers } m \geq l\,.$$

P11.2

Establish the identities

$$\sum_{i=a}^{s-b} \binom{s-i}{b} \cdot \binom{i}{a} = \binom{s+1}{a+b+1}, \quad \text{for natural numbers } s \geq a+b\,.$$

[8] However, readers familiar with these techniques will be able to find shorter solutions – due to more efficient use of generating functions!

P11.3

We will need the multi-index notation for the problems below. A multi-index is an ordered set of natural numbers (including zero) of a fixed (finite) dimension $\alpha = (\alpha_1,\ldots,\alpha_l)$; its norm is $|\alpha| = \sum\alpha_i$. For a numerical sequence $y = (y_1,\ldots,y_l)$ of length l we define $y^\alpha := \prod {y_i}^{\alpha_i}$. For α, β of equal dimensions, $\alpha \geq \beta$ if $\alpha_i \geq \beta_i$ for each component; also, we use the notation $\binom{\alpha}{\beta} := \prod \binom{\alpha_i}{\beta_i}$.[9] *Establish* the following generalization of identities in section P11.2:

$$\sum_{\beta \geq \alpha,\ |\beta|=s} \binom{\beta}{\alpha} = \binom{s+n-1}{|\alpha|+n-1},\ \forall s \geq 0,\ \forall n>0,\ \forall \alpha \textit{ of } \text{dimn and } \textit{norm} \leq s.$$

State and *prove* a similar identity having $\sum_{\beta \geq \alpha,\ |\beta| \leq s} \binom{\beta}{\alpha}$ on the left-hand side.

P11.4

Establish a generalization of identities in section P11.1:

$$\sum_{\beta \geq \alpha,\ |\beta| \leq |\alpha|+k} (-1)^{|\beta|-|\alpha|} \binom{|\alpha|+k+n}{|\beta|+n} \binom{\beta}{\alpha} = 1,\ \forall k \geq 0,\ \forall n>0,\ \forall \alpha \text{ of dim } n$$

(section P11.1 is related to $n = 1$).

P11.5

Establish a generalization of section P11.4:

$$\sum_{\beta \geq \alpha,\ |\beta| \leq |\alpha|+k} (-1)^{|\beta|-|\alpha|} \binom{|\alpha|+k+n+m}{|\beta|+n+m} \binom{\beta}{\alpha} = \binom{k+m}{m},\ \forall k,m \geq 0,\ \forall n>0,\ \forall \alpha \text{ of dim } n\,.$$

[9] In particular, this symbol vanishes when α is not greater than or equal to β (i.e., $\beta_i > \alpha_i$ for some i).

P11.6[*]

Denote by $p_{n,k}(m)$ the number of k-element subsets of $\{0, \ldots, n-1\}$ with the sum of all elements in the subset equal to m.[10] *Establish* the following properties of the generating series $P_{n,k}(t) := \sum_{m=0}^{\infty} p_{n,k}(m)\, t^m$:

- For any n and k, the series $P_{n,k}$ terminates (is a polynomial; of what degree?);
- $P_{n,0} = 1$ and $P_{n,k} = 0$ when $k > n$;
- The polynomial $P_{n,k}$ is divided by $t^{(k-1)k/2}$;
- The polynomial $t^{-(k-1)k/2} P_{n,k}$ is reciprocal [a polynomial $a_d t^d + a_1 t^{d-1} + \ldots. + a_0$ ($a_d \neq 0$) with real coefficients is called reciprocal if $a_i = a_{d-i}$, $\forall i = 0,\ldots, d$];
- Coefficients of $P_{n,k}$ are stabilized with growing n [that is, the numbers $p_{n,k}(m)$ cease to depend on n when n is large enough for fixed k and m].

P11.7[*]

Establish the formula

$$P_{n,k}(t) = t^{(k-1)k/2} \cdot \frac{(1-t^{n-k+1})\ldots(1-t^n)}{(1-t)\ldots(1-t^k)}, \quad k>0.$$

P11.8[*]

Coefficients of a polynomial $\sum_{m=0}^{k(n-k)} p_{n,k}\left(\frac{(k-1)k}{2}+m\right) t^m = t^{-(k-1)k/2} P_{n,k}(t) = \frac{(1-t^{n-k+1})\ldots(1-t^n)}{(1-t)\ldots(1-t^k)}$ will be the same as of the **Euler series** (or the **Euler function**)

$$\varphi(t) := \prod_{i=1}^{\infty} (1-t^i)^{-1} = \sum_{m=0}^{\infty} p(m)\, t^m = 1 + t + 2t^2 + 3t^3 + 5t^4 + 7t^5 + \ldots.$$

while $m \leq \min(n-k, k)$. (*Prove.*)

[10] In the terminology of combinatorial analysis, $p_{n,k}(m)$ is the number of partitions of a natural number m into k distinct natural summands, not exceeding $n-1$.

It is easy to establish that $p(0) = 1$ and for $m > 0$, $p(m)$ is the number of all partitions of m into (positive) natural summands (partitions differing only in the order of summands are considered to be the same partition). One can quickly obtain the coefficients $p(m)$ of the Euler series from recursion equations:

$$\begin{aligned}
&p(0) = 1,\\
&p(1) \ - p(0) = 0,\\
&p(2) \ - p(1) \ - p(0) = 0,\\
&p(3) \ - p(2) \ - p(1) = 0,\\
&\ldots\\
&p(m) \ - p(m-1) - p(m-2) + p(m-5) + p(m-7) \ - \ \ldots = 0.\\
&\ldots
\end{aligned}$$

These equations are obtained from the functional equation

$$1 = \prod_{i=1}^{\infty}(1-t^i)\cdot\varphi(t) = (1 - t - t^2 + t^5 + t^7 - \ldots)\cdot\big(p(0) + p(1)t + p(2)t^2 + \ldots\big).$$

The identity used in the preceding equation is the famous **Euler identity**:

$$\prod_{i=1}^{\infty}(1-t^i) = 1 - t - t^2 + t^5 + t^7 - \ldots = 1 + \sum_{i=1}^{\infty}(-1)^i\left(t^{\frac{3i^2-i}{2}} + t^{\frac{3i^2+i}{2}}\right).$$

This identity can be proved in an elementary way using the bijection between partitions such that the number of summands differs by 1 (this proof was discovered by F. Franklin). (Try to derive this proof yourself or find it in the literature listed below.) However, there are remarkable identities for degrees of the Euler function $[\varphi(t)]^n$ with selected $n > 1$ (namely, n = 3,8,10,14,15,21,24,26,28,35,36,...) discovered by C.F. Gauss, C.G.J. Jacobi, F. Klein, and other scientists, proofs of which are many times more complicated! The subject is very deep and has a far-reaching further development; we will indicate three directions of it: **combinatorial analysis**, **number theory,** and **homological algebra**. Readers are encouraged to familiarize themselves with the Euler identity and related topics, as presented in Franklin (1881), Polya (1954), Hall (1967), Stanley (1999), Andrews (1976), Andrews and Eriksson (2004), Hardy and Wright (1960), Rademacher (1973), Niven and Zuckerman (1973), Gupta (1980), Grosswald (1983), Apostol (1990), Dyson (1972), Fuchs (1981, 1984), MacDonald (1972, 1979), Kac (1978, 1994), and references therein.

*P11.9**

Write out the polynomials $\overline{P}_{n,q}(t) := t^{-(q-1)q/2}P_{n,q}(t)$ for prime q = 2,3,5 and $n = 2q$,3q in the usual form $\sum a_m t^m$. {To make your computation short, use section P11.7 to find higher coefficients and reciprocity for lower ones; also, you may write out $\bar{P}_{n,q}(t)$ for a larger number of primes q and $n = 2q, 3q, 4q, 5q,\ldots$ using a computer.} In all those $\bar{P}_{n,q}(t)$, sum the coefficients whose indices have equal remainders modulo q. *Formulate*, based on what you have found, the general result in this direction. *Compare* your statement with that in section P11.10*.

P11.10[*]

Establish that for a prime q, the number of q-element subsets of $\{0, \ldots, kq - 1\}$, with the sums of elements having the remainder i modulo q, is equal to

$$\frac{\binom{kq}{q} - k}{q}, \text{ when } q>2 \,\&\, i>0 \text{ or } q = 2 \,\&\, i = 0,$$

$$\frac{\binom{kq}{q} - k}{q} + k, \text{ when } q>2 \,\&\, i = 0 \text{ or } q = 2 \,\&\, i = 1.$$

P11.11[*]

Prove that for a prime q and $k > 1$,

$$\sum_{l=1}^{k} (-1)^{k-l} \binom{k}{l} \cdot \binom{lq}{q} \equiv 0 \mod q^k .$$

(The left-hand side becomes zero when $k > q$; *why*?)

P11.12[**]

This and the next three problems are related to the renewal processes in probability theory.[11] Let $f = \{f_n\}_{n = 0,1,\ldots}$ be a nonnegative sequence with $f_0 = 0$ and $s_f = \sum f_n \leq 1$. Define a sequence $u = \{u_n\}_{n = 0,1,\ldots}$ by the recursion relations

$$u_0 = 1, \quad u_n = \sum_{k=0}^{n} f_k \, u_{n-k} \ (n>0).$$

Prove that $s_f < 1$ if and only if $s_u := \sum u_n < \infty$, and in this case $s_f = (s_u - 1)/s_u$.

[11] Here we are interested in a quite elementary combinatorial aspect. The probabilistic meaning and a far-reaching development are exhaustively discussed in Feller (1966) and references therein.

P11.13[**]

We will call a **period** of a nonnegative-element series $\sum f_n$ the greatest common divisor of indices n with $f_n > 0$ and call the series **periodic** (**nonperiodic**) if the period is greater than (resp. equal to) 1. *Prove* that under conditions of the previous problem, $\lim\limits_{n\to\infty} u_n = 0$ if $s_f < 1$ and $\lim\limits_{n\to\infty} u_n = \mu^{-1}$ if $\sum f_n$ is nonperiodic and $s_f = 1$, where $\mu = \sum n f_n$ (in particular, $u_n \to 0$ for $\mu = \infty$).

P11.14[**]

(Generalization of section P11.12** and P11.13**.) Let $a = \{a_n\}_{n\,=\,0,1,\ldots}$ be a nonnegative sequence with $a_0 < 1$, and let $\sum b_n$ be a convergent series with nonnegative elements b_n some of which are strictly positive. Define a sequence $u = \{u_n\}_{n\,=\,0,1,\ldots}$ by the recursion relations

$$u_n = b_n + \sum_{k=0}^{n} a_k\, u_{n-k} \quad (n \geq 0)$$

(more explicitly, $u_0 = \frac{b_0}{1-a_0}$ and $u_n = \frac{b_n + a_1 u_{n-1} + \ldots + a_n u_0}{1-a_0}$, $n>0$). *Prove* the following claims (1)–(4):

1. The sequence u is bounded if and only if $s_a := \sum a_n \leq 1$.
2. For $s_a = 1$, u does not tend to zero if $a_0 + \ldots + a_N = 1$ and $a_{N+1} = a_{N+2} = \ldots = 0$ for some natural N.
3. The series $\sum u_n$ converges if and only if $s_a < 1$, and in this case $s_u = s_b/(1 - s_a)$ ($s_b = \sum b_n$).
4. If the series $\sum a_n$ is nonperiodic, then:
   - For $s_a = 1$, $\lim\limits_{n\to\infty} u_n = s_b \mu^{-1}$, with $\mu = \sum n a_n$ (in particular, $u_n \to 0$ for $\mu = \infty$),
   - For $s_a > 1$ (including $s_a = \infty$), if the sequence a is bounded, an equation $A(t) = 1$, with the generating function of a on the left-hand side, has a unique positive root $t_0 < 1$, and in this case $u_n \sim \frac{B(t_0)}{A'(t_0)} \cdot t_0^{-n}$ [here $B(t)$ is the generating function of $\{b_n\}$ and the symbol ~ means that the ratio of its left- and right-hand sides tends to 1].[12]

[12] In particular, u grows by an exponential law. Note that the derivative $A'(t_0)$ is finite. [*Why*? Use arguments similar to those in section E11.13, using convergence of the power series $A(t)$ for $|t| < 1$.]

P11.15**

Let, under the conditions of section P11.14**, $\sum a_n$ be a series of period $\lambda > 1$. Define a subsequence $a^{(0)} := \{a_{n\lambda}\}_{n=0,1,\ldots}$ of the sequence $\{a_n\}$ and for $j = 0, \ldots, \lambda - 1$, subsequences $u^{(j)} := \{u_{n\lambda+j}\}_{n=0,1,\ldots}$, $b^{(j)} := \{b_{n\lambda+j}\}_{n=0,1,\ldots}$ of the sequences $u = \{u_n\}$, $b = \{b_n\}$, respectively. *Prove* that three sequences $a^{(0)}$, $b^{(j)}$, $u^{(j)}$ satisfy the recursion relations from section P11.14**, substituting a, b, u, respectively. *Deduce* from it that $\lim\limits_{n\to\infty} u_{n\lambda+j} = \lambda\, s_{b^{(j)}}\mu^{-1}$ for $s_a = 1$, where $s_{b^{(j)}} = \sum b_n^{(j)}$ (which is called the **asymptotic periodicity** of the sequence u).

Hint

H11.0

If a sequence obeys a recursion relation $f_{n+k} = m_1 f_{n+k-1} + \ldots + m_k f_n$ ($m_k \neq 0$) for $n \geq n_f$, then its generating function satisfies the equation $\left[t^k \chi(t^{-1})\right] \cdot \left(F - \sum\limits_{n=0}^{n_f-1} f_n t^n\right) = t^{n_f} p(t)$, where $\chi(t)$ is the same as in section P11.0 and p is a polynomial of degree less than or equal to $k - 1$, defined by using $f_{n_f}, \ldots, f_{n_f+k-1}$ and $m_1, \ldots, m_k$; this polynomial is linear on a tuple $\left(f_{n_f}, \ldots, f_{n_f+k-1}\right)$ and has the constant term f_{n_f}.

The first method to prove the linear independence of the sequences belonging to the canonical basis (so that the "canonical basis" really is a basis): $\chi(t)$ is the characteristic polynomial for the matrix $\begin{pmatrix} m_1 & \cdots & \cdots & m_k \\ 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ & & 1 & 0 \end{pmatrix}$ which has the type discussed in sections S7.10 and S10.21 (resp. from the "2×2 Matrices That Are Roots of Unity" and "One-Parameter Groups of Linear Transformations" problem groups discussed previously). When they are considered vectors in $\mathbb{R}^k$, the tuples of the first k elements of the sequences that belong to the canonical basis form a Jordan basis for this matrix; thus, these k-tuples are linearly independent. (The independence follows from the proposition in section S10.21 because this matrix has a single Jordan box in the complexified space corresponding to any eigenvalue; we leave it to the reader to work out all the details.) Hence, the sequences belonging to the canonical basis are linearly independent.

Second method: let some linear combination of elements of the canonical basis with numerical coefficients equal zero. Using an identity $A\cos\theta + B\sin\theta = C\cos(\theta - \theta_0)$ $\left(C = \sqrt{A^2 + B^2},\ \cot\theta_0 = \frac{A}{B}\right)$, this linear combination can be transformed into a linear combination of sequences with common terms $n^s \rho_i^n \cdot \cos(\omega_i n - \theta_{0i})$ ($\rho_i = |\lambda_i|$, $\omega_i = \arg\lambda_i$, $s = 0, \ldots, k_i - 1$, and θ_{0i} is not an odd multiple of $\pi/2$ if ω_i is a multiple of 2π). (In the proof of Proposition 2 in section E10.21, a similar transformation was applied to quasipolynomials.) If ω_i is commensurable with 2π, then $|\cos(\omega_i n - \theta_{0i})| \geq \cos d > 0$ for infinitely many n, where d is the distance

between θ_{0i} and the closest element (distinct from itself) of a cyclic group as generated by ω_i and 2π. If ω_i is incommensurable with 2π, then $\omega_i n - \theta_{0i}$ can become arbitrarily small (as discussed previously in section H10.13.D, the "One-Parameter Groups of Linear Transformations" problem group), so that $|\cos(\omega_i n - \theta_{0i})| > ½$ for infinitely many n. Hence, $\cos(\omega_i n - \theta_{0i})$ is bounded away from zero for infinitely many n. Using this fact, we can eliminate from the linear combination all elements having $\rho_i \neq 1$ or $s > 0$ by induction based on their different growths as $n \to \infty$: we make the term-by-term division by a current maximum of ${\rho_i}^n$ and by n^s, where s equals the maximal $k_j - 1$ for all j such that ρ_j are equal to this maximal ρ_i. Thus, the linear independence problem can be reduced to a special case where all λ_i are simple (nonmultiple) roots of the characteristic polynomial χ and the desired result follows from the nondegeneracy of the Vandermonde matrix $({\lambda_i}^n)_{i\,=\,1,\ldots,\,k,\ n\,=\,0,\ldots,\,k-1}$; we leave it to the reader to fill in all details. (***Warning:*** C and θ_0 in the foregoing deduction are functions of n).

Third method: readers who can use convergent power series may proceed as follows. The correspondence between sequences and their generating functions is a linear isomorphism, so let us prove the linear independence of the generating functions. An inductive argument shows that a sequence obeying a linear recursion relation cannot grow faster than at a geometric progression, that is, $|\lambda|^n$, such that $|\lambda|^k \geq |m_1\lambda^{k-1}| + \ldots + |m_k|$; therefore, the generating function is given by a power series having nonzero convergence radius. In particular, for the elements of a canonical basis these series converge to rational functions $(1 - \lambda_i t)^{-s}$: $i = 0,\ldots, l$, $s = 0,\ldots, k_i$, respectively. Thus, we must prove the linear independence of these functions on a common domain $\{t: |t| < \min |\lambda_i|^{-1}\} \subset \mathbb{C}$. Let some linear combination with numerical coefficients of these functions be equal to zero; we can eliminate these functions from this combination by an inductive procedure based on their different behavior near the point $t = {\lambda_i}^{-1}$ having the minimal modulus over the λ_i that remained after the previous step. (We leave it to the reader to work out the details.)

A proof of the fact that a finite-dimensional vector space of sequences consists of all sequences obeying the same linear recursion relation if and only if a sequence remains in the space after removal of the first element from the sequence without using generating functions. A sequence obeys a recursion relation starting from any element, and after removing the first element it will remain in the space. Conversely, sequences $\{f_0, f_1, f_2, \ldots\}$, $\{f_1, f_2, \ldots\}, \ldots, \{f_n, \ldots\}, \ldots$ belonging to a finite-dimensional space cannot be linearly independent, and therefore $\{f_n\}$ obeys a linear recursion relation. The space of all sequences obeying the shortest of these relations is generated by the foregoing sequences (why?), and hence, it is a subspace of the considered space. If this subspace coincides with the whole space, then the proof is complete. If not, we must take a sequence $\{g_n\}$ that does not belong to the subspace and build a similar subspace with $\{g_n\}$. The sum of these subspaces possesses a canonical basis that is the union of the canonical bases of the summands; thus it is the space of all sequences obeying some linear recursion relation. Therefore, the proof may be completed by induction on the spatial dimension. Note that *the shortest recursion relation is the same for a sequence and for the same sequence with* the *first element removed*. (We leave it to the interested reader to fill in all the details.)

H11.1

For fixed m, introduce the generating function $\sum_{l=0}^{m} a_l t^l$, where a_l denotes the left-hand sides of the identities to be proved. Changing the order of summation, show that all coefficients equal 1.

H11.2

Show that the left-hand side depends on $a + b$, but not on a and b separately; then confirm identity by an elementary computation, replacing a by 0 and b by $a + b$.

H11.3

Apply induction on n: for $n = 1$ the equality is trivial, and the transition of induction is performed using the identities in section P11.2. Alternatively, one may argue as follows: the left-hand side depends only on $/\alpha/$ (which can be proved as in section H11.2, using a multigenerating function). Considering $\alpha = (0, \ldots, 0, a)$ one has

$$\sum_{\beta \geq \alpha,\, |\beta| = s} \binom{\beta}{\alpha} = \sum_{b=a}^{s} \sum_{\beta_1, \ldots, \beta_{n-1} \geq 0,\, \sum_{i=1}^{n-1} \beta_i = s-b} \binom{b}{a} = \sum_{b=|\alpha|}^{s} \binom{s-b+n-2}{n-2} \binom{b}{|\alpha|}$$

$$= \binom{s+n-1}{|\alpha|+n-1},$$

by virtue of the identities in section P11.2. The summation on $|\beta|$ yields

$$\sum_{\beta \geq \alpha,\, |\beta| \leq s} \binom{\beta}{\alpha} = \binom{s+n}{|\alpha|+n}, \ \forall s \geq 0, \ \forall n > 0, \ \forall \alpha \text{ of dim } n \text{ and } \mathit{norm} \leq s\,.$$

H11.4

Using the identities in section P11.3, reduce the identities in section P11.4 to their special case in section P11.1.

H11.5

Apply induction on $k + m$. Under $m = 0$, the left-hand sides are equal to unity by virtue of the identities in section P11.4; and under $k = 0$, they contain only one term with $\beta = \alpha$ and are also unity. Therefore, the values of the left-hand sides fill the

Pascal triangular table. (Why?) On the other hand, values of the right-hand sides fill the same table in the same sequence.

H11.6

The degree of $P_{n,k}(t)$ is equal to $(n-k)+\ldots+(n-1)=(2n-k-1)\ k/2$. $P_{n,k}(t)$ is divisible by $t^{(k-1)k/2}$ since the minimal sum of the elements of a k-element subset is $0+\ldots+(k-1)=(k-1)\ k/2$. The reciprocity of $t^{-(k-1)k/2}P_{n,k}$ is bound with the involution on the set of k-element subsets of $\{0,\ldots,n-1\}$[13]:

$$\{x_1,\ldots,x_k\}\ \mapsto \{n-1-x_1,\ldots,n-1-x_k\} \qquad (x_1<\ldots<x_k),$$

which transforms the sum of the elements into $s\mapsto\ (2n-k-1)\ k/2-s$.

H11.7

Establish the recursion formula

$$P_{n,k}(t)\cdot(t^k-1)=(t^n-1)t^{k-1}\cdot P_{n-1,k-1}(t)$$

$(k>0)$, which is equivalent to the explicit formula being proved.

H11.10

Consider the classes of q-element subsets with equal remainders of the sums of elements modulo q. Denote them by $R_0,\ldots,R_{q-1}$. We must establish that all, except one, R_i have the same number of elements, but the exceptional class is larger by k. We must find that class. To do this, represent the set $M=\{0,\ldots,kq-1\}$ as a disjunctive union of its subsets $M_l=\{(l-1)q,\ldots,lq-1\}, l=1,\ldots,k$. Let, for $L\subseteq\{1,\ldots,k\}$, $R_{i,L}$ be the subclass of R_i consisting of the q-element subsets of M that intersect with M_l

[13] An involution is an automorphism of a set equal to its own inversion.

if and only if $l \in L$. Thus, we have the disjunctive unions $R_i = \bigcup_{L \subseteq \{1,\ldots,k\}} R_{i,L}, \quad \forall i$.
Now, for a given $L \subseteq \{1,\ldots,k\}$, define a cyclic group of transformations of $\bigcup_{l \in L} M_l$

$$G = G_L := \{\sigma, \sigma^2, \ldots, \sigma^q = e\} \cong \mathrm{Z}(q),$$

where σ permutes points of M_l with the least index $l \in L$ by the formula

$$\sigma(x) \in M_l, \quad \sigma(x) \equiv (x+1) \bmod q \quad (x \in M_l)$$

and leaves the elements of $\bigcup_{j \in L \setminus \{l\}} M_j$ fixed. G permutes the elements of $\bigcup_{i=0}^{q-1} R_{i,L}$, thereby acting on the union.[14] Investigate the behavior of orbits of the induced action.[15] How many times do the orbits $Gs = \{\sigma s, \ldots, \sigma^q s = s\}$ $\left(s \in \bigcup_{i=0}^{q-1} R_{i,L}\right)$ intersect with a class $R_{i,L}$, for a fixed $i \in \{0, \ldots, q-1\}$, when $\#L > 1$, and when $\#L = 1$?

H11.11

Represent the set $M = \{0, \ldots, kq-1\}$ as a union of $M_l = \{(l-1)q, \ldots, lq-1\}$, as in section H11.10. Also, let K be the class of all q-element subsets of M that intersect with all M_l. (How many elements does K consist of?) Consider cyclic groups of transformations of M, $G_l \cong \mathbb{Z}(q)$, $l = 1,\ldots, k$, that permute points of, respectively, M_l in the same way as in section H11.10 and leave the points of the remaining M_j fixed. Consider the induced action of the direct product $G = \prod G_l$ on K. Show that K is invariant with respect to this action (K consists of entire orbits of G) and that all orbits in K are **regular** (q^k-element).

H11.12

We can prove this by applying the relationship between the generating functions of the sequences f and u, respectively, $F(t) = \sum f_n t^n$ and $U(t) = \sum u_n t^n$.

[14] This action is referred to as an **induced** action.

[15] Orbits of an action of a group H on a set S are equivalence classes with respect to the equivalence relation $s_1 \equiv s_2 \Leftrightarrow \exists\ h \in H\colon hs_1 = s_2$. An equivalent definition is that the orbits are the sets $Hs = \{hs\colon h \in H\}$ $(s \in S)$.

H11.13

According to the relationship between the generating functions $F(t)$, $U(t)$, as obtained in section E11.12, $U(t) = (1 - F(t))^{-1}$ for $| t | < 1$. Moving on to the limit as $t \to 1 - 0$ shows that the series $\sum u_n = U(1)$ converges when $s_f < 1$, and so $u_n \to 0$. Calculating the limit of u_n for $s_f = 1$, pay attention to the fact that $\mu = \sum n f_n$ is the derivative of F at $t = 1$, $\mu = F'(1)$. By the relationship between $F(t)$ and $U(t)$,

$$(1 - t) \cdot U(t) = \tfrac{1-t}{1-F(t)} \quad (|t|<1).$$

oving on to the limit when $t \to 1 - 0$, using l'Hôpital's rule on the right-hand side yields

$$\lim_{t \to 1-0} \lim_{n \to \infty} [U_n(t) - t \cdot U_{n-1}(t)] = \lim_{t \to 1-0} (1 - t) \cdot U(t) = \mu^{-1},$$

where $U_n(t) = \sum_{k=0}^{n} u_k t^k$ are partial sums of series $U(t)$. If the limits are permutable with one another, the left-hand side will be equal to $\lim_{n\to\infty} [U_n(1) - U_{n-1}(1)] = \lim_{n\to\infty} u_n$.

One could think that to complete the solution, we need to substantiate the permutability of those limits. But in fact this would be incorrect reasoning because so far we have not considered the nonperiodicity of the series $\sum f_n$. And the importance of this condition is obvious because a similar statement is invalid in the periodic case (as will emerge subsequently in section P11.15**)! We suggest proceeding as in Feller (1966). Start from an easy observation followed by induction from the recursion relations $0 \le u_n \le 1$. In particular, u_n have their upper and lower limits lying in the segment [0,1]. Show that both of them are equal to μ^{-1}. (***Warning***: in the periodic case these limits will be distinct unless $\mu = \infty$.) Do this in the following steps:

Step 1. Prove that if a subsequence $\{u_{n_k}\}_{k=0,1,\ldots}$ tends to the upper (lower) limit, then for any j, such that $f_j > 0$, $u_{n_k - j}$ will tend to the same limit.

Step 2. Consider the tails (remainders) of the series $\sum f_n$, $r_n := f_{n+1} + f_{n+2} + \ldots$. Deduce from the recursion relations the identities

$$r_0 u_n + r_1 u_{n-1} + \ldots + r_n u_0 = 1, \qquad \forall n.$$

Next, establish an identity $\mu = \sum r_n$, that is, prove the following lemma.

Lemma 1. *For a convergent series $\sum a_n$ with nonnegative elements, the equality $\sum r_n = \sum n a_n$ with a series of the tails on the left-hand side, $r_n := \sum\limits_k a_k$,*[16] *holds including the case where one (and then both) of the sides is infinite.*

(An outline of the proof can be found subsequently in this section below. For further details, readers may turn to section E11.13.)

Step 3. Derive, using the result of step 1 and the identities established in step 2, that

$$\mu \cdot \underline{\lim}\, u_n \quad \geq \quad 1 \quad \geq \quad \mu \cdot \overline{\lim}\, u_n,$$

which will complete the solution of section P11.13**. At first, readers may derive these inequalities under the simplified assumption that $f_1 > 0$, then drop this simplification using the following lemma.

Lemma from number theory (J.J. Sylvester). *Let $k_1, \ldots, k_m$ ($m \geq 2$) be relatively prime positive integers, with at most one of them equal to one if $m > 2$. Then any positive integer exceeding $\prod k_i$ can be represented by a linear combination with positive integer coefficients of k_i.*

Proving this lemma is a fairly good exercise in the geometry of numbers. However, readers who are not yet advanced enough may look at the outline of the proof at the end of section H11.13.

Outline of proof of Lemma 1. The equality $\sum r_n = \sum n a_n$ can be derived using the commutative convergence of nonnegative-element series. This was discussed previously, including cases of infinities, in the "One-Parameter Groups of Linear Transformations" problem group (section S10.18) and can also be found in handbooks on analysis, e.g., Schwartz and Huet (1961) and Fichtengoltz (1970). Also, this equality can be established using the relationship between the generating functions $A(t) = \sum a_n t^n$ and $R(t) = \sum r_n t^n$.

A case of Lemma 1 corresponding to $\sum n a_n < \infty$ is a special case of a more general assertion holding independently of absolute convergence:

Lemma 1′. *If a series (with arbitrary a_n) $\sum n a_n$ converges, then $\sum a_n$ and the series of its tails $\sum r_n$ also converge and $\sum r_n = \sum n a_n$.*

[This is a fine lemma, and proving it by elementary methods is a nice exercise in mathematical analysis. Pay attention to the fact that convergences of $\sum a_n$ and $\sum r_n$ do not imply convergence of $\sum n a_n$, as appears from a simple example when $a_n := (-1)^n \left(\frac{1}{n+1} + \frac{1}{n+2}\right)$. Examples like this are similar to those of divergent double series, while the related iterated series converge.]

[16] Convergence of a series implies convergence of the tails.

Outline of proof of Sylvester's lemma. Start with the following two observations, the first of which, we believe, will prevent the reader from proceeding by false methods:

1. The product of k_i cannot be replaced by their least common multiple (provide examples).
2. The natural numbers starting from $\sum k_i$ can be represented by linear combinations with positive integer coefficients of k_i if some $k_i = 1$. With this, an induction shows that $\sum k_i \leq \prod k_i + 1$ if $k_i = 1$ only for one of i (which holds including $m = 1$ but may be broken when $k_i = 1$ for more than one i).

Considering the last observation, we may restrict ourselves to the case where all $k_i > 1$. Apply induction on m as follows:

When $m = 2$, a vector $\vec{k} = (k_1, k_2)$ of relatively prime components can be included as $\vec{k}^0$ in a basis $\vec{k}^0$, $\vec{k}^1$ of the integer grid $\mathbb{Z} \times \mathbb{Z} \subset \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$. (Why? Also, generalize this claim for any m). There exists a **dual basis** of the same grid consisting of the vectors $\vec{\kappa}^0, \vec{\kappa}^1$ such that $\left\langle \vec{\kappa}^i, \vec{k}^j \right\rangle = \kappa^i_1 k^j_1 + \kappa^i_2 k^j_2 = \delta_{ij}$. (Why? Also, generalize this claim for any m). Use this basis to show that the right triangle cut from the first (positive) quadrant by the straight line $\left\{ (\xi_1, \xi_2) : \left\langle \vec{\xi}, \vec{k} \right\rangle = 1 \right\}$ does not contain integer points except for the origin, and the integer points $\vec{\xi}^1$, $\vec{\xi}^2$ of this line lying in the interior of the second and *fourth quadrants*, respectively, that are closest to the origin are removed from one another by a distance of $\left|\vec{k}\right| = \sqrt{k_1^2 + k_2^2}$. (Produce a figure.) This the vectors $\vec{\eta}^1, \vec{\eta}^2$ parallel to $\vec{\xi}^1, \vec{\xi}^2$, respectively, having the endpoints (k_2,0), (0,k_1) have a common starting point (which is an integer point lying in the interior of the first quadrant). (Why?) Use this statement to define, by induction, a sequence of points $\vec{\zeta} = \vec{\zeta}^n$ in the interior of the first quadrant lying on the lines $\left\{ (\zeta_1, \zeta_2) : \left\langle \vec{\zeta}, \vec{k} \right\rangle = k_1 k_2 + n \right\}$ (n = 1,2,...), respectively. (Readers should make use of their figure.) **QED**.

$m \geq 2$. Establish an elementary inequality

$$d \cdot \left[k_1 + \prod_{i>1} (k_i/d) \right] \leq \prod_{i=1}^{m} k_i,$$

where d is the greatest common divisor of $k_2, \ldots, k_m$ ($m > 2$ and $k_i > 1$ for each i). Using this inequality we will find for $N > \prod_{i=1}^{m} k_i$ that $N = k_1 d + N_1 + d \cdot \prod_{i>1} (k_i/d)$ with $N_1 \in \mathbb{Z}^+ = \{1,2,\ldots\}$. By the inductive hypothesis, $k_1 d + N_1 = z_1 k_1 + z' d$ (z_1, $z' \in \mathbb{Z}^+$), so $N = z_1 k_1 + z_0 d$ with $z_0 = z' + \prod_{i>1} (k_i/d) > \prod_{i>1} (k_i/d)$. Applying

the inductive hypothesis once again yields $z_0 = \sum_{i>1} z_i \cdot (k_i/d)\,(z_i \in \mathbb{Z}^+)$ and, finally, $N = \sum_{i=1}^{m} z_i k_i$, **QED.**

H11.14

The boundedness (unboundedness) of u for $s_a \leq 1$ (resp. $s_a > 1$) is obtained by induction directly from recursion relations. To be exact, show that $(0 \leq)\ u_n \leq (1 - a_0)^{-1} \sum_{k=0}^{n} b_k$ for $s_a \leq 1$, and for $s_a > 1$, $u_{m+N}, \ldots, u_{m+2N-1} \geq qM$ when $u_m, \ldots, u_{m+N-1} \geq M$ and N is large so that $\sum_{n=1}^{N} a_n \geq q\,(1 - a_0)$ $(q > 1)$. A similar computation yields $u_n \geq M$, $\forall n \geq m$, if $u_n \geq M$ for $n = m, \ldots, m + N - 1$, and N is large enough so that $\sum_{n=0}^{N} a_n \geq 1$. Divergence of $\sum u_n$ for $s_a > 1$ is produced by the unboundedness of u. The remaining claims in section P11.14 are formal consequences of the relationships among the generating functions of the sequences a, b, and u, respectively, $A(t) = \sum a_n t^n$, $B(t) = \sum b_n t^n$, and $U(t) = \sum u_n t^n$, which are equivalent to the recursion relations from section H11.14 and of statements in sections P11.12 and P11.13. No additional fine analytic consideration is necessary.

H11.15

(Feller 1966). Show, using the relationship among generating functions $A(t)$, $B(t)$, $U(t)$ from section H11.14, that for any $j = 0, \ldots, \lambda - 1$ the terms of $U(t)$ of the indices $n\lambda + j$, $n = 0,1,\ldots$ depend only on the terms of $B(t)$ of the indices $m\lambda + j$, $m = 0,1,\ldots$. Explicitly, $U^{(j)}(t) = \frac{B^{(j)}(t)}{1 - A(t)}$, with $U^{(j)} := \sum_{n=0}^{\infty} u_{n\lambda+j}\, t^{n\lambda}$ and $B^{(j)} := \sum_{n=0}^{\infty} b_{n\lambda+j}\, t^{n\lambda}$. The change in variables $t^{\lambda} \to t$ makes it possible to establish the relationships among the generating functions of $a^{(0)}$, $b^{(j)}$, $u^{(j)}$, which are equivalent to the recursion relations being proved. Asymptotic periodicity now follows from section P11.14**. (Why?) Explain why λ appears in the asymptotic formula.

Explanation

E11.1

A generating function is

$$\sum_{l=0}^{m} t^l \cdot \sum_{i=l}^{m} (-1)^{i-l} \binom{m+1}{i+1} \cdot \binom{i}{l} = \sum_{i=0}^{m} \binom{m+1}{i+1} \cdot \sum_{l=0}^{i} \binom{i}{l} (-1)^{i-l} t^l$$
$$= \sum_{i=0}^{m} \binom{m+1}{i+1} \cdot \{t-1)^i$$
$$= (t-1)^{-1} \cdot \left\{ [(t-1)+1]^{m+1} - 1 \right\} = \sum_{l=0}^{m} t^l,$$

and therefore, all $a_l = 1$.

E11.2

Dependence of the left-hand side on only $a + b$ may be established by various methods. For example, the equality

$$\sum_{i=a}^{s-b} \binom{s-i}{b} \cdot \binom{i}{a} = \sum_{i=a+1}^{s-b+1} \binom{s-i}{b-1} \cdot \binom{i}{a+1}$$

may be checked directly, after a series of elementary computations with binomial coefficients. Proceeding by a different method, apply the double-generating function

$$\sum_{a,b \geq 0, a+b \leq s} f(a,b) \cdot t^a u^b = \sum_{i=0}^{s} \sum_{a=0}^{i} \binom{i}{a} t^a \cdot \sum_{b=0}^{s-i} \binom{s-i}{b} u^b$$
$$= \sum_{i=0}^{s} (t+1)^i (u+1)^{s-i},$$

where $f(a,b) = \sum_{i=a}^{s-b} \binom{s-i}{b} \binom{i}{a}$. We must prove that the coefficients at the monomials $t^a u^b$ depend only on $a + b$, or, in different terms, that $f(a - 1,b) = f(a,b - 1)$. Prove this using the obtained explicit expression for the generating function.

E11.4

Writing the left-hand side of identities from section P11.4 which are being proved as $\sum\limits_{s:\ |\alpha|\le s\le|\alpha|+k}(-1)^{s-|\alpha|}\binom{|\alpha|+k+n}{s+n}\sum\limits_{\beta\ge\alpha,\ |\beta|=s}\binom{\beta}{\alpha}$ and using identities from section P11.3 for the internal sum brings the identities in section P11.4 to those in section P11.1.

E11.7

Multiplication of $P_{n,k}(t)$ by the binomial t^k-1 yields

$$\begin{aligned}P_{n,k}(t)\cdot(t^k-1)&=\sum_{A\subseteq\{0,\ldots,n-1\},\ \#A=k}t^{\sum\limits_{x\in A}x}\cdot(t^k-1)\\&=\sum_{A\subseteq\{0,\ldots,n-1\},\ \#A=k}\left(t^{\sum\limits_{x\in A}(x+1)}-t^{\sum\limits_{x\in A}x}\right)\end{aligned}$$

($k>0$). The terms on the right-hand side that correspond to those subsets of $\{0,\ldots,n-1\}$ that contain neither 0 nor $n-1$ will annihilate each other. (Why?) Thus, we will obtain

$$\begin{aligned}P_{n,k}(t)\cdot(t^k-1)&=\sum_{A\subseteq\{0,\ldots,n-1\},\ \#A=k,\ n-1\in A}t^{\sum\limits_{x\in A}(x+1)}\\&\quad-\sum_{A\subseteq\{0,\ldots,n-1\},\ \#A=k,\ 0\in A}t^{\sum\limits_{x\in A}x}=(t^n-1)\sum_{A\subseteq\{1,\ldots,n-1\},\ \#A=k-1}t^{\sum\limits_{x\in A}x}\\&=(t^n-1)\,t^{k-1}\sum_{A\subseteq\{0,\ldots,n-2\},\ \#A=k-1}t^{\sum\limits_{x\in A}x}=(t^n-1)\,t^{k-1}P_{n-1,k-1}(t)\,.\end{aligned}$$

E11.10

Since G is a group of a prime order, the orbits consist of either q or 1 element of $\bigcup\limits_{i=0}^{q-1}R_{i,L}$ (being, respectively, **regular** or **singular**). Any orbit is regular and intersects only once with all $R_{i,L}$ when $\#L>1$. (Why?) For $\#L=1$ the orbits are

singular. The corresponding k q-element subsets of M (which ones?) have sums of elements of the remainder 0 (1) modulo q if $q > 2$ (resp. $q = 2$) so that R_0 (resp. R_1) is larger by k than the remaining R_i.

E11.11

By the principle of inclusion–exclusion, K consists of

$$\#K = \sum_{l=0}^{k} (-1)^l \binom{k}{l} \cdot \binom{(k-l)\,q}{q} = \sum_{l=1}^{k} (-1)^{k-l} \binom{k}{l} \cdot \binom{lq}{q}$$

elements. (In particular, $\#K = q^q$ for $k = q$, and K is empty, $\#K = 0$, for any $k > q$.) The invariance of K is obvious. Next, some $x \in K$ are brought by two elements $g_1 = \langle {\sigma_1}^{i_1}, \ldots, {\sigma_k}^{i_k} \rangle$, $g_2 = \langle {\sigma_1}^{j_1}, \ldots, {\sigma_k}^{j_k} \rangle$ of G to the same point if and only if ${\sigma_l}^{i_l}(x \cap M_l) = {\sigma_l}^{j_l}(x \cap M_l)$, $\forall l = 1, \ldots, k$, that is, when $g_1 = g_2$. (Why?) Deduce from the preceding statement that the orbits of G on K are q^k-element and, therefore, that $\#K$ is divisible by q^k.

E11.12

(Feller 1966) A power series $F(t)$ converges for $| t | \leq 1$. (Why?) In turn, $U(t)$ converges for $| t | < 1$; indeed, from th recursion relation we have, by induction, the boundedness of u,

$$(0 \leq)\ u_k \leq C \text{ for } k \Rightarrow (0 \leq)\ u_n \leq f_1 \cdot u_{n-1} + \ldots + f_n \cdot u_0 \leq C,$$

and therefore,

$$\sum |u_n t^n| \leq C \cdot \sum |t^n| = C \cdot (1 - |t|)^{-1} < \infty \qquad (|t| < 1)$$

(what does this mean?). The recursion relation on u with regard to $u_0 = 1$ imposes the following relationship on the generating functions:

$$U(t) - 1 = U(t) \cdot F(t).$$

$U(t)$ grows with t when $t > 0$ (why?); hence, for any N

$$\sum_{n=0}^{N} u_n \leq \lim_{t \to 1-0} U(t) \leq \sum_{n=0}^{\infty} u_n = s_u,$$

and thus, $\lim\limits_{t\to 1-0} U(t) = s_u$. But $U(t) \to (1 - s_f)^{-1}$ if $s_f < 1$ and $U(t) \to \infty$ if $s_f = 1$,[17] which completes the proof.

E11.13

Step 1. If u_{n_k-j} did not tend to the same limit (say, upper limit), one could find arbitrarily large n so that

$$u_n > \lambda - \varepsilon, \qquad u_{n-j} < \lambda' < \lambda.$$

Choose N so large that $\sum\limits_{n=N+1}^{\infty} f_n < \varepsilon$. Since $u_k \le 1$, we have

$$u_n \quad \le \quad f_0 u_n + \ldots + f_N u_{n-N} + \varepsilon.$$

For large n any u_k on the right-hand side is less than $\lambda + \varepsilon$; therefore,

$$\begin{aligned} u_n \quad &< \quad (f_0 + \ldots + f_{j-1} + f_{j+1} + \ldots + f_N) \cdot (\lambda + \varepsilon) + f_j \lambda' + \varepsilon \\ &\le (1 - f_j) \cdot (\lambda + \varepsilon) + f_j \lambda' + \varepsilon \quad < \quad \lambda + 2\varepsilon - f_j (\lambda - \lambda'), \end{aligned}$$

which, provided that ε is so small that $f_j(\lambda - \lambda') > 3\varepsilon$, contradicts the inequality $u_n > \lambda - \varepsilon$. The case of the lower limit is processed similarly.

Step 2. The identities $r_0 u_n + r_1 u_{n-1} + \ldots + r_n u_0 = 1$ are obtained by induction from the recursion relations on sequence u. A detailed proof of Lemma 1 is at the end of this section.

Step 3. According to step 1, if $f_j > 0$, then

$$u_{n_k} \to \lambda \;\Rightarrow\; u_{n_k-j} \to \lambda \;\Rightarrow\; u_{n_k-2j} \to \lambda \;\Rightarrow\; u_{n_k-3j} \to \lambda.$$

When $j = 1$, $u_{n_k-i} \to \lambda$ for any i. Applying the identities obtained in step 2 to $n = n_k$, we have $\sum\limits_{i=0}^{N} r_i u_{n_k-i} \le 1$. Consider the upper limit $\lambda = \underline{\lambda} := \overline{\lim}\, u_n$. For a fixed N any $u_{n_k-i} \to \bar{\lambda}$, hence, $\bar{\lambda} \cdot \sum\limits_{i=0}^{N} r_i \le 1$ and, because of the arbitrariness of N, $\bar{\lambda}\mu \le 1$. Specifically for $\mu = \infty$, u_n will have zero upper limit, so $u_n \to 0$ (as $u_n \ge 0$).

[17] As $U(t) = (1 - F(t))^{-1} = \sum\limits_{n=0}^{\infty} [F(t)]^n$ for $|t| < 1$ and $[F(t)]^n > ½$ for t close to $1 - 0$, if $s_f = 1$, then $U(t) > 1 + n/2$ for t close to $1 - 0$ when $s_f = 1$; what does this imply?

Suppose now that $\mu < \infty$, and consider the lower limit $\lambda = \underline{\lambda} := \underline{\lim}\, u_n$. If N is large enough so that $\sum_{i=N+1}^{\infty} r_i < \varepsilon$, then $\sum_{i=0}^{N} r_i\, u_{n_k - i} + \varepsilon \geq 1$ (as $0 \leq u_n \leq 1$). Since $u_{n_k - i} \to \underline{\lambda}$ for any i, $\underline{\lambda} \cdot \sum_{i=0}^{N} r_i + \varepsilon \geq 1$, therefore, $\underline{\lambda}\, \mu \geq 1$.

If $f_1 = 0$, then using the nonperiodicity condition we can find a finite set of indices $j_1, \ldots$ with the greatest common divisor 1 such that $f_{j_l} > 0$, $\forall l$. We have

$$u_{n_k} \to \lambda \;\Rightarrow\; u_{n_k - m_1 j_1 - m_2 j_2 - \ldots} \to \lambda, \quad \forall \textit{fixed}\ m_1 \geq 0,\ m_2 \geq 0,\ \ldots.$$

According to the cited lemma from number theory, any natural number exceeding $\prod j_l$ can be represented by a linear combination of j_l with positive integer coefficients. Check that we will arrive at the same inequalities $\bar{\lambda}\mu \leq 1$ and $\underline{\lambda}\, \mu \geq 1$ by applying the identities obtained in step 2 to $n_k{}' = n_k - \prod j_l$ in place of n_k. This completes step 3.

Proof of lemma 1 with the generating functions $A(t)$, $R(t)$ (Feller 1966). Convergence of a number series $\sum a_n$ means convergence of the power series $A(t)$ for $t = 1$. Abel's famous lemma implies uniform, over $| t | \leq \theta$ [$\forall \theta \in (0,1)$], convergence of $A(t)$ itself and a series of $A(t)$'s derivatives, which provides the differentiability $A'(t) = \sum n a_n t^{n-1}$ for $| t | < 1$.[18] (Why?) As a consequence of nonnegativity, $A'(t) \to \sum n a_n$ and $R(t) \to \sum r_n$ when $t \to 1 - 0$.[19] The equations $a_n = r_{n-1} - r_n$, $n = 1,2,\ldots$ mean that $A(t) - a_0 = t \cdot R(t) - R(t) + r_0$, that is,

$$R(t) = \frac{a_0 + r_0 - A(t)}{1 - t}.$$

By the Lagrange intermediate value theorem, $\forall t \in (0,1)$, $\exists \theta \in (t,1)$: $R(t) = A'(\theta)$, and moving on to the limit as $t \to 1 - 0$ will yield the required equality.

Proof of Lemma 1′. Proceed using a method similar to that in Polya et al. (1964). Let d_n be the tails of $\sum n a_n$: $d_n = (n+1)\, a_{n+1} + (n+2)\, a_{n+2} + \ldots$. We will have $a_n = \frac{d_{n-1}}{n} - \frac{d_n}{n+1}$, hence,

$$a_1 + a_2 + a_3 + \ldots = d_0 - d_1 + \tfrac{1}{2}(d_1 - d_2) + \tfrac{1}{3}(d_2 - d_3) + \ldots\ .$$

The series on the right-hand side converges, as can be seen from the following lemma.

[18] By induction, we will have infinite differentiability.

[19] For a nonnegative-element series $\sum c_n$, if $c_n > 0$, then $\sum_{k=0}^{n} c_k\, t^k > \sum_{k=0}^{n-1} c_k$ for t close to $1 - 0$. Hence, $\lim_{t \to 1-0} \sum c_n t^n = \sum c_n$, including the case where $\sum c_n = \infty$. (Why?) A similar limit relation holds for a convergent series with arbitrary elements $\sum c_n$ (which is proved on the basis of Lemma 2′ (see below on this page); prove it!).

Lemma 2′. (Fichtengoltz 1970). *If a sequence* α_n *is monotone, then*

$$\left| \sum_{i=1}^{m} \alpha_i \beta_i \right| \leq \max_{1 \leq n \leq m} |B_n| \cdot (\,|\alpha_1| + 2|\alpha_m|\,),$$

where $B_n := \sum_{i=1}^{n} \beta_i$.

Prove Lemma 2′ using **Abel's transform** (a discrete analog of an integration-by-parts formula). Thus, the series $\sum a_n$ converges. For the tails we find

$$\begin{aligned} b_0 &= d_0 - d_1 + \tfrac{1}{2}(d_1 - d_2) + \frac{1}{3}(d_2 - d_3) + \ldots, \\ b_1 &= \frac{1}{2}(d_1 - d_2) + \frac{1}{3}(d_2 - d_3) + \ldots, \\ b_2 &= \frac{1}{3}(d_2 - d_3) + \ldots, \end{aligned}$$

and so on. Therefore, $\sum_{n=0}^{m} b_n = d_0 - d_{m+1} + (m+1) \cdot \left(\frac{d_{m+1} - d_{m+2}}{m+2} + \frac{d_{m+2} - d_{m+3}}{m+3} + \ldots \right)$. With regard to $d_n \to 0$, as $n \to \infty$, Lemma 2′ shows that $\sum b_n$ converges to d_0. **QED**.

E11.14

(Feller 1966). The desired relationship between generating functions is

$$U(t) = \frac{B(t)}{1 - A(t)},$$

generalizing the relationship in section E11.12. Convergence for $s_a < 1$ and divergence for $s_a = 1$ of $\sum u_n$ is proved by arguments similar to those used in section E11.12.

Now let the series $\sum a_n$ be nonperiodic. For $s_a = 1$, let v_n be coefficients at t^n in the expansion

$$\begin{aligned} [1 - A(t)]^{-1} &= \sum_{n=0}^{\infty} [A(t)]^n \\ &= \frac{a_0}{1 - a_0} + \frac{a_1 t}{(1 - a_0)^2} + \left[\frac{a_1}{(1 - a_0)^3} + \frac{a_2}{(1 - a_0)^2} \right] t^2 + \ldots. \end{aligned}$$

In accordance with section P11.13, $v_n \to \mu^{-1}$. Next,

$$u_n = v_n b_0 + v_{n-1} b_1 + \ldots + v_0 b_n.$$

For any fixed k, $v_{n-k}b_k \to \mu^{-1}b_k$ as $n \to \infty$. In addition, v_n are bounded (because v_n has a finite limit). From the foregoing discussion, u_n differs arbitrarily little from $v_n b_0 + \ldots + v_{n-N}\, b_N$ for large N; in turn, the last value tends to $(b_0 + \ldots + b_N)\cdot\mu^{-1}$, which differs arbitrarily little from $s_b\mu^{-1}$. **QED**.

For $s_a \geq 1$, when the sequence a is bounded, the power series $A(t)$ converges for $|\,t\,| < 1$ (at the rate of a geometric progression in t) and monotonically increases with the growth of t. This, with $a_0 < 1$, provides the unique solvability of the equation $A(t) = 1$ for $t > 0$, with the root $t_0 < 1$. The recursion relation from section P11.14** is satisfied with the sequences $\left\{a_n t_0^n\right\}_{n=0,1,\ldots}$, $\left\{b_n t_0^n\right\}_{n=0,1,\ldots}$, $\left\{u_n t_0^n\right\}_{n=0,1,\ldots}$ in place of a, b, u, respectively. Application of the claim to $s_a = 1$ completes the proof.

Completing the Solution

S11.2

A generating function is $\frac{(t+1)^{s+1}-(u+1)^{s+1}}{t-u}$. Therefore, extending the symbol $f(a,b)$ by zero to the domain $\{a < 0\}\cup\{b < 0\}$, we have

$$\sum_{a,b\geq 0,\ a+b\leq s+1} [f(a-1,b) - f(a,b-1)]\, t^a u^b = (t+1)^{s+1} - (u+1)^{s+1}.$$

This proves the equalities $f(a-1,b) = f(a,b-1)$ since the right-hand side does not contain terms with $t^a u^b$ for $a > 0$ and $b > 0$. **QED**.

Also, it provides the complete solution of section P11.2 by comparing coefficients of the monomials on both sides, yielding $(f(d,0) = f(d-1,1) = \ldots =) f(0,d) = \binom{s+1}{d+1}$. (We leave it to the reader to work out the details.)

S11.5

Inserting decomposition $\binom{|\alpha|+k+n+m}{|\beta|+n+m} = \binom{|\alpha|+k+n+(m-1)}{|\beta|+n+(m-1)} + \binom{|\alpha|+(k-1)+n+m}{|\beta|+n+m}$ into the left-hand side of the identitiesresults, by the inductive hypothesis, in $\binom{k+m-1}{m-1} + \binom{k-1+m}{m} = \binom{k+m}{m}$. **QED**. (We leave it to the reader to fill in the details.)

S11.6

The stabilization of coefficients, meaning that $p_{n,k}(m) = p_{n+1,k}(m) = p_{n+2,k}(m) = \ldots$, arises for $k = 0$ starting from $n = 1$ and, for $k > 0$, when $n \geq m + 1 - (k - 1)(k - 2)/2$. (Why?)

S11.10

Proof of divisibility of order (cardinality) of a finite group of transformations by cardinality of any of its orbits. Let a group H act on a set S. Consider an orbit H_s. Let H_s be a **stationary subgroup** of the element s, $H_s = \{h \in H: hs = s\}$. (Verify that it is a subgroup.) The left **residue classes** modulo H_s, hH_s ($h \in H$) coincide or do not intersect (why?), and, evidently, each of them has the same number of elements as H_s. Hence, the order of H equals the product of the order of H_s and the number of the left (or right) residue classes. (This assertion is usually referred to as **Lagrange's theorem**.) Thus, since the points of the orbit H_s correspond one-to-one to the left residue classes (why?),[20] the cardinality of the orbit divides the order of H. **QED**.

Let $\#L > 1$ and l be the least of L. A q-element subset A intersects with M_l by $m < q$ elements, so that applying the transformations $\sigma, \sigma^2, \ldots, \sigma^q = e$ will change the remainder modulo q summed over the elements of A by $m, 2m, \ldots, qm \bmod q$, respectively; all these numbers are distinct since $rm \not\equiv 0 \bmod q$ for $r,m < q$. (Why?) Therefore, any $R_{i,L}$ is once, and only once, mapped onto any other $R_{j,L}$. (We leave it to the reader to work out the details.) **QED**.

S11.11

The famous principle of inclusion–exclusion from elementary combinatorics gives for a finite set A and its subsets A_i ($i \in I$) the cardinality of the complement to the union of A_i as

$$\#(A \setminus \bigcup A_i) = \#A - \sum_i \#A_i + \sum_i \#(A_i \cap A_j) - \ldots + (-1)^{\#I} \#\bigcap A_i.$$

[20] We will have a one-to-one correspondence of the orbit to the set of right residue classes if we denote the group's action by h: $s \mapsto sh$ ($h \in H$).

(Readers who have not yet come across this formula may establish it on their own using quite obvious arguments, or they may find a proof in handbooks on combinatorial analysis.) Apply this formula to A, which is the total class of q-element subsets of $\{0,\ldots, kq-1\}$, and A_i, which are the subclasses consisting of all of those subsets that do not intersect with M_i, respectively, which yields a description of K just as $K = A \setminus \cup A_i$. (Work out the details.)

The same approach as in section S11.10 shows that applying distinct transformations from the group G_l to $S \in K$ will change the summarized remainders modulo q of the elements of $S \cap M_l$ by distinct values. (Fill in the details using the primality of q.[21]) Hence, only identity transformation from G leaves the set S invariant; therefore, the orbits are regular, as claimed. **QED**. (We leave it to the reader to fill in the details.)

S11.12

For a convergent nonnegative-element series $\sum f_n$, a power series $\sum f_n t^n$ converges (even absolutely) for $|t| \leq 1$ since it **is majorized** by the first series. (Work out the details.)

However, readers may give simple examples showing that a **convergent radius** (the supremum of numbers r such that $\sum f_n t^n$ converges for $|t| \leq r$) greater than 1 is not guaranteed by the convergence of $\sum f_n$.

S11.13

Step 2. Proof of identities $r_0 u_n \pm \ldots \pm r_n u_0 = 1, \forall n = 0,1,\ldots$. For $n = 0$, it holds that $f_0 = 0$, $\sum f_n = 1$ and $u_0 = 1$. If these identities hold $\forall n < k$, then we find

$$\begin{aligned} r_0 u_k + \ldots + r_k u_0 &= r_0(f_1 u_{k-1} + \ldots + f_k u_0) + r_1(f_1 u_{k-2} + \ldots + f_{k-1} u_0) + \ldots \\ &+ r_{k-1} f_1 u_0 + r_k u_0 = f_1(r_0 u_{k-1} + \ldots + r_{k-1} u_0) + \ldots \\ &+ f_k r_0 u_0 + r_k u_0 = f_1 + \ldots + f_k + r_k u_0 = \sum_{n=0}^{\infty} f_n = 1, \end{aligned}$$

QED.

[21] This does not hold for composite numbers, which may be verified immediately by considering the first one, 4. (Do this!) Accordingly, the claim of section P11.11* does not hold for composite q. (Consider an example of $q = 4$ and $k = 2$.)

On the proof of Lemma 1 from section H11.13 with the generating functions $A(t), B(t)$. Readers may establish that the *series* $\sum c_n$ *converges* $\Rightarrow \sum c_n t^n \to \sum c_n$ *as* $t \to 1 - 0$[22] using the following arguments. Given $\varepsilon > 0$, let N be large so that $\left|\sum_{n=N+1}^{M} c_n\right| < \varepsilon$ for $M > N$. Using Lemma 2′ from section E11.13 yields, for $0 \le t < 1$, an estimate

$$\left|\sum_{n=N+1}^{M} c_n t^n\right| \le \max_{N+1 \le n \le M} \left|\sum_{i=N+1}^{n} c_i\right| \cdot (t^{N+1} + 2t^M) < 3\varepsilon$$

(which also proves the convergence of $\sum c_n t^n$ without employing Abel's lemma; furnish the details). Thus, we have $\left|\sum_{n=1}^{\infty} c_n - \sum_{n=1}^{N} c_n\right| \le \varepsilon$ and $\left|\sum_{n=1}^{\infty} c_n t^n - \sum_{n=1}^{N} c_n t^n\right| \le 3\varepsilon$. (Why?) Using this, and applying once again the previously cited Lemma 2′, yields

$$\left|\sum_{n=1}^{\infty} c_n - \sum_{n=1}^{\infty} c_n t^n\right| \le \varepsilon + 3\varepsilon + \left|\sum_{n=1}^{N} c_n (1 - t^n)\right|$$

$$\le 4\varepsilon + 2N \cdot (1 - t) \cdot \max_{1 \le n \le N} \left|\sum_{i=1}^{n} c_i\right|,$$

which is less than 5ε for t close to $1 - 0$. (Work out the details; note that an estimate $\left|\sum_{n=1}^{N} c_n (1 - t^n)\right| < \varepsilon$ for t close to 1 may also be obtained using such an obvious argument as the continuity of a polynomial $\sum_{n=1}^{N} c_n t^n$) **QED**.

Proof of Lemma 2′ from section E11.13. Abel's transform moves a sum of products to the form

$$\sum_{i=1}^{m} \alpha_i \beta_i = \alpha_1 B_1 + \alpha_2 (B_2 - B_1) + \ldots + \alpha_m (B_m - B_{m-1})$$

$$= \alpha_m B_m - \sum_{i=1}^{m-1} (\alpha_{i+1} - \alpha_i) B_i.$$

[22] The series $\sum c_n t^n$ will converge (even absolutely) for $|t| < 1$ by Abel's lemma.

For a monotone sequence of α_i, the differences on the right-hand side will be of the same sign, so we have, with the notation $M := \max_{1 \le i \le m} |B_i|$, the estimate

$$\left|\sum_{i=1}^{m} \alpha_i \beta_i\right| = \left|\alpha_m B_m - \sum_{i=1}^{m-1} (\alpha_{i+1} - \alpha_i) B_i\right| \le M\left(|\alpha_m| + \left|\sum_{i=1}^{m-1} (\alpha_{i+1} - \alpha_i)\right|\right) \le M(|\alpha_1| + 2|\alpha_m|),$$

as claimed. **QED**.

Step 3. To prove the relations $\bar{\lambda}\mu \le 1$, and, for $\mu < \infty$, $^{-}\lambda\mu \ge 1$, in the case of $f_1 = 0$, readers may proceed by the same method as for $f_1 > 0$ (section E11.13). That is, considering $n_k' = n_k - \prod j_l$ and taking into account the lemma from number theory in section H11.13 yields the equalities $n_k' - i = n_k - (n_k - n_k') - i = n_k - \sum m_{il} j_l$ with fixed, for fixed $i \ge 0$, nonnegative (for $i > 0$, even positive) integers m_{il}, showing that $u_{n'_k - i}$ of any fixed $i \ge 0$ tends to the same limit, as does u_{n_k}; andthus, the rest of the proof may be accomplished in the same way as for $f_1 > 0$ in section E11.13. (Complete the proof.)

S11.15

In the asymptotic formula for $u_{n\lambda+j}$, $\mu^{(0)}$ arises, which plays the same role for the sequence $a^{(0)}$ as μ does for a: $\mu^{(0)} = \sum n a_n^{(0)}$. But since $\mu^{(0)} = \sum n a_{n\lambda} = \mu/\lambda$, we have $\lim_{n\to\infty} u_{n\lambda+j} = s_{b^{(j)}} \left(\mu^{(0)}\right)^{-1} = \lambda\, s_{b^{(j)}} \mu^{-1}$.

Least Squares and Chebyshev Systems

Problems

P12.0[*]

Preliminaries. As readers know, polynomials of degree n, in other words linear combinations of $n + 1$ monomials $1,\ldots, t^n$, may have at most n real zeros. A far-reaching generalization of this fact raises a fundamental concept of **nonoscillatory by Sturm**, or **Chebyshev systems**, **T-systems** for short (T is the first letter of German spelling of the discoverer's surname, which is "Tchebyscheff"). Those systems are defined as follows. For a set (or system) of functions $F = \{f_0,\ldots\}$, a linear combination of a finite number of elements, $f = \sum c_i f_i$, is called a **polynomial on *F*** (considered nonzero when $\exists i$: $c_i \neq 0$). A system of $n + 1$ functions $F = \{f_0,\ldots,f_n\}$ on an interval (or a half-interval, or a non-one-point segment) $I \subseteq \mathbb{R}$ is referred to as having the **Chebyshev property**, or T-system, when nonzero polynomials on F may have at most n distinct zeros in I (zeros of extensions outside I are not counted). Geometrically, the Chebyshev property means that the curve $t \mapsto [f_0(t),\ldots, f_n(t)]$ in $\mathbb{R}^{n+1}$ may have at most n distinct common points with any hyperplane passing through the origin. (*Why?*)

The elements of a T-system are always linearly independent. (*Prove.*)

Many works are devoted to the history, comprehensive theory, and numerous applications of T-systems; interested readers may consult Karlin and Studden (1966), Krein and Nudelman (1973), Gantmacher and Krein (1950)and multiple references therein. However, the basic ideas of the theory of T-systems may be understood by readers with relatively limited experience. In this problem group we focus both on these ideas and on some nice applications of this theory such as estimation of numbers of zeros and critical points of functions (in analysis and geometry: sections P12.15[***] and P12.25[***]), real-life problems in tomography (section P12.17[**]), interpolation theory (sections P12.18[*] and P12.19[*]), approximation theory (sections P12.26[***], P12.27[***], P12.28[***], P12.29[***], P12.30[***], and

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8_12, © Springer Science+Business Media, LLC 2013

P12.31***), and, of course, least squares (sections P12.32**, P12.33**, P12.34**, P12.35**, P12.36**, P12.37**, P12.38**, P12.39***, P12.40***, and P12.41***).

We undertake a comprehensive discussion of the linear least-squares technique. We present its analytic, algebraic, and geometric versions (section P12.32**), and we examine its connections with linear algebra (sections P12.33** and P12.34**) and combinatorial analysis (in this connection we discuss some problems previously considered in "A Combinatorial Algorithm in Multiexponential Analysis" and "Convexity and Related Classic Inequalities" problem groups: section P12.35**, from a different point of view). We also discuss some features of the least-squares solutions (e.g., passing through determinate points, asymptotic properties; sections P12.36**, P12.37**, and P12.38**). In addition, we outline ties of the least squares with probability theory and statistics (section P12.40***). Finally, we discuss real-life applications to polynomial interpolation (such as finding the best polynomial fitting for two-dimensional surfaces: section P12.39***),[1] and signal processing in nuclear magnetic resonance technology (such as finding covariance matrices for maximal likelihood estimating parameters: section P12.41***).

*P12.1**

Show that $F = \{f_0,\ldots,f_n\}$ is T-system on I if and only if $\det (f_i(t_j))_{i,j\,=\,0,\ldots,\,n} \neq 0$ for any distinct points $t_0,\ldots, t_n \in I$. (Therefore, for continuous functions f_i, these determinants have the same signs while an order of the points is kept unchanged (say, $t_0 <\ldots< t_n$); *prove*.) To *what* is this determinant equal for $F = \{1,\ldots, t^n\}$?

From the preceding discussion *derive* that for a nondegenerate matrix $A = (a_{ij})_{i,j\,=\,0,\ldots,\,n}$, $AF = \left\{\sum_j a_{ij} f_j\right\}_{i=0,\ldots,n}$ will be a T-system if F is. Therefore, the Chebyshev property remains after replacing f_i with $k_i f_i$, with constant $k_i \neq 0$. Also, this property remains after multiplying all f_i by another such function that does not have zero valrcues.

[1] The problem of polynomial interpolation (or polynomial fitting) appears in numerous applications, both theoretical and practical, including high-order numerical methods, signal and image processing, and a variety of other applications. As a very small sample, see Wiener and Yomdin (2000) and Yomdin and Zahavi (2006) for an analysis of the stability of polynomial fitting and applications in high-order numerical solutions of linear and nonlinear partial differential equations (PDEs), Elichai and Yomdin (1993), Briskin et al. (2000), and Haviv and Yomdin (pers. communication: Model based representation of surfaces) for high-order polynomial fitting in image processing, and Yomdin (submitted) and Brudnyi and Yomdin (pers. communication: Remez Sets) for a geometric analysis of "fitting sets," as appear in Problem P12.39***. (We thank Yosef Yomdin for kindly giving us the details.)

P12.2[*]

The property of being a T-system is connected to I. *Provide* examples of losing this property if I is extended by one point.

P12.3[*]

Provide examples showing that subsystems of T-systems may not be T-systems themselves. Actually, T-systems for which some or all of their subsystems are T-systems deserve special consideration. A system of functions $F = \{f_0,\ldots,f_n\}$ on $I \subseteq \mathbb{R}$ is referred to as having the **Markov property**, or being a **Markov system** (**M-system**), when subsystems $\{f_0,\ldots,f_m\}$ are T-systems on I for $0 \le m \le n$ [which is equivalent, according to section P12.1[*], to $\det (f_i(t_j))_{i,j\,=\,0,\ldots,\,n} \ne 0$ for $0 \le m \le n$ and distinct points $t_0,\ldots,t_m \in I$]. The same system of functions is referred to as having the **Descartes property**, or being a **Descartes system** (**D-system**), when all its subsystems are T-systems on I [which is equivalent, according to section P12.1[*], to $\det \left(f_i(t_j)\right)_{\substack{i=i_0,\ldots,i_m \\ j\,=\,0,\ldots,m}} \ne 0$ for $0 \le m \le n$, distinct $t_0,\ldots,t_m \in I$, and sequences $0 \le i_0 <\ldots< i_m \le n$], with the signs of these determinants for fixed $t_0,\ldots, t_m$ being the same for all sequences $0 \le i_0 <\ldots< i_m \le n$. (For continuous f_i, these signs are the same also for any $t_0 <\ldots< t_m$; *prove*.) Obviously, the subsystems of D-systems themselves are D-systems.

Provide examples showing that the inclusion {M-systems} $\subseteq$ {T-systems} is strict.

Provide examples of losing the Markov property with an appropriate permutation of f_i. (Therefore, the inclusion {D-systems} $\subseteq$ {M-systems} is strict.)

Obviously, the Markov (Descartes) property remains after replacement of f_i by $k_i f_i$, with constant $k_i \ne 0$ (resp. the constant k_i of the same sign), and after multiplying each f_i by a function (the same for all i) that does not have zero values. (*Why*?) *Verify* a more general statement: for a lower-triangular matrix $A = (a_{ij})_{i,j\,=\,0,\ldots,\,n}$ with $a_{ii} \ne 0$, $AF = \left\{\sum_j a_{ij} f_j\right\}_{i=0,\ldots,n}$ will be an M-system if F is.

P12.4[*]

Show that monomials $1,\ldots, t^n$ form a D-system on $(0,\infty)$, and the same holds for polynomials $p_i = \sum_{j=k_{i-1}+1}^{k_i} a_j t^j$ $(\sum k_i = n)$ having all coefficients in their totality either positive or negative. (They will simply form an M-system if these coefficients obey

a weaker condition $sign(a_{k_{i-1}+1}) = \ldots = sign(a_{k_i}),\ \forall i$, but the Markov property will remain following any permutations.)

P12.5[*]

Establish a generalization of the statement in section P12.4[*]: polynomials on a D-system $F = \{f_0,\ldots,f_n\}$, $p_i = \sum_{j=k_{i-1}+1}^{k_i} a_j f_j$ ($\sum k_i = n$), will form a D-system if all their coefficients together are either positive or negative and an M-system when those coefficients obey a weaker condition $sign(a_{k_{i-1}+1}) = \ldots = sign(a_{k_i}),\ \forall i$.

P12.6[*]

Prove that a system F of functions on $I \subseteq \mathbb{R}$ is a D-system if and only if it obeys the **Descartes rule of signs**: the number N of distinct zeros of a polynomial on F in I does not exceed the number W of sign changes in the sequence of coefficients of this polynomial, $N \leq W$ (the changes are counted between adjacent coefficients and zero coefficients are not counted in the sequence) (Gantmacher and Krein 1950).

P12.7[*]

Let I and J be intervals (or half-intervals, or non-one-point segments) of $\mathbb{R}$. A continuous function K: $I \times J \to \mathbb{R}$ is referred to as a **kernel of regular sign** when $\det(K(s_i,t_j))_{i,j\,=\,0,\ldots,\,n} \neq 0$ for $n \geq 0$ and distinct $s_0,\ldots,s_n \in I$ and $t_0,\ldots,t_n \in J$. *Show* that $K(s_i,.)$ form a D-system for any n and distinct $s_0,\ldots,s_n \in I$.

P12.8[*]

A T-system of n times differentiable functions on I, $F = \{f_0,\ldots,f_n\}$, is referred to as an **ET-system** when nonzero polynomials on F have at most n zeros, considering multiplicity. Consider, for $t_0,\ldots,t_n \in I$, a determinant $W = W_{n+1}(f_i)(t_j)_{i,j\,=\,0,\ldots,\,n}$ with the columns $(f_i(t_j))_{i\,=\,0,\ldots,\,n}$ for the distinct t_j and $(f_i'(t_j))_{i\,=\,0,\ldots,\,n}$,…, $(f_i^{[r-1]}(t_j))_{i\,=\,0,\ldots,\,n}$ for the second,…, rth repetitions of t_j, respectively. Show that $F = \{f_0,\ldots,f_n\}$ is an ET-system on I if and only if $W \neq 0$ for any (perhaps equal) $t_0,\ldots,t_n \in I$. [Therefore, for functions f_i having a continuous nth derivative,

these determinants have the same signs while an order of the points is kept unchanged (say, $t_0 \leq \ldots \leq t_n$); *prove*.[2]]

From the preceding discussion *derive* that for a nondegenerate matrix $A = (a_j)_{i,j\,=\,0,\ldots,\,n}$, $AF = \left\{\sum_j a_{ij} f_j\right\}_{i=0,\ldots,n}$ will be an ET-system if F is one as well. Therefore, this property will remain after replacing f_i by $k_i f_i$, with constant $k_i \neq 0$. Also, this property remains after multiplying all of f_i by a function (the same for all i) that does not have zero values.

P12.9[*]

Formulate definitions of EM-systems and ED-systems in terms of numbers of zeros and their equivalents in terms of signs of determinants. [Therefore, the property of being an EM- (ED-) system remains after replacing f_i by $k_i f_i$, with constant $k_i \neq 0$ (resp. constant k_i of the same sign), or multiplying them by a n times differentiable function (the same for all i) that does not have zero values.]

The preceding examples of T- (M-, D-) systems were actually ET- (EM-, ED-) systems (e.g., $1,\ldots, t^n$ form an ED-system on $I = (0,\infty)$: section P12.10[*].) *Show* that $1,\ldots, t^{n-1}, f(t)$ form an ET- (and so, EM-) system on $I \subseteq \mathbb{R}$ if nth derivative of f keeps sign in I. (Further examples are discussed below.)

Establish, for ED-systems, analogs of the statements in sections P12.5[*] and P12.6[*] [in the second, it is stated that a set of n times differentiable functions $F = \{f_0,\ldots,f_n\}$ is an ED-system on I if and only if the number of zeros of a polynomial on F in I, considering multiplicity, does not exceed the number of changes in signs in the sequence of coefficients of this polynomial (the changes are counted between adjacent coefficients and zero coefficients are not counted in the sequence)].

P12.10[*]

Prove that the functions $e^{s_0 t}, \ldots, e^{s_n t}$ $(s_0 < \ldots < s_n)$ form an ED-system on $I = \mathbb{R}$. Next, *derive* that $t^{s_0}, \ldots, t^{s_n}$ form an ED-system on $I = (0,\infty)$ for any $s_0 < \ldots < s_n$.

[2] For $n > 0$, this sign is unchanged regardless of the continuity of the nth derivatives. Readers may prove this with the help of the lemma from section E12.11 below, taking into account the continuity of differentiable functions.

P12.11[*]

The **Wronskian** of n times differentiable functions $f_0,\ldots, f_n$ is a determinant $W_{n+1}(f_0,\ldots,f_n)(t) = \det\begin{pmatrix} f_0(t) & \ldots & f_n(t) \\ \vdots & & \vdots \\ f_0^{[n]}(t) & \ldots & f_n^{[n]}(t) \end{pmatrix}$. *Find* it for the exponential functions in section P12.10[*].

Establish a generalization of the statements in section P12.10[*] claiming that:

1. *n times differentiable functions* $f_0,\ldots, f_n$ *form an EM-system if and only if* $\forall t\in I$, $W_{m+1}(f_0,\ldots, f_m)\,(t) \neq 0$ *for* $0 \leq m \leq n$.[3]
2. *n times differentiable functions* $f_0,\ldots, f_n$ *form an ED-system if and only if* $\forall t\in I$, $W_{m+1}(f_{i_0},\ldots,f_{i_m})(t) \neq 0$ *for* $0 \leq m \leq n$ *and* $0 \leq i_0 <\ldots< i_m \leq n$, *and* W_{m+1} *at a fixed* $t\in I$ *has the same sign for all these sequences. In this case, this sign is kept throughout* I *when* $n > 0$, *or when* $n = 0$ *and* f_0 *is continuous.*[4]
3. *Let a function* $K(s,t)$ *be continuous in* (s,t) *and n times differentiable in t. K is a "kernel of regular sign (defined in section P12.7*[*]*), considering multiplicity," that is, functions* $f_i(t) = K(s_i,t)$ *form an ED-system for distinct* $s_0,\ldots, s_n$, *if and only if* $\forall t\in I, W_{m+1}(f_{i_0},\ldots,f_{i_m})(t) \neq 0$ *for* $0 \leq m \leq n$ *and* $0 \leq i_0 <\ldots< i_m \leq n$. *In this case,* W_{m+1} *has the same sign for all these sequences and* $t\in I$.

A number of classic examples of ED-systems, some variations, and generalizations of the Rolle theorem and other related topics are discussed in Polya and Szegö (1964).

P12.12[*]

Show that $g_0 = -\frac{df_0}{df_m},\ldots, g_{m-1} = -\frac{df_{m-1}}{dt\ f_m}$, $g_m = -\frac{df_{m+1}}{dt\ f_m},\ldots, g_{n-1} = -\frac{df_n}{df_m}$ (for a fixed $m\in\{0,\ldots, n\}$) form an ED-system on I if $f_0,\ldots, f_n$ do (Polya and Szegö 1964).

---

[3] This is a multivariate version of the Rolle theorem. ***Warning***: a single inequality $W_{n+1}(f_0,\ldots,f_n) \neq 0$, considered separately from the whole family of these inequalities for $0 \leq m \leq n$, does not guarantee the Chebyshev property of $f_0,\ldots, f_n$; *provide* counterexamples.

[4] For $n = 0$, the continuity of f_0 needs to be assumed also because differentiability of the zero order commonly means nothing. Of course, it might be defined as continuity since representability of $f(x + h)$ by a polynomial of degree n in h with accuracy $o(|h|^n)$ is equivalent to continuity at x for $n = 0$ and the nth-order differentiability at x for $n > 0$. However, the definition of continuity as the zero-order differentiability is not commonly accepted.

P12.13[*]

Show that, for a D-system consisting of continuous functions on I, the ratios $f_j(t)/f_i(t)$ are positive in I, and for an ED-system, those ratios with $i < j$ all together strictly increase or strictly decrease (Polya and Szegö 1964).

P12.14[*]

Prove, for polynomials on an ED-system $\{f_0,\ldots\}$, that $W \equiv N \bmod 2$ (where W is as defined in section P12.6[*] and N is the number of roots in I, considering multiplicity) if the ratios $f_j(t)/f_i(t)$ with $i < j$ have zero limits at one end of I and infinite limits at the other end. (This condition is satisfied for the systems from section P12.10[*].)

P12.15[***]

This problem is addressed to readers familiar with linear differential equations. A linear combination of exponential functions with polynomial coefficients $f = \sum p_i(t)\, e^{\lambda_i t}$ is referred to as a **quasipolynomial**. *Prove* that a quasipolynomial has at most $\sum(\deg p_i + 1) - 1$ real zeros, considering multiplicity. *Does* a similar claim hold for quasipolynomials of a more general form $f = \sum (p_i(t)\cos\omega_i t + q_i(t)\sin\omega_i t)\, e^{\lambda_i t}$?

The estimation of the number of zeros is very important in multiple applications. A number of classic examples of such estimations are discussed in Polya and Szegö (1964). Far-reaching developments of this subject are presented in Khovanskii (1997).

P12.16[*]

The claims from section P12.10[*] may be reformulated to state that the functions e^{st} and t^s are "kernels of regular sign, considering multiplicity" [in the same sense as in section P12.11[*]; for these functions, respectively, $I = J = (-\infty,\infty)$ and $I = J = (0,\infty)$]. *Prove* the same claim for the functions $(s + t)^{-1}$ (the Cauchy kernel) using $I = J = (0,\infty)$ or any other intervals (or half-intervals, or non-one-point segments) I, J such that $0 \notin I + J = \{s + t\colon s \in I, t \in J\}$ and $e^{-(s+t)^2}$ (the Gaussian kernel), using $I = J = (-\infty,\infty)$.

P12.17[**]

Prove that $h_i(\theta) = (\cos\theta)^i(\sin\theta)^{n-i}$ $(i = 0,\ldots, n)$ form an ET-system on $[0,\pi)$.

This paragraph shows a real-life example from computerized tomography and is addressed to readers familiar with ***Radon transform***. The Chebyshev property of the foregoing trigonometric polynomials makes it possible to determine the **moments of** a planar density **distribution** (which are proportional to the Taylor coefficients of a Fourier image) via moments of its Radon image, which finds practical applications in tomography. Specifically, for a function $K(s)$ such that $K(x_1\cos\theta + x_2\sin\theta)$ is summable with the density $f(x)$ $(x = (x_1, x_2) \in \mathbb{R}^2)$, there is the equality

$$\int_{\mathbf{R}^2} K(x_1\cos\theta + x_2\sin\theta) f(x_1, x_2)\, dx_1 dx_2 = \int_{-\infty}^{\infty} K(s) p(\theta, s)\, ds,$$

where $p(\theta,s)$ is the Radon image of $f(x)$ (Gelfand et al. 1959–1962; Helgason 1980, 1984). (A proof is immediately obtained using the Fubini theorem.) Inserting here a function $K(s) = s^n$ yields the formula known in tomography as the **moment-projection theorem** (Smith et al. 1977; Herman 1980, and references therein):

$$\int_{\mathbf{R}^2} (x_1\cos\theta + x_2\sin\theta)^n f(x_1, x_2)\, dx_1 dx_2 = \int_{-\infty}^{\infty} s^n p(\theta, s)\, ds.$$

Taking the integral on the right-hand side of this equality for distinct values $\theta_0,\ldots, \theta_n$ yields a system of $n + 1$ linear equations on the desired moments $\int_{\mathbf{R}^2} x_1^i x_2^{n-i} f(x_1, x_2)\, dx_1 dx_2$ $(i = 0,\ldots, n)$ with the matrix of coefficients $\left(\binom{n}{i} h_i(\theta_j)\right)_{i,j=0,\ldots,n}$, which is uniquely resolvable because of the Chebyshev property of $\{h_i\}$.[5]

P12.18[*]

Prove the existence, for a system $F = \{f_0,\ldots,f_n\}$ of functions on I and given distinct points $t_1,\ldots, t_m \in I$ $(m \le n)$, of a nonzero polynomial on F having those points among its zeros. *Prove* that for a T-system these polynomials form an $n - m + 1$-dimensional vector subspace. Specifically, for $m = n$ this polynomial is defined uniquely, up to proportionality; define it explicitly via f_i and t_j. *Make* equivalent geometric formulations regarding a curve $t \mapsto (f_0(t),\ldots, f_n(t))$ in $\mathbb{R}^{n+1}$.

[5] Similarly, setting $K(s) = e^{-irs}$ yields an expression of the two-dimensional Fourier image of $f(x)$ via the one-dimensional Fourier image of $p(\theta,s)$ with respect to variable s, known in tomography as the **slice-projection theorem**. All these formulas have obvious generalizations for multivariable distributions.

P12.19[*]

Show that, given a system $F = \{f_0,\ldots,f_n\}$ of functions on I, distinct points $t_0,\ldots,$ $t_n \in I$, and values $y_0,\ldots,y_n$, one cannot guarantee the existence (or uniqueness, if it exists) of a polynomial on F taking values y_j at t_j, respectively; and actually, the unique existence of this polynomial, for any $t_0,\ldots,t_n$ and $y_0,\ldots,y_n$, is equivalent to the Chebyshev property. *Interpret* this statement in terms of an $n+1$-dimensional vector space of functions on an $n+1$-point set and its vector subspace corresponding to the polynomials on F. More generally, *make* a similar description for a set of the polynomials on a T-system F on I taking given values $y_0,\ldots,y_{m-1}$ at distinct points $t_0,\ldots, t_{m-1} \in I$, respectively ($m \le \#F$).

Come up with a formulation regarding the points $\big(f_0(t_j),\ldots,f_n(t_j),y_j\big) \in \mathbf{R}^{n+2}_{x_0,\ldots,x_n,y}$ ($j = 0,\ldots, n$) equivalent to the unique existence of a polynomial taking the values $y_0,\ldots,y_n$, at $t_0,\ldots,t_n$. Polynomials on a T-system F, taking given values at $n+1$ ($=\#F$) given distinct points, are referred to as **Lagrange interpolating polynomials** (on F). Their coefficients, $c_0,\ldots,c_n$, are defined by a system of linear equations $\sum c_i f_i(t_j) = y_j$, $j = 0,\ldots, n$. For large n, a direct solution becomes impractical; fortunately, in some important cases there are different ways to find the solution. For $p(t) = \sum\limits_{i=0}^{n} c_i t^i$ *obtain* the representation $p(t) = \sum\limits_{j=0}^{n} p(t_j) \cdot \prod\limits_{i\ne j} \frac{t-t_i}{t_j-t_i}$, decomposing $p(t)$ with respect to a δ-basis in the space of functions on $n+1$-point set $\{t_0,\ldots, t_n\}$.[6]

Notice that a leading coefficient c_n may be defined using the Jacobi identities (section P1.1[**] above):

$$c_n = c_n \cdot \sum_{j=0}^{n} t_j^n \cdot \prod_{k\ne j} (t_j - t_k)^{-1} = \sum_{j=0}^{n} \left(\sum_{i=0}^{n} c_i t_j^i \right) \cdot \prod_{k\ne j} (t_j - t_k)^{-1} = \sum_{j=0}^{n} p(t_j) \cdot \prod_{i\ne j} (t_j - t_i)^{-1}.$$

Warning. The Lagrange polynomial using $\{t_0,\ldots, t_n\}$ may have its degree smaller than n and may even become equal to a zero polynomial!

P12.20[***]

A real-valued continuous function f is referred to as changing signs passing through a point if $f \ge 0$ but does not vanish identically in one of that point's half-neighborhoods and if $f \le 0$ but does not vanish identically in the other. Following

[6] δ-basis $p_0,\ldots, p_n$ is defined as $p_i(t_j) = \delta_{ij} = \begin{cases} 1 & \text{for } \ i = j \\ 0 & \text{otherwise} \end{cases}$. These functions also provide restrictions to the set $\{t_0,\ldots, t_n\}$ of certain Lagrange polynomials; we leave it to the reader to find out which polynomials are restricted in this way.

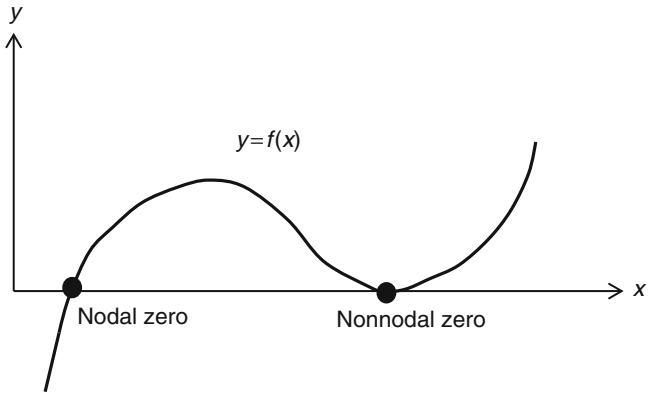


Fig. 1 Nodal and nonnodal zeros

Krein and Nudelman (1973), we will call zeros that do (do not) change signs **nodal** (resp. **nonnodal**) (Fig. 1).

Interpret, for a polynomial on $f_0,\ldots, f_n$, nodal and nonnodal zeros geometrically, in terms of a curve $t \mapsto (f_0(t),\ldots, f_n(t))$ in $\mathbb{R}^{n+1}$. *Prove* the following theorem generalizing a statement from section P12.18* to T-systems consisting of continuous functions. In this theorem, in the cases where the domain of f is a segment, zeros at the endpoints are considered nodal:

Theorem *For a T-system $F = \{f_0,\ldots, f_n\}$ on I consisting of continuous functions, given distinct points $t_1,\ldots, t_{k+l} \in I$ with $t_1,\ldots, t_k \in \operatorname{int} I$, there exists a polynomial on F having only zeros $t_1,\ldots, t_{k+l}$, specifically, $t_1,\ldots,t_k$ nonnodal, and $t_{k+1},\ldots,t_{k+l}$ nodal, only if $2k + l \leq n$. If $n \equiv l \bmod 2$ or I is a segment, then this condition is also sufficient.*

[(Therefore, for even n or $I = [a,b]$, there is a positive polynomial ($k = l = 0$).]

*P12.21*****

For a continuous function $I \xrightarrow{f} \mathbf{R}$ a number of sign changes is defined as a supremum of natural numbers n for which there exist $n + 1$ distinct points such that the signs of f on these points alternate. Let f change sign $n < \infty$ times and $t_0 <\ldots< t_n$ be such points that sign $f(t_{i+1}) = -\text{sign } f(t_i) \neq 0$, $i = 0,\ldots, n$. Show the unique existence, for each i, of two (perhaps coinciding) points $a_i \leq b_i$ of the interval (t_i,t_{i+1}) with the following properties:

1. $f = 0$ on $[a_i,b_i]$.
2. For certainty, if $f(t_i) > 0$ and $f(t_{i+1}) < 0$, then $f \geq 0$ on $[t_i,a_i]$ and $f \leq 0$ on $[b_i,t_{i+1}]$, and there are points in any left half-neighborhood of a_i where $f > 0$ and points in any right half-neighborhood of b_i where $f < 0$.

Show that the set of nodal zeros of f (defined in section P12.20***) coincides with $\bigcup[a_i,b_i]$ (a "strict," meaning from plus to minus or vice versa, sign change corresponds to passing through the whole segment $[a_i,b_i]$).

Prove that a continuous function $[a,\ b] \xrightarrow{f} \mathbf{R}$, not vanishing identically, changes sign in (a,b) at least n times if $\int_a^b f(t)\,dt = 0, \ldots, \int_a^b f(t)\,t^{n-1}dt = 0$ (Hurwitz's theorem).

Prove the following generalization: a continuous function $[a,\ b] \xrightarrow{f} \mathbf{R}$, not vanishing identically, changes sign in (a,b) at least n times if it is orthogonal to the elements of a T-system of continuous functions on $[a,b]$, $F = \{f_0,\ldots,f_{n-1}\}$ with respect to a scalar product $\langle f,\ g\rangle = \int_a^b f(t)g(t)\,dt$. (The same holds using more general scalar products $\langle f,\ g\rangle = \int_a^b f(t)g(t)w(t)\,dt$, with a positive (except perhaps on a set having zero volume) weight functions w.)

*P12.22**

The functions $f_0,\ldots$ on the segment $I = [a,b]$ taking equal values at the ends and forming a T- (ET-) system on $[a,b)$ (equivalently, on $(a,b]$) are referred to as forming a **periodic** T- (ET-) system on I, or, using different terminology, a T- (ET-) system on the circle $\mathbb{R}$ mod $(b-a)$. *Give* examples of periodic T-systems on $[a,b]$ that are not T-systems on $[a,b]$. *Show* that a periodic T-system consisting of continuous functions has an odd number of elements.[7]

*P12.23**

Show that the trigonometric harmonics $1,\ \cos\theta,\ldots,\ \cos n\theta,\ \sin\theta,\ldots,\ \sin n\theta$ form a periodic ET-system on the unit circle $\mathbb{S}^1 = \mathbb{R}$ mod 2π.

*P12.24****

The definition for the number of sign changes of a real-valued continuous function in section P12.21*** is not applicable to the functions on a circle. (*Why*?) Let us

[7] Therefore, a reasonable way of defining periodic M- (EM-) systems $\{f_0,\ldots,f_{2n}\}$ is as follows: subsystems $\{f_0,\ldots,f_{2n}\}$ must be periodic T- (ET-) systems on I for $0 \le m \le n$.

modify to make the definition applicable. For a finite subset P in the domain of function f let $\overline{\overline{P}}$ denote a set of the pairs of geometrically adjacent elements of P such that $f > 0$ on one of them and $f < 0$ on another. Call the number of sign changes the supremum, over all P, of cardinalities $\#\overline{\overline{P}}$. *Verify* that for functions on intervals (or half-intervals, or segments) this new definition is equivalent to the definition from section P12.21***. *Show* that the number of sign changes on a circle is even if it is finite. *Establish* that if the number of sign changes on a circle equals n, then the set of nodal zeros of f coincides with a union of the segments $[a_i,b_i] \subset (t_i,t_{i+1})$ ($i = (0,\ldots, n-1) \bmod n$) (defined in section P12.20***).

Prove that a real-valued continuous function on a circle not vanishing identically changes sign at least $2n$ times if it is orthogonal to the elements of a T-system of continuous functions on this circle, $F = \{f_0,\ldots, f_{2n-2}\}$, with respect to a scalar product $\langle f, g\rangle = \oint f(\theta)g(\theta)\,d\theta$. [The same holds using a scalar product of a more general kind $\langle f, g\rangle = \oint f(\theta)g(\theta)w(\theta)\,d\theta$, with a positive (except perhaps on a set having zero volume) weight function w.]

P12.25***

A famous theorem states that a 2π-periodic (in other words, defined on $\mathbb{S}^1$) continuous (or at least square-integrable on $\mathbb{S}^1$) function f is expanded in a Fourier series, which is convergent, in the sense of term-by-term integrability with any 2π-periodic continuous (or at least square-integrable on $\mathbb{S}^1$) multipliers g:

$$f(\theta) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos n\theta + b_n \sin n\theta :$$

$$\int_0^{2\pi} f(\theta)g(\theta)\,d\theta = \frac{a_0}{2}\int_0^{2\pi} g(\theta)\,d\theta + \sum_{n=1}^{\infty} a_n \int_0^{2\pi} \cos n\theta \cdot g(\theta)\,d\theta + b_n \int_0^{2\pi} \sin n\theta \cdot g(\theta)\,d\theta.$$

Prove that f has at least as many distinct zeros on the circle as its first nonvanishing harmonics:

$$f(\theta) \sim \sum_{n=N}^{\infty} a_n \cos n\theta + b_n \sin n\theta \quad \Rightarrow \quad \#\{\theta \bmod (2\pi) : f(\theta) = 0\} \geq 2N,$$

and, moreover, that f changes signs at least $2N$ times (C. Sturm, A. Hurwitz).

This paragraph is addressed to readers familiar with linear differential equations. Trigonometric harmonics are eigenfunctions of a periodic boundary value problem $-\ddot{f} = \lambda f, f(0) = f(2\pi)$ ($\lambda_n = n^2$, $n = 0,1,\ldots$) (for ordinary differential equations, we use dots to denote derivatives, as is

usual in physics). Many widely applicable periodic and nonperiodic EM-systems are formed by eigenfunctions of corresponding **Sturm-Liouville boundary eigenvalue problems**. In such problems, the linear differential operator of the second order is symmetric with respect to an integral (perhaps weighted) scalar product in the space of functions satisfying the respective boundary conditions, so that the eigenfunctions corresponding to distinct eigenvalues are orthogonal. The **Sturm theorem** states that zeros of the eigenfunctions separate zeros of any other eigenfunctions that correspond to smaller eigenvalues, so the number of zeros of the eigenfunction increases with the next eigenvalue by one in the nonperiodic case and by two in the periodic case. (In the nonperiodic case, the eigenvalues are simple. In the periodic case, the eigenvalue of the minimal magnitude is simple, and the rest are double.) Interested readers will find discussions of this subject, its further development, and related topics in Gantmacher and Krein (1950), Karlin and Studden (1966), Courant and Hilbert (1953–1962)), and references therein.

Application to the four-vertex problem. *This paragraph is addressed to readers familiar with the fundamentals of differential geometry of planar curves.* A continuous, periodic, of period L, function $\kappa(s)$ is a curvature of a connected naturally parameterized planar curve of length L

$$x(s) = x_0 + \int_0^s \cos\theta(t)\, dt, \qquad y(s) = y_0 + \int_0^s \sin\theta(t)\, dt,$$

where θ is a polar angle of a tangent vector (x',y') (the prime denotes d/ds), $\theta(s) = \theta_0 + \int_0^s \kappa(t)\, dt$. [Taking into account that $\rho := |(x',\, y')| = 1$, the curve is defined by the **Frenet equations** $\rho' = 0$, $\theta' = \kappa$, or, using Cartesian coordinates, $x'' = -\kappa y'$, $y'' = \kappa x'$.] Functions $x(s)$, $y(s)$ are L-periodic, and $\theta(s)$ increases on $[0,L]$ by $2\pi n$ (a multiple of 2π). Assume that $\kappa > 0$ and $n = 1$. Such a curve is referred to as a smooth planar **oval**; it is free of self-intersections and bounds a convex set. (*Prove.*) *Verify* that a continuously differentiable substitution of independent variable s by θ yields the preceding parametric equations to the form (we may always assume that $\theta_0 = 0$)

$$x(\theta) = x_0 + \int_0^\theta R(\eta)\cos\eta\, d\eta, \qquad y(\theta) = y_0 + \int_0^\theta R(\eta)\sin\eta\, d\eta$$

[$R(\theta) = \kappa^{-1}$ is the radius of curvature]. Take an oscillatory component of R, $R_0 = R - (2\pi)^{-1}\int_0^{2\pi} R\, d\theta$. *Verify* that $\int_0^{2\pi} R_0(\theta)\, d\theta = \int_0^{2\pi} R_0(\theta)\cos\theta\, d\theta = \int_0^{2\pi} R_0(\theta)\sin\theta\, d\theta = 0$. Hence, by the preceding theorem, R_0 changes sign on $\mathbb{S}^1$ at least four times. Therefore, R_0 has at least four extrema (among them, two maxima and two minima) (*why*?), and the same is true for $R = R_0 + const$ and $\kappa = R^{-1}$. Thus, we have proved

The four-vertex theorem: The curvature of a smooth planar oval has at least four extrema.

[We recommend that readers analyze circles, ellipses (Fig. 2), and further special cases.]

The name of this theorem derives from the fundamental geometrical fact that critical points of a curve's curvature correspond to cusps ("vertices") of its evolute. This classic theorem, proved in the early twentieth century by S. Mukhopadhyaya and A. Kneser, lies at a junction of several powerful mathematical theories. A proof similar to the preceding one appeared in a work of S.L. Tabachnikov. Interested readers will become familiar with the impressive far-reaching development of the subject and find the names of the researchers that contributed to it in Arnol'd (1994, 2002) and references therein.

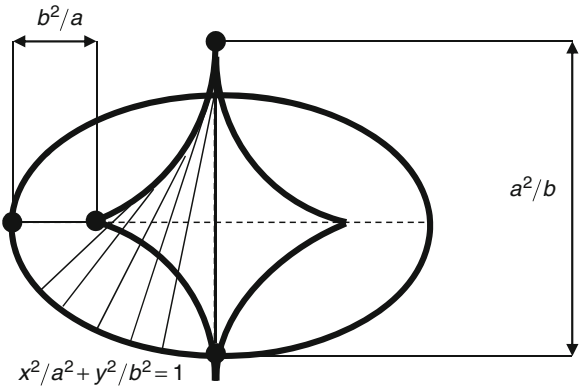


Fig. 2 Evolute of an ellipse

P12.26[***]

Definitions. In section P12.19[*] we encountered an interpolation problem for finite-dimensional spaces of functions on finite-point sets. (For the same spaces, a problem of a more general kind, the famous least-squares problem, is discussed below.) In sections P12.26[***], P12.27[***], P12.28[***], P12.29[***], P12.30[***], and P12.31[***] we consider an approximation problem for an (infinite-dimensional) space of continuous functions on a segment. Given a system $F = \{f_0,\ldots\}$ of continuous functions on $I = [a,b]$, a polynomial p on F with the least value of $\max\limits_{t\in I} |f(t) - p(t)|$ over all those polynomials is referred to as a **polynomial of best approximation for the (continuous) function *f* on *I***. In geometrical terms, the preceding definition means the following. Consider a subspace of the space of continuous functions on I spanned on F and f. Let H be a hyperplane spanned on F. (H consists of the polynomials on F.) The polynomial of best approximation is an element of H closest to f if the distance between functions is defined as $\max\limits_{t\in I} |f(t) - g(t)|$. Hence, a polynomial of best approximation exists on I for any f and finite F (*prove*) but may not exist with infinite F (*give* an example). *Verify* that p is a polynomial of best approximation if and only if $f - p$ is an element of an affine hyperplane $f + H$ **of the least deviation from zero** (the closest to the origin) **on *I***.

P12.27[***]

Prove the central claim regarding the polynomials of best approximation:

Theorem (P.L. Chebyshev) *A polynomial p on a T-system is a polynomial of best approximation for f on* $[a,b]$ *if and only if* $f(t) - p(t)$ *takes, in turn, values* $+d$ *and* $-d$ *at some* $\#F + 1$ *points* $a \le t_0 < \ldots < t_{\#F} \le b$, *where* $d = \max\limits_{a\le t\le b} |f(t) - p(t)|$.

P12.28[***]

Prove the following complement to the previous claim:

Theorem (A. Haar) *Given a system F of continuous functions on I = [a,b], a polynomial of best approximation is unique for any continuous function f on I if and only if F is a T-system.*

P12.29[***]

The famous **Chebyshev polynomials** $T_n(t)$ ($n = 0,1,2,\ldots$) are defined as $T_0 = 1$ and, for $n > 0$, as polynomials $\sum_{i=0}^{n} a_{n-i}t^i$, of the least deviation from zero on $[-1,1]$ over the polynomials of degree n in t with leading coefficients $a_0 = 2^{n-1}$. In other words, $T_n(t) = 2^{n-1}t^n - p_{n-1}(t)$, where p_{n-1} is a polynomial of degree $n - 1$ in t of best approximation for $2^{n-1}t^n$ on $[-1,1]$. The Haar theorem gives the unique existence of $T_n(t)$, $\forall n = 1,2,\ldots$. *Derive* from it the evenness of $T_n(t)$ for even n and the oddness of $T_n(t)$ for odd n. (Therefore, the extrema are located symmetrically, with respect to zero.)

By the Chebyshev theorem, $T_n(t)$ has at least $n + 1$ extrema in $[-1,1]$. Actually, there are exactly $n + 1$ extrema, and the first and last ones coincide with the endpoints. (*Why*?)

P12.30[***]

Verify that cos (n arccos t) are polynomials in t and satisfy the condition of the Chebyshev theorem. From this *deduce* that $T_n(t) = \cos$ (n arccos t). Derive the explicit representations for the Chebyshev polynomials (Fig. 3):

$$T_1 = t, \quad T_2 = 2t^2 - 1, \quad T_3 = 4t^3 - 3t, \quad T_4 = 8t^4 - 8t^2 + 1,$$
$$T_5 = 16t^5 - 20t^3 + 5t, \ldots.$$

Interested readers will find more about the properties of Chebyshev polynomials and their role in approximation theory in multiple publications; the following classic guides are recommended: Polya and Szegö(1964) and Berezin and Zhidkov (1959) and the monographs by Szegö (1981) and Erdelyi (1953). Also, Chebyshev polynomials play a very important role in mathematical physics, including such topics as vibrations of circles, Poisson integrals for disks, and others. In connection with this function, the Chebyshev polynomials represent in dimension two a hierarchy having representatives in other dimensions as well. In this hierarchy, the Chebyshev polynomials are called **cyclotomic**. The representatives in dimension three are the famous **Legendre**, or **spherical**, polynomials. The members of the hierarchy in aggregate are called **Gegenbauer**, **hyperspherical**,

or **ultraspherical** polynomials. Gegenbauer polynomials corresponding to a fixed dimension are orthogonal on [−1,1], with respect to properly weighted integral scalar products; therefore, since any $n + 1$ polynomials of degrees 0,. . ., n ($n = 0,1,\ldots$) form T-systems, Gegenbauer polynomials of degree n have n simple zeros in (−1,1) (by the claim from section P12.21***). Gegenbauer polynomials determine the fundamental vibrations of hyperspheres. These vibrations are zonal, that is, spherical zones between fixed latitudes (determined by zeros of the corresponding Gegenbauer polynomial) vibrate, but the points on those latitudes themselves remain immovable. Unfortunately, a detailed introduction to this beautiful branch of mathematical physics incorporating methods of differential equation theory, algebraic geometry, and group representation theory is beyond the scope of the present problem book. The interested reader may consult Arnol'd (1997), Vilenkin (1965), Szegö (1981), and Erdelyi (1953).

*P12.31****

Following Polya and Szegö (1964), define $\mu_n(a,b) := \max\limits_{a \le t \le b} |p(t)|$ for the polynomial $p(t)$ of degree n in t with leading coefficient 1, which is of the least deviation from zero on a segment [a,b] over all such polynomials. The explicit formula from section P12.30*** gives $\mu_n(-1,1) = 2^{1-n}$ ($n = 1,2,\ldots$). (*Prove.*) *Evaluate* $\mu_n(a,b)$ for any segment [a,b]. In particular, *find* that $\lim\limits_{n\to\infty} \mu_n(a,b) = 0 \Leftrightarrow b - a < 4$.

*P12.32***

Definitions. The famous **least-squares technique** comprises various recipes for evaluating/adjusting/tuning constant parameters in prescribed formulas (e.g., expressing physical laws) via minimizing the sum of squared deviations between the predicted and factual outputs in a series of experiments. In this problem group, we will analyze formulas that depend linearly on parameters to be adjusted.[8] The corresponding minimization problem, which is referred to as the (**linear**) **least-squares problem**, allows for different equivalent formulations. Here, we introduce the three most important formulations; let us agree to call them **algebraic**, **analytic**, and **geometric**, respectively:

Algebraic formulation: Given a linear operator $C \xrightarrow{A} Y$ *between Euclidean spaces and a point* $y \in Y$, *find* $x \in C$ *with a minimal value of* $|Ax - y|$. *(Obviously, the*

[8] For nonlinear models, see Press et al. (1992) and references therein, paying special attention to the brilliant **Levenberg-Marquardt method**.

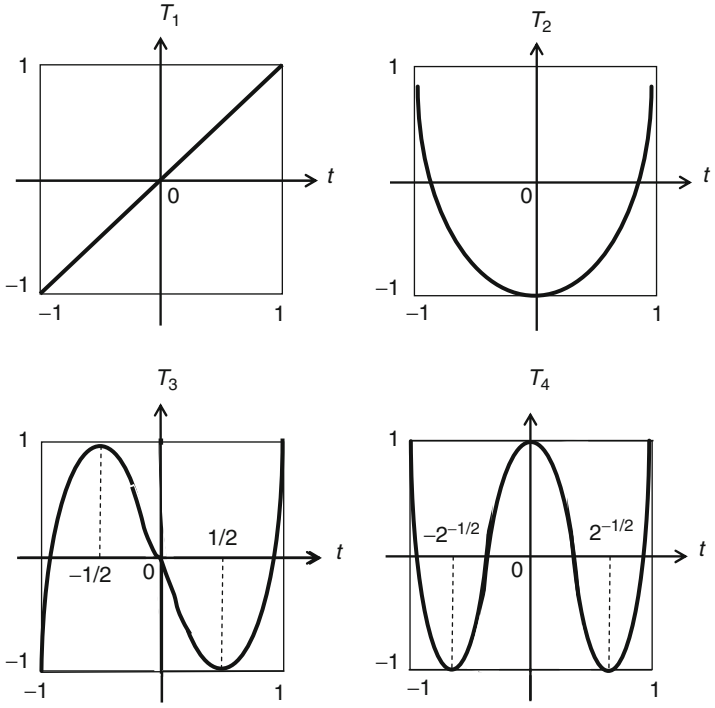


Fig. 3 The graphs of the first four Chebyshev polynomials

problem is equivalent to $|Ax - y|^2 = \min$, *with which the name "least squares" is bound.)* ***Legend****: x is a vector of the parameters, A is the theoretical relationship, Ax is the theoretical output, and y is the experimental result.*

Analytic formulation: Given a set of functions $F = \{f_0(t),\ldots, f_n(t)\}$ *and assumed values* $y_1,\ldots, y_m$ *experimentally found at* $t_1,\ldots, t_m$, *respectively, define a polynomial on F* $p(t) = \sum x_i f_i$ *so that* $\sum(p(t_j) - y_j)^2 = \min$. f_i *may be defined on any set of t not supplied with any topological, algebraic, or other structure, but most often* $t_1,\ldots, t_m$ *are the points of a grid inscribed into a real interval (or half-interval or segment) or a planar body.* ***Legend for*** t_j*: they are labels of the measurement acts (for example, they may denote moments of time).*

Geometric formulation. Given points $(u_{0j},\ldots, u_{nj}, y_j) \in \mathbf{R}^{n+2}_{u_0,\ldots u_n,y}$: $j = 1,\ldots, m$, *draw a nonvertical hyperplane (i.e., described by an equation of a form* $y = \sum x_i u_i$*) with a minimal, in a vertical (y-) direction, total squared deviation from the data*

$$\sum_j \left(\sum_i x_i u_{ij} - y_j\right)^2.$$

The equivalence of the algebraic and analytic versions is immediately obvious from the interpretation of C, Y, and A (similar to that in section H12.19): C as the space of the polynomial coefficients, Y as the space of the functions on the points t_1,. . ., and A as the linear operator $\begin{pmatrix} c_0 \\ c_1 \\ \vdots \end{pmatrix} \mapsto \begin{pmatrix} f_0(t_1) & f_1(t_1) & \cdots \\ f_0(t_2) & & \\ \vdots & & \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \end{pmatrix}$. The equivalence of the analytic and geometric versions is established by identifying $f_i(t_j)$ with u_{ij}. (*Do* this!)

From the algebraic formulation find that the least-squares problem is always solvable and, in general, nonuniquely. Meanwhile, a practically important case is the case of uniqueness. The uniqueness condition is $\ker A = 0$. In particular, $\dim Y \geq \dim C$ is necessary: the number of distinct measurements should not be less than the number of unknowns. In general, this condition is not sufficient. It is convenient to formulate the necessary and sufficient conditions of the uniqueness for the analytic and geometric versions of the problem. Readers will briefly *verify* the correctness of the following formulations:

Let F be a set of functions on a real interval (or half-interval, or segment). The least-squares problem with respect to an unknown polynomial on F is solvable uniquely, given the results of any $m \geq n + 1 = \#F$ distinct measurements, if and only if F is a T-system.

(In particular, the solution for $m = n + 1$ is the Lagrange interpolating polynomial considered in section P12.19*.)

The least-squares problem with respect to an unknown nonvertical hyperplane in $\mathbb{R}^{n+2}_{u_0,\ldots u_n,y}$ is solvable uniquely, given m points, if and only if their projections to a horizontal coordinate subspace $\mathbb{R}^{n+1}_{u_0,\ldots u_n}$ do not belong to a hyperplane (a vector subspace of codimension one) of this subspace.

In a practically important case, one of F's elements, say, f_0, is constant, $f_0 \equiv 1$. *Verify* that the related geometric version of the least-squares problem allows for a reformulation as follows: *given points* $(u_{1j}, \ldots, u_{nj}, y_j) \in \mathbb{R}^{n+1}_{u_1,\ldots u_n,y}$: $j = 1,\ldots, m$, *draw an affine hyperplane of the form* $y = x_0 + \sum_{i=1}^{n} x_i u_i$ *with a minimal* (*in the vertical* (*y*-) *direction*) *total squared deviation from the data:* $\sum_j \left(x_0 + \sum_{i=1}^{n} x_i u_{ij} - y_j \right)^2 = \min$. In this case, the necessary and sufficient conditions of the uniqueness look as follows:

The least-squares problem with respect to an unknown nonvertical affine hyperplane in $\mathbb{R}^{n+1}_{u_1,\ldots u_n,y}$ is solvable uniquely, given m points, if and only if their projections to a horizontal coordinate subspace $\mathbb{R}^{n}_{u_1,\ldots,u_n}$ do not belong to an $n-1$-dimensional affine plane of this subspace.

P12.33[**]

Prove that the least-squares problem is equivalent, with the preceding (in section P12.32[**]) notations, to an equation ${}^tA(Ax - y) = 0$, where a linear operator $Y \xrightarrow{{}^tA} C$ is adjoint to A. (Readers who have not worked with adjoint operators will find the necessary information in section E6.3 above.) *Calculate* ${}^tA \circ A$ for a T-system $\{t^n, \ldots, t, 1\}$.

P12.34[**]

Show that $\ker A = 0 \Leftrightarrow {}^tA \circ A$ *is invertible*. Therefore, the solution of a uniquely solvable least-squares problem is found as $x = ({}^tA \circ A)^{-1}({}^tAy)$.

P12.35[**]

Show that ${}^tA \circ A$ is positive definite for a nondegenerate matrix A. Prove, using this and statements in sections P12.32[**], P12.33[**], and P12.34[**], that $\det \begin{pmatrix} \sum t_j^{2n} & \cdots & \sum t_j^n \\ \vdots & & \vdots \\ \sum t_j^n & \cdots & m \end{pmatrix} = 0$ for $m = \#\{j\} \le n$ but positive for $m > n$ and distinct t_j, and, more generally, $\det \begin{pmatrix} \sum t_j^{2n} & \cdots & \sum t_j^n s_j^n \\ \vdots & & \vdots \\ \sum t_j^n s_j^n & \cdots & \sum s_j^{2n} \end{pmatrix} = 0$ for $m \le n$, but positive for $m > n$ if there exists an $n + 1$-element sequence $j_0 < \ldots < j_n$ of indices j such that either all s_j with these indices are nonzero and the ratios t_j/s_j are distinct or all t_j with these indices are nonzero and the ratios s_j/t_j are distinct (in fact, $\begin{pmatrix} \sum t_j^{2n} & \cdots & \sum t_j^n s_j^n \\ \vdots & & \vdots \\ \sum t_j^n s_j^n & \cdots & \sum s_j^{2n} \end{pmatrix} = {}^tV \circ V$ with $V = \begin{pmatrix} t_1^n & t_1^{n-1}s_1 & \cdots & s_1^n \\ \vdots & \vdots & & \vdots \\ t_m^n & t_m^{n-1}s_m & \cdots & s_m^n \end{pmatrix}$). (Thus, the reader will obtain a new proof, different from those discussed previously in the "A Combinatorial Algorithm in Multiexponential Analysis" and "Convexity and Related Classic Inequalities" problem groups.)

P12.36[**]

Perform a rectilinear data fitting in actuality, or, in other words, solve the least-squares problem with respect to an unknown straight line $y = x_0 + x_1u$ in $\mathbb{R}^2_{u,y}$, given data (u_j,y_j): $j = 1,\ldots, m$ (with at least two distinct u_j). *Establish* the following property of this line:

Given $(u_j,y_j)\in\mathbb{R}^2$: $j = 1,\ldots, m$ *(with at least two distinct* u_j*), the straight line* $y = x_0 + x_1u$*, which is the solution of the least-squares problem, passes through a point* $(\sum u_j/m, \sum y_j/m)$ *and, if* $\sum u_j \neq 0$*, through a point* $(\sum u_j^2/\sum u_j, \sum u_jy_j/\sum u_j)$.[9]

(For $m = 2$ it is trivial; *why*?)

P12.37[**]

Establish a generalization of the claim in section P12.36[**]:

Given $(u_{1j},\ldots, u_{nj}, y_j) \in \mathbf{R}^{n+1}_{u_1,\ldots u_n,y}$: $j = 1,\ldots, m$ *(with projections to a horizontal coordinate hyperplane* $\mathbf{R}^n_{u_1,\ldots u_n}$ *belonging to no* $n - 1$*-dimensional affine planes), an affine hyperplane* $y = x_0 + \sum_{i=1}^{n} x_iu_i$*, which is the solution of the least-squares problem, passes through* $n + 1$ *points* $(\sum\alpha_{kj}u_{1j},\ldots, \sum\alpha_{kj}u_{nj}, \sum\alpha_{kj}y_j)$*, where* $\alpha_{0j} = m^{-1}$ *and* $\alpha_{kj} = u_{kj}\Big/\sum_l u_{kl}$ *for* $k = 1,\ldots, n$ *(provided that the denominators do not vanish).*[10]

(For $m = n + 1$ it is trivial; *why*?)

P12.38[**]

Consider the uniform grid $N_0 \times\ldots\times N_n$ with edges parallel to the coordinate axes on a parallelepiped $\Pi_a = \{-a_i \le u_i \le a_i,\ \forall i\} \subset \mathbf{R}^{n+1}_{u_0,\ldots,u_n}$ $(a_i \ge 0, \forall i; a = (a_0,\ldots, a_n))$; denote this grid $G^{n+1}(N,a)$ $(N = (N_0,\ldots, N_n); N_i > 1$ if $a_i > 0)$. Now let us consider the set of all hyperplanes $y = \sum_{i=0}^{n} x_iu_i$ in $\mathbf{R}^{n+2}_{u_0,\ldots u_n,y}$ that are solutions of the least-squares problem for data sets $\{(u_{0j},\ldots, u_{nj},y_j)$: $j = 1,\ldots, \prod N_i\}$ such that $(u_{0j},\ldots, u_{nj})$ are the nodes of $G^{n+1}(N,a)$. *Prove* that parameters x_i of those hyperplanes decrease like a_i^{-1}

[9] Experienced readers familiar with projective geometry may eliminate the restriction of $\sum u_j \neq 0$, as the point with $\sum u_j = 0$ belongs (at infinity) to the projective closure of this line in $\mathbb{R}P^2$.

[10] Experienced readers familiar with projective geometry may eliminate this restriction considering points at infinity of the projective closure as well.

irrespective of the values $N_0,\ldots,N_n$, as $a_i \to \infty$ and all y_j remain bounded; more precisely:

There exists $K > 0$ *such that* $|x_i| \le KCa_i^{-1}$ $(i = 1,\ldots, n)$ *if* $N_0,\ldots,N_n > 1$ *and* $|y_i| \le C$, $\forall j$.

Regarding the grid $G^n(N,a) \subset \mathbf{R}^n_{u_1,\ldots,u_n}$ $[N = (N_1,\ldots, N_n)$, $a = (a_1,\ldots, a_n)]$, a similar statement holds true for the least-squares problem with respect to an affine hyperplane $y = x_0 + \sum_{i=1}^{n} x_i u_i$ in $\mathbf{R}^{n+1}_{u_1,\ldots u_n,y}$ (here a_0 should be set equal to one): the solution becomes a horizontal hyperplane as $a_1,\ldots, a_n \to \infty$.

P12.39***

Find the minimal $m(d)$ of natural m with the following property: there exist m points $(s_j,t_j) \in \mathbb{R}^2$, $j = 1,\ldots, m$, such that the least-squares problem with respect to a polynomial $p(s,t) = \sum_{k+l \le d} x_{kl} s^k t^l$ is uniquely solvable, with any given values y_j that refer to (s_j,t_j), respectively. *Describe* an algebraic algorithm for a determination if the solvability is unique, given $m \ge m(d)$ points. For $d = 2$, also *find* a geometric algorithm.

P12.40***

This and the next problem are addressed to readers familiar with elements of mathematical statistics. Dealing with empirical data, the following classical problem arises: Let "true" values of a random function Y be determined by unknown parameters $c = (c_0,\ldots, c_n)$: $\bar{y}_j = A_j(c)$ (with known functions A_j). How is it possible to assess c with a sufficient store of values of Y, that is, what integer m and map $(y_1,\ldots,y_m) \overset{\Phi}{\mapsto} (x_0,\ldots,x_n)$ provide a good estimate $x = (x_0,\ldots, x_n)$ for c, where $y = (y_1,\ldots, y_m)$ denotes a vector of observed values of the function Y? Legendre suggested obtaining the estimates $x = \Phi(y)$ with the least-squares technique

$$\sum \big(A_j(x) - y_j\big)^2 = \min .$$

Gauss, who investigated this problem in depth, provided several probabilistic foundations for Legendre's method, among them the following:

Assume that Y is normally distributed, its first,..., mth values are sampled independently from one another, and the samplings are free of systematic errors. This means that they form independent Gaussian random values $Y_1,\ldots,Y_m$ *of the same distribution function (which is assumed to be known), and the "true" values*

$\bar{y}_1, \ldots, \bar{y}_m$ are their mathematical expectations (mean values), respectively. This means that the probability of an event $y_j < Y_j \le y_j + dy_j, \forall j = 1,\ldots, m$ equals $p(y,c)$ $dy_1 \ldots dy_m$, where

$$p(y,c) := \left(\sqrt{2\pi}\,\sigma\right)^{-m} e^{-\sum_{j=1}^{m} \left(y_j - A_j(c)\right)^2 / 2\sigma^2}.$$

More formally, we assume we are dealing with the multidimensional random value $(Y_1,\ldots, Y_m)$ distributed with the Gaussian density $p(y,c)$ $dy_1 \ldots dy_m$ of the mean values' vector y, which is linked by a known law with an unknown parameter's vector: $\bar{y} = A(c)$. Given these conditions, the least-squares method yields an estimate of the value of parameters x such that $p(y,x)$ attains the maximum for a fixed given y.[11]

Fisher generalized Gauss's approach to the famous **maximal likelihood principle** of parameter estimation (Van der Waerden 1957; Cramér 1946).[12]

Just a few lines below we ask the reader to prove that Gauss's "maximal likelihood" estimate $x = \Phi(y)$ is **unbiased** (free of systematic errors); that is, considered as a random value (a function of random y), it has a "true" parameter as its mathematical expectation: $\int\limits_{\mathbf{R}^m} \Phi(y) p(y,c)\, dy_1 \ldots dy_m = c$.

Let Y be a (perhaps multidimensional) random value distributed with a density $\Omega(y)$ (say, the preceding $p\, dy_1 \ldots dy_m$). In the space of functions of Y that are square-integrable with Ω, consider a scalar product $\langle \varphi, \psi \rangle = \int \varphi(y)\psi(y)\Omega(y)$. Recall that a **covariance matrix** of $\varphi_1(Y),\ldots, \varphi_k(Y)$ has the entries $\mathrm{cov}_{ij} = \left\langle \varphi_i - \bar{\varphi}_i,\ \varphi_j - \bar{\varphi}_j \right\rangle$ [where $\bar{\varphi} = \int \varphi(y)\Omega(y)$ is the mathematical expectation]. Therefore, what follows is an exercise for readers familiar with a multivariate integration. Let $(Y_1,\ldots,Y_m)$ be distributed with Gaussian density $p\, dy_1 \ldots dy_m$ of the mean values linked by a known law with unknown parameters. Let it be a linear law $\bar{y} = Ac$, with a nondegenerate matrix A. *Establish* the unbiasedness of the maximal likelihood estimate $x = \Phi(y)$ for c and the coincidence of the covariance matrix for $\Phi_0(Y),\ldots, \Phi_n(Y)$ with $\sigma^2({}^tA \circ A)^{-1}$.

Gauss also proved that this estimate has the least variance over unbiased estimates: $\sum\limits_{i=0}^{n} \int\limits_{\mathbf{R}^m} (\Phi_i(y) - c_i)^2 p(y,c)\, dy_1 \ldots dy_m = \min = \sigma^2 \mathrm{tr}({}^tA \circ A)^{-1}$. Further development of the subject brought about in the twentieth century the famous **Frechét-Fisher-Cramér-Rao** (FFCR) lower bound for variances of unbiased parameter estimates under rather general conditions. That is, for a distribution $\Omega = \omega(y,c)\, dy_1 \ldots dy_m$, the bound equals tr G^{-1}, where G is a matrix of the entries $G_{ij} = \left\langle \partial \ln \omega / \partial c_i,\ \partial \ln \omega / \partial c_j \right\rangle$.

---

[11] The case where Y_j have distinct variances ${\sigma_j}^2$, attaining the maximum probability over x, for a fixed given y relates to the weighted least-squares formula $\sum (A_j(x) - y_j)^2 / {\sigma_j}^2 = \min$.

[12] Researchers using the least-squares and related principles in practice should never forget about their conditionality. Those principles will yield practical results as long as the observations are free of systematic errors, independent, and normally distributed.

The discoverers of this bound determined it as the largest eigenvalue of a quadratic form on the space of $(n + 1) \times (n + 1)$ matrices, $r(T,T) = (\mathrm{tr}\, T)^2$, with respect to a positive definite quadratic form on the same space, $q(T, T) = \mathrm{tr}(T \circ G \circ {}^tT)$: $\max\limits_{T \neq 0} \frac{r(T,T)}{q(T,T)} = \max\limits_{q(T,T)=1} r(T, T) = \mathrm{tr}\, G^{-1}$. In other words, this is the largest eigenvalue of the linear operator $L : T \mapsto LT$ defining r via q: $r(T,T) = q(LT,T)$, $\forall T$. [In reality, r is a factorable form, so it is of rank one (*why*?) and has a single nonzero eigenvalue; or L has a single nonzero eigenvalue; it just equals $\mathrm{tr}\, G^{-1}$, of the eigenline spanned on G^{-1}, since $LT = (\mathrm{tr}\, T) \cdot G^{-1}$.]

Readers may *verify* that $G = \sigma^{-2}\, ({}^tA \circ A)$ if $\omega = p$; so the least-squares parameter estimate has $\sigma^2\, ({}^tA \circ A)^{-1} = G^{-1}$ as its covariance matrix, and the FFCR bound is obtained in this case. And, generally, the unbiased estimate of the variance reaching FFCR bound, if it exists, must be $\Phi(Y) = G^{-1}(\partial \ln \omega / \partial c) + c$, so it will have G^{-1} as its covariance matrix (*verify*); however, those bounds are obtained with no unbiased estimates, unless we use special probabilistic densities. Interested readers will find discussions of this and related topics in Van der Waerden (1957) and Cramér (1946).

*P12.41****

Let us give a real-life example of signal processing in nuclear magnetic resonance (NMR) technology of composite analysis (see previous section P3.0 in the "A Combinatorial Algorithm in Multiexponential Analysis" problem group). A "true" NMR signal is a linear combination of exponential functions, $\bar{y}(t) = \sum\limits_{i=0}^{n} c_i e^{-t/T_i}$. In it, parameters T_i are referred to as relaxation terms; they correspond to different components and are assumed to be distinct. Coefficients c_i characterize the relative weights of the components in the composite and are referred to as amplitudes. The signal is registered at fixed moments $t_1, \ldots, t_m$ (called time echoes; they usually form arithmetic progressions $t_j = t_0 + j\Delta t$), which give random values $Y_j = Y(t_j)$. Assume these values to be independent, Gaussian, of equal variances σ^2, and to have "true" values $\bar{y}_j = \bar{y}(t_j)$ as their mathematical expectations. Let the relaxation terms be known exactly; it is desirable for us to find out how well we can estimate the amplitudes. For that, *find* the covariance matrix for the maximum likelihood estimate $x = \Phi(y)$ of the amplitudes' vector c for $n = 0$ and $n = 1$ ($m \geq n + 1$). Using it, *prove* the intuitively clear facts that the indistinctness of the exponential terms makes their amplitudes practically inestimable (estimable with infinite errors) and the estimates become entirely correlated (dependent).

Hint

H12.1

Indeed, the Chebyshev property means a unique (that is, trivial) solvability, with respect to c_i, of the linear systems $\sum\limits_i c_i f_i(t_j) = 0$: $t_0 < \ldots < t_n$.

An $n+1$-dimensional set $\{(t_0,\ldots,t_n)\in I^{n+1}: t_0<\ldots<t_n\}$ is arcwise connected (actually, its points are connectable by line segments; provide the details). In turn, since the function $\det(f_i(t_j))_{i,j=0,\ldots,n}$ is continuous, it will have a constant sign on the whole arc if it does not vanish in some point of it. (Why?)

Advanced readers may prefer the following argument. A real-valued continuous function f on a connected topological space A that does not have zero values retains the same sign throughout A. Indeed, a subset $B=\{f>0\}$ is open (because a continuous function cannot change its sign in a small neighborhood of a point) and closed (by the same argument, applied to a limit point of B), so $B=A$ or $B=\emptyset$.

For $F=\{1,\ldots,t^n\}$, $\det(f_i(t_j))_{i,j=0,\ldots,n}$ is the Vandermonde determinant. (Therefore, it is positive for $t_0<\ldots<t_n$.)

H12.2

The functions $t,\ldots,t^n$ form a T-system on $(0,\infty)$ but not on $[0,\infty)$.

H12.3

The functions $t+c$, $t+d$ with $c\neq d$ form a T-system on $(-\infty,\infty)$ together but not separately; therefore, $F=\{t+c,t+d\}$ is not an M-system on $(-\infty,\infty)$ (nor is its permutation $\{t+d,t+c\}$). $F=\{1,\ldots,t^n\}$ is not a D-system on $(-\infty,\infty)$.

H12.4

We must establish the positiveness of a generalized Vandermonde determinant

$$V_{n+1}(\vec{t},\vec{k})=\det\begin{pmatrix} t_0^{k_0} & \cdots & t_n^{k_0}\\ \vdots & & \vdots\\ t_0^{k_n} & \cdots & t_n^{k_n}\end{pmatrix}$$ for $0<t_0<\ldots<t_n$ and natural numbers $k_0<\ldots<k_n$. First, prove that $V_{n+1}(\vec{t},\vec{k})\neq 0$, in other words (by virtue of section P12.1*), that monomials t^{k_i} form a T-system on $(0,\infty)$. Readers may proceed as follows. Let $t_1<\ldots<t_l$ be the positive zeros of a polynomial $p=\sum c_i t^i$; we must prove that p has at least $l+1$ nonzero coefficients. Polynomial p is divided by the polynomial $q=\sum_{i=0}^{l} a_i t^i=\prod_{j=1}^{l}(t-t_j)$, so $p=qr$, with $r=\sum b_i t^i$,

where the coefficients of p and r are bound by a system of linear equations, which in a matrix form looks like $Ab = c$:

$$\begin{pmatrix} a_0 & & \\ \ddots & \ddots & \\ \ddots & a_1 & a_0 \\ \ddots & \ddots & \vdots \\ & a_l & a_{l-1} \\ & & a_l \end{pmatrix} \begin{pmatrix} b_0 \\ \\ \\ b_m \end{pmatrix} = \begin{pmatrix} c_0 \\ \\ \\ c_{l+m} \end{pmatrix}$$

($m = \deg r$). Show that the image of operator A has zero intersections with any coordinate subspaces of l or fewer nonzero coordinates c_i.

Second, establish, using the proved relations $V_{n+1} \neq 0$ for any n, the desired more exact relations $V_{n+1} > 0$ by induction on n (Polya and Szegö 1964; also, this inequality follows from a claim in section P12.10*).

Lastly, find that $\det \begin{pmatrix} \sum a_{0j} t_0^{k_{0j}} & \cdots & \sum a_{0j} t_n^{k_{0j}} \\ \vdots & & \vdots \\ \sum a_{nj} t_0^{k_{nj}} & \cdots & \sum a_{nj} t_n^{k_{nj}} \end{pmatrix} > 0$ ($0 < t_0 < \ldots < t_n$, $a_{ij} > 0$, $k_{i_1 j_1} < k_{i_2 j_2}$ if $i_1 < i_2$) using the multilinearity of determinants with respect to rows.

H12.5

Use the multilinearity of determinants with respect to rows, as in section P12.4*.

H12.6

The number of sign changes in the sequence of coefficients cannot exceed the number of coefficients decreased by 1. Hence, systems obeying the Descartes rule of signs are M-systems, and moreover, the Markov property remains following any permutations of the elements. (Why?) The same arguments show that obeying the Descartes rule of signs implies a similar property of D-systems, as established in section P12.5*: the (arbitrarily ordered) sets of polynomials $p_i = \sum_{j=k_{i-1}+1}^{k_i} a_j f_j$ ($\sum k_i = n$) with coefficients satisfying the condition $sign(a_{k_{i-1}+1}) = \ldots = sign(a_{k_i})$, $\forall i$

form M-systems. It is enough that they form T-systems because the Chebyshev property already shows that

$$\det\left(f_{i_0}(t_j)\ldots f_{i_{k-1}}(t_j)\ \ \alpha f_{i_k}(t_j)+(1-\alpha)f_{i_{k+1}}(t_j)\ \ f_{i_{k+2}}(t_j)\ldots f_{i_{m+1}}(t_j)\right)_{j=0,\ldots,m}\neq 0$$
$$(t_0<\ldots<t_m)$$

for $0\leq\alpha\leq 1$. On the other hand, this determinant continuously depends on α, so it has the same sign for $\alpha=0$ and $\alpha=1$. (Why? Use the Bolzano-Cauchy theorem for your answer.) Therefore, the Descartes rule of signs implies the Descartes property. Conversely, a polynomial on a D-system is, according to section P12.5*, a polynomial on a T-system formed by the sums of the terms corresponding to the sections with constant signs of coefficients. Thus, the number of its zeros cannot exceed the number of those sections diminished by 1, which completes the proof.

H12.7

$\det\left(K(s_i,t_j)\right)_{i,j\,=\,0,\ldots,\,n}$ has the same sign for all $s_0<\ldots<s_n$ and $t_0<\ldots<t_n$ by virtue of continuity. [Provide the details, which are similar to those in section H12.1; the only difference is that you will have to deal with a function on a $2n+2$-dimensional arcwise connected set $\{(s_0,\ldots,s_n,t_0,\ldots,t_n)\in I^{n+1}\times J^{n+1}\colon s_0<\ldots<s_n, t_0<\ldots<t_n\}$.]

H12.8

Indeed, the property of being an ET-system means a unique (trivial) solvability, with respect to c_i, of linear systems corresponding to any $t_0\leq\ldots\leq t_n$, which consist, for t_j of multiplicity r_j, of the equation groups $\sum_i c_i f_i(t_j)=0,\ldots,\sum_i c_i f_i^{[r_j-1]}(t_j)=0$.

To prove that the property of being an ET-system remains after multiplying all f_i by the same nonvanishing and n times differentiable function, use the following lemma.

Lemma *Let functions $f, f_0,\ldots,f_n$ be n times differentiable. Then*

$$W_{n+1}(f\cdot f_i)(t_j)_{i,j=0,\ldots,n}=\left(\prod_j f(t_j)\right)\cdot W_{n+1}(f_i)(t_j)_{i,j=0,\ldots,n}.$$

The reader can obtain this through direct computation using the multilinearity and skew symmetry of the determinants with respect to columns.

H12.9

The system $1,\ldots,t^{n-1}, f(t)$ is an ET-system because if a polynomial on it has at least $n+1$ zeros, considering multiplicity, then the nth derivative of this polynomial, which is equal to $f^{[n]}$, would have at least one zero (by the Rolle theorem) (Krein and Nudelman 1973).

The proofs of the analogs of the statements in sections P12.5* and P12.6* for ED-systems are similar to those in sections H12.5 and H12.6, using $W_{n+1}(f_i)(t_j)_{i,j=0,\ldots,n}$ instead of $\det(f_i(t_j))_{i,j=0,\ldots,n}$.

H12.10

We will prove that exponential functions $e^{s_0 t},\ldots,e^{s_n t}$ form an ED-system for any n and $s_0 <\ldots< s_n$ if we can prove that they form an ET-system for any n and $s_0 <\ldots< s_n$. (Why? Use the continuity of the determinants to derive that the signs of those determinants are preserved for any $s_0 <\ldots< s_n$ and $t_0 <\ldots< t_n$.) Thus, we must show that a polynomial $f=\sum_{i=0}^{n} c_i e^{s_i t}$ has at most n real zeros, considering multiplicity. $e^{-s_0 t} f = c_0 + \ldots + c_n e^{(s_n - s_0)t}$ has the same number of zeros as f, so, by the Rolle theorem, its derivative will have the number of zeros decreased by at least 1 (considering multiplicity). Complete this proof by induction using the fact that the number of exponential terms in the derivative is smaller by 1 than the number of terms in the function since one of these terms is constant.

Show that $t^{s_0},\ldots,t^{s_n}$ form an ED-system on $(0,\infty)$ by inserting $\log t$ for t into all $W_{n+1}(e^{s_i t})(t_j)_{i,j=0,\ldots,n}$ and using the following lemma.

Lemma *Let the functions $g, f_0,\ldots,f_n$ be n times differentiable. Then*

$$W_{n+1}(f_i \circ g)(t_j)_{i,j=0,\ldots,n} = \left[\prod_k \left(g'(t_{j_k})\right)^{\binom{\kappa_{j_k}}{2}}\right] \cdot W_{n+1}(f_i)\left(g(t_j)\right)_{i,j=0,\ldots,n},$$

where κ_j are multiplicities of t_j occurring in the sequence $t_0 \leq\ldots\leq t_n$ and the product is taken over representatives of all subsets of equal t_j.

(*An equivalent formulation* (*or, to be absolutely accurate, a corollary*) *is as follows*: the property of being an ET- or EM-system remains following independent variable substitution provided that the substitution function is n times differentiable and its first derivative does not go to zero; the same will hold for the property of being an ED-system, provided that this derivative is positive.)

The reader can obtain this by a direct computation differentiating compositions of functions and using the multilinearity and skew symmetry of the determinants with respect to columns.

H12.11

The Wronskian of exponential functions is proportional to a Vandermonde determinant. (Work out the details.)

Proof of Statement 1. Taking into account the claim in section P12.8*, we must prove only sufficiency. By the condition, f_0 does not have zero values (why?), so, by virtue of the lemma from section H12.8, $W_{m+1}(g_0,\ldots, g_m) \neq 0$ for $0 \leq m \leq n$, with $g_i = f_0^{-1} f_i$. Employ the Rolle theorem to complete the proof by induction on n.

Proof of Statement 2. Here we must also prove only sufficiency. Applying Statement 1 yields $W_{m+1}(f_{i_0},\ldots,f_{i_m})(t_0,\ldots,t_m) \neq 0$ for $t_0 \leq \ldots \leq t_m$ and $0 \leq i_0 < \ldots < i_m \leq n$. Therefore, the only step remaining is to verify the implication as follows: if the sign of $W_{m+1}(f_{i_0},\ldots,f_{i_m})(t)$ is constant for any fixed t and all sequences $0 \leq i_0 < \ldots < i_m \leq n$, then the sign of $W_{m+1}(f_{i_0},\ldots,f_{i_m})(t_0,\ldots,t_m)$ is constant for any $t_0 \leq \ldots \leq t_m$ and $0 \leq i_0 < \ldots < i_m \leq n$.

Proof of Statement 3. Applying Statement 1 yields $W_{m+1}(f_{i_0},\ldots,f_{i_m})(t_0,\ldots,t_m) \neq 0$ for $t_0 \leq \ldots \leq t_m$ and $0 \leq i_0 < \ldots < i_m \leq n$. Therefore, for any $t_0 < \ldots < t_m$ and $0 \leq i_0 < \ldots < i_m \leq n$, the sign is constant due to the continuity of $K(s,t)$. (Fill in the details.) Extend the preceding statement to any $t_0 \leq \ldots \leq t_m$ by using the lemma in section E12.11 (proof of Statement 2) and taking into account that the determinants do not vanish.

H12.12

Verify that the Wronskian of $g_0,\ldots,g_{n-1}$ satisfies the sufficient conditions from Statement 2 in section P12.11*. (In fact, it has already been discussed in sections H12.11 and E12.11: analyze the proof of Statement 1.)

H12.13

The conditions for Wronskians that are equivalent to the property of being a D-system (resp. ED-system) imply that, for $m = 0$, the signs of $W_1(f_i) = f_i$ are the same for all i, and for $m = 1$ the signs of $W_2(f_i,f_j) = f_i^2 \cdot (d/dt)(f_i(t)/f_j(t))$ are the same for all $i < j$.

H12.14

Given a polynomial on $\{f_0,\ldots\}$, group together its adjacent terms that have coefficients of the same sign. By analogy with the claim from section P12.5* for ED-systems, the resulting polynomial is defined on the new ED-system (of grouped together terms), with adjacent coefficients having opposite signs. Obviously, such grouping preserves W and N. Verify that this grouping preserves the conditions of the statement that is being proved.

Case of $W = 0$. $N = 0$ (as $0 \le N \le W$), so that $W - N$ is even.

Case of $W = 1$. We are dealing with a polynomial $p = \pm (f_0 - f_1)$. Use section P12.13* to show that under the conditions of section P12.14* a function $\pm p/f_0 = 1 - f_1/f_0$ on I has N zeros, decreases, and tends to 1 and $-\infty$ at the left and right ends of I, respectively. Therefore, $N = 1$, so that $W - N$ is even.

Case of $W > 1$. Consider the polynomial after its division by a middle term: $q = \frac{p}{\pm f_m} = \ldots - \frac{f_{m-1}}{\pm f_m} + 1 - \frac{f_{m+1}}{\pm f_m} + \ldots$. $q(t)$ is W times differentiable. Use section P12.13* to show that, under the conditions of section P12.14*, $|q(t)| \to \infty$ at the ends of I. Replacing $q(t)$ by $q(t) - 1$ diminishes by 2 the number of sign changes in a series of the coefficients. Prove that this substitution cannot affect the number of zeros (considering multiplicity) modulo 2 and complete the proof by induction on W.

H12.15

A sequence of quasimonomials $\left\{e^{\lambda_1 t}, \ldots, e^{\lambda_1 t} t^{\deg p_1}, \ldots, e^{\lambda_n t}, \ldots, e^{\lambda_n t} t^{\deg p_n}\right\}$ is an EM-system on $I = \mathbb{R}$ by Statement 1 in section P12.11*. Indeed, Wronskians of the first m elements of this sequence $[1 \le m \le \sum(\deg p_i + 1)]$ cannot vanish since those functions form fundamental bases for solutions of systems of (autonomous) linear differential equations.

Readers should analyze this proof and verify its inapplicability to quasipolynomials containing trigonometric functions, which may have infinitely many real zeros!

H12.16

Wronskians of the Cauchy kernel are proportional to a Vandermonde determinant: for $f_i(t) = (s_i + t)^{-1}$, we will have

$$W_{n+1}(f_0,\ldots,f_n)(t) \propto \prod (s_i+t)^{-1} \cdot \det \begin{pmatrix} 1 & \cdots & (s_0+t)^{-n} \\ \vdots & & \vdots \\ 1 & \cdots & (s_n+t)^{-n} \end{pmatrix}.$$

A weaker property of the Cauchy kernel – that this kernel is a kernel of regular sign in the sense of section P12.7* – can be established without using Wronskians. Indeed, this follows directly from an elementary algebraic identity

$$\det\left((s_i+t_j)^{-1}\right)_{i,j=0,\ldots,n} = \prod_{0\le i\le n}(s_j-s_i)(t_j-t_i) \Big/ \prod_{i,j=0}^{n}(s_i+t_j)$$

when $s_i + t_j \neq 0$, $\forall i,j$ (see section E1.4, in the previously discussed "Jacobi Identities and Related Combinatorial Formulas" problem group).

Similar arguments are applicable to the Gaussian kernel. A short calculation yields $\frac{d^k}{dt^k}e^{-t^2} = e^{-t^2}p_k(t)$, where p_k is a polynomial of degree k with integral coefficients depending only on k. Using the multilinearity and skew symmetry of determinants with respect to columns, we have, for $f_i(t) = e^{-(s_i+t)^2}$,

$$W_{n+1}(f_0,\ldots,f_n)(t) \propto e^{-\sum (s_i+t)^2} \cdot \det \begin{pmatrix} 1 & \cdots & (s_0+t)^{n} \\ \vdots & & \vdots \\ 1 & \cdots & (s_n+t)^{n} \end{pmatrix},$$

as was just stated.

H12.17

Verify that $\theta + \pi$ will be a root of a polynomial $\sum c_i h_i$ if θ is also a root. Because of the linear independence of h_i,[13] $\sum c_i h_i$ vanishes identically if and only if $c_i = 0$, $\forall i$. Therefore, the claim from section P12.17** follows from the following proposition.

Proposition *A trigonometric polynomial of order n,* $\sum_{i+j\le n} c_{ij}(\cos\theta)^i(\sin\theta)^j$, *that does not vanish identically may have at most* $2n$ *zeros in* $[0, 2\pi)$, *considering multiplicity.*

We suggest proving this proposition using either of the following two methods:

(1) Algebraic method: Apply the **Bézout theorem**, which states that two algebraic curves defined by the polynomial equations $f(x,y) = 0$, $g(x,y) = 0$ of degrees m, n, respectively, intersect by, at most, mn points, considering multiplicity, or have a common component. (Readers not familiar with this theorem should consider section E12.17.) In our case, one of the equations defines the cosine and sine, $f = x^2 + y^2 - 1$, and the second one is $g = \sum_{i+j\le n} c_{ij}x^i y^j$. Therefore,

[13] Readers may prove this on their own or refer to Lemma 1 in E12.39 below.

for a homogeneous polynomial $g = \sum_{i+j=n} c_{ij} x^i y^j$, the number of common zeros cannot exceed $2n = \deg f \cdot \deg g$ (because f is an **irreducible** polynomial, which cannot divide a homogeneous polynomial; geometrically, the curve $x^2 + y^2 = 1$ consists of one component (the unit circle), and a nonzero homogeneous polynomial cannot vanish on all of it; work out the details).

(2) Analytic method: Make use of an equivalent representation of the trigonometric polynomials by the Fourier polynomials $\sum_{k=0}^{n} a_k \cos k\theta + b_k \sin k\theta$ with real coefficients.[14] Prove that a nonzero Fourier polynomial may have at most $2n$ zeros, considering multiplicity, in $[0, 2\pi)$ [thus, the harmonics 1, cos θ,..., cos $n\theta$, sin θ,..., sin $n\theta$ form an ET-system on $[0, 2\pi)$]. Do this as follows (Polya and Szegö 1964). A complex polynomial $p(z) = \sum u_{n-i} z^i$ of degree n is called **reciprocal** if $p(z) = z^n \bar{p}(z^{-1})$, in other words, simultaneously substituting z by z^{-1} and the coefficients by their conjugates and multiplication by z^n do not affect $p(z)$. (An equivalent requirement consists in $u_i = \bar{u}_{n-i}$, $\forall i = 0, \ldots, n$.) Prove that the trigonometric polynomials $G(\theta)$ of order n with real coefficients, and only these polynomials, are presentable as $G(\theta) = e^{-in\theta}\, p(e^{i\theta})$, with a reciprocal polynomial $p(z)$ of degree $2n$. However, $p(z)$ has $2n$ complex roots, considering multiplicity, and, consequently, at most $2n$ of them belong to a unit circle. Therefore, $G(\theta)$ has at most $2n$ zeros in $[0, 2\pi)$, considering multiplicity.

H12.18

A desired polynomial is arrived at by solving the system of $m < n + 1$ homogeneous linear equations on $n + 1$ unknown coefficients $\sum_{i=0}^{n} c_i f_i(t_j) = 0, j = 1,..., m$. For a T-system, the $m \times m$ minors of the matrix for this system cannot vanish all together. (Why?) Therefore, the solutions form an $n - m + 1$-dimensional space. **QED**. [For $m = n$, an explicit formula (up to proportionality) is found decomposing $\det (f_i(t)\, f_i(t_1) \ldots f_i(t_n))_{i=0,\ldots,n}$ with respect to its first column.] Geometrically, this is equivalent to the unique existence of an m-dimensional vector subspace in $\mathbb{R}^{n+1}$ having common points with a curve $t \mapsto (f_0(t), \ldots, f_n(t))$ for $t = t_1,..., t_m$. (Why?) (For $m = n$ and continuous f_i, this curve "pierces through" that plane at the common points, as follows from the statements in section P12.20*** below.)

[14] In this representation, the space of homogeneous polynomials $\sum_{i+j=n} c_{ij} (\cos\theta)^i (\sin\theta)^j$ corresponds to the space of Fourier polynomials $\sum_{0 \le k \le n,\ k \equiv n \bmod 2} a_k \cos k\theta + b_k \sin k\theta$. (Why? Compare the behavior of the trigonometric polynomials to that of the Fourier polynomials with respect to the involution caused by the substitution $\theta \mapsto \theta + t$.)

H12.19

Apply the Kronecker-Capelli theorem in its geometric version. Consider a vector space Y of functions on an $n+1$-point set $\{t_0,\ldots, t_n\}$ and a linear operator $C \xrightarrow{A} Y$ mapping a vector of coefficients $(c_0,\ldots, c_n)\in C$ onto $(\sum c_i f_i(t_0),\ldots, \sum c_i f_i(t_n)) \in Y$. The existence of a polynomial taking values $\sum c_i f(t_j) = y_j$, respectively, is the same as the solvability of the equation $Ac = y$ ($y = (y_0,\ldots, y_n)$). The solvability with any right-hand side $y\in Y$ means that $\operatorname{im} A = Y$; in turn, uniqueness (for $y \in \operatorname{im} A$) means that $\ker A = 0$. Since $\dim C = \dim Y$, existence with any y is equivalent to uniqueness; therefore, existence with any $y = (y_0,\ldots, y_n)$, referred to any distinct $t_0,\ldots,t_n$, is equivalent to the Chebyshev property (according to section P12.1*).

Similar arguments accounting for the existence of nonvanishing $m \times m$ minors (section H12.18) show that for a T-system F a linear operator $C \xrightarrow{A} Y_m$ onto the space of functions on an m-point set ($m \le \#F$) has rank m. Therefore, the polynomials on F taking given values $y_0,\ldots,y_{m-1}$ at distinct points $t_0,\ldots,t_{m-1}$, respectively, form a $(\#F - m)$-dimensional affine plane in the space C.

The unique existence of the polynomial $p(t)$ taking values $y_0,\ldots,y_n$ at distinct points $t_0,\ldots,t_n$ is equivalent to the unique existence of a hyperplane of the form $y = \sum c_i x_i$ in $\mathbb{R}^{n+2}_{x_0,\ldots,x_n,y}$ containing the points $(f_0(t_j),\ldots,f_n(t_j), y_j)$, $j = 0,\ldots, n$. [That is, c_i must be the coefficients of $p(t)$.]

The previously indicated representation of the Lagrange polynomial on $\{1,\ldots, t^n\}$ is obtained by a direct verification that the polynomials $p_j(t) = \prod\limits_{i\ne j} \frac{t-t_i}{t_j-t_i}$ (restricted to $\{t_0,\ldots,t_n\}$) form the required δ-basis.

H12.20

A curve $t\mapsto(f_0(t),\ldots,f_n(t))$ in $\mathbb{R}^{n+1}$ "pierces through" a hyperplane $\sum c_i x_i = 0$ at the nodal zeros of a polynomial $\sum c_i f_i$ and remains on the same side of this plane in a whole neighborhood of the nonnodal ones.

Proof of theorem (Krein and Nudelman 1973). Necessity. Let $t_1,\ldots,t_k$ and $t_{k+1},\ldots, t_{k+l}$ be nonnodal and nodal zeros of a polynomial $p(t)$. Find a Lagrange interpolating polynomial $q(t)$ such that $p(t) - q(t)$ would be nonzero and have zeros $t_{k+1},\ldots,t_{k+l}$ and two zeros in a small neighborhood of each $t_1,\ldots,t_k$. By virtue of the Chebyshev property we have $2k + l \le n$.

Sufficiency. Case of $n \equiv l \bmod 2$. First assume that $l = 0$ and $n = 2k$. Take a convergent sequence $\left(t_1^{(\nu)},\ldots, t_k^{(\nu)}\right) \to (t_1,\ldots, t_k)$, as $\nu \to \infty$, such that $t_j^{(\nu)} < t_j$, $\forall j$, and define the polynomials $p^{(\nu)} = \sum c_i^{(\nu)} f_i$ with n distinct zeros $t_1^{(\nu)} < t_1 < \ldots {}^{(\nu)}_{k} < t_k$ (for a fixed ν, all of these polynomials, according to section P12.18*, are

proportional), normalized by two conditions: (1) $p^{(v)}(t)$ have the same sign, say, $p^{(v)}(t) > 0$, for $t_1^{(v)}$, and (2) $\sum \left(c_i^{(v)}\right)^2 = 1$. (Provide a figure.) Prove the existence of a limit polynomial and show that it will have nonnodal zeros $t_1,\ldots,t_k$ and no others.

Now assume that $l = 0$ and $n = 2m > 2k$. Using arguments similar to the preceding ones define a nonnegative polynomial $p'(t)$ with the required nonnodal zeros $t_1,\ldots,t_k$ and extra nonnodal zeros $t'_{k+1},\ldots,t'_m$, located as follows: $t_k < t'_{k+1} < \ldots < t'_m$. Similarly, define a nonnegative polynomial $p''(t)$ with the required nonnodal zeros $t_1,\ldots,t_k$ and extra nonnodal zeros $t''_{k+1},\ldots,t''_m$, located as follows: $t'_m < t''_{k+1} < \ldots < t''_m$. Obviously, the polynomial $p' + p''$ has nonnodal zeros $t_1,\ldots,t_k$ and no others. (Provide a figure.)

Finally, assume that $l > 0$. The only difference with the former situations is that polynomials $p^{(v)}(t)$ are defined as possessing nodal zeros $t_{k+1},\ldots,t_{k+l}$ as well, and therefore they will have n distinct zeros; a limit polynomial will also have nodal zeros $t_{k+1},\ldots,t_{k+l}$. (Provide a figure and complete the details.)

Sufficiency. Case of $n \not\equiv l \bmod 2$ and $I = [a,b]$. First assume that $2k + l = n - 1$. There are three possibilities: (1) $l = 0$ or $t_{k+1},\ldots,t_{k+l} \in (a,b)$; (2) one t_{k+j} coincides with an end of I (a or b); or (3) two of them coincide with the ends.

In the first case, define $p'(t)$ with nonnodal zeros $t_1,\ldots,t_k$, nodal zeros $t_{k+1},\ldots,t_{k+l}$, and a. Define $p''(t)$ with nonnodal zeros $t_1,\ldots,t_k$, nodal zeros $t_{k+1},\ldots,t_{k+l}$, and b so that $p'(t)$ and $p''(t)$ would have equal signs in a right half-neighborhood of a. The polynomial $p' + p''$ has nonnodal zeros $t_1,\ldots,t_k$, nodal zeros $t_{k+1},\ldots,t_{k+l}$, and no others. (Provide a figure and fill in the details.)

In the third case, define a sequence of polynomials $p^{(v)} = \sum c_i^{(v)} f_i$ with $\sum \left(c_i^{(v)}\right)^2 = 1$, nonnodal zeros $t_1,\ldots,t_k$, nodal zeros $t_{k+1},\ldots,t_{k+l}$, and an extra nodal zero $t^{(v)} \to b$, so that they would have equal signs in a right half-neighborhood of a (which is possible, as we have just proved). A limit polynomial will have nonnodal zeros $t_1,\ldots,t_k$, nodal zeros $t_{k+1},\ldots,t_{k+l}$ and no others.

The second case is the hardest nut to crack. For simplicity, let $k = 0$. For definiteness, assume that $t_1 = a$. We might take a Lagrange polynomial having nodal zeros $t_1,\ldots,t_l$ (and taking some values at two extra points), but we cannot be certain that an extra nodal zero does not exist because its existence is not prohibited by the "sum of indices condition," as $2k + l < n$ in this case! Similarly to the first case, we might define a sequence of polynomials $p^{(v)} = \sum c_i^{(v)} f_i$ with $\sum \left(c_i^{(v)}\right)^2 = 1$, nodal zeros $t_2,\ldots,t_l$, extra nodal zero $t_1^{(v)} \to a$, and no other zeros (so with equal signs in a left half-neighborhood of b) and take a limit polynomial $p(t)$. This brings us close to a solution. A remaining problem is that we cannot control $p(b)$; meanwhile, a nodal zero at b is not prohibited by the "sum of indices condition." Try avoiding this problem as follows. Normalize $p^{(v)}$ so as $p^{(v)}(b) = 1$, instead of $\sum \left(c_i^{(v)}\right)^2 = 1$, which will imply $p(b) = 1$, if, of course, a limit polynomial exists. When does it exist? Consider an affine straight line $\Lambda^{(v)} \subset \mathbb{R}^{n+1}_{c_0,\ldots,c_n}$ defined by the

equations $p(t_1^{(v)}) = p(t_2) = \ldots = p(t_l) = 0$, $p(b) = 1$. [$\Lambda^{(v)}$ is a shift of a one-dimensional vector subspace defined by linear homogeneous equations $p(t_1^{(v)}) = p(t_2) = \ldots = p(t_l) = p(b) = 0$.] Prove that a distance from the origin to $\Lambda^{(v)}$ [in other words, the minimum of $(\sum {c_i}^2)^{1/2}$ over the elements of $\Lambda^{(v)}$] is bounded above, uniformly on v, so a sequence of $p^{(v)}$ corresponding to those minima has a limit polynomial. (Work out the details.)

For $k > 0$, take $t_{k+1}^{(v)} \to t_{k+1} = a$ and a sequence $\left(t_1^{(v)}, \ldots, t_k^{(v)}\right) \to (t_1, \ldots, t_k)$ such that $t_j^{(v)} < t_j$, $\forall j$, and define $p^{(v)} = \sum c_i^{(v)} f_i$ corresponding to the minimum of the distance from the origin of $\mathbb{R}^{n+1}_{c_0,\ldots,c_n}$ to an affine straight line

$$\Lambda^{(v)} = \left\{ p(t_1^{(v)}) = p(t_{k+1}^{(v)}) = p(t_1) = \ldots = p(t_k) = p(t_{k+2}) = \ldots \right.$$
$$\left. = p(t_{k+l}) = 0, \quad p(b) = 1 \right\}.$$

(First prove the uniform boundedness of these distances.) A limit polynomial will have nonnodal zeros $t_1,\ldots,t_k$, nodal zeros $t_{k+1},\ldots,t_{k+l}$, and no others.

For $2k + l < 2m + l = n - 1$, define the required polynomial using extra nonnodal zeros as was done previously.

H12.21

$a_i := \inf \{t \in (t_i, t_{i+1}): f \le 0 \text{ on } [t, t_{i+1}]\}$, $b_i := \sup \{t \in (t_i, t_{i+1}): f \ge 0 \text{ on } [t_i, t]\}$ are the required points. Indeed, by this definition, there are points where $f > 0$ in any left half-neighborhood of a_i and where $f < 0$ in any right half-neighborhood of b_i, $f \ge 0$ on $[t_i, a_i]$ and $f \le 0$ on $[b_i, t_{i+1}]$; hence, $a_i \le b_i$ (as otherwise f changes sign at least $n + 1$ times; why?), and so, $f = 0$ on $[a_i,b_i]$. Obviously, all the points of $[a_i,b_i]$ are f's nodal zeros, while the remaining points of (t_i,t_{i+1}) are not (because the sign of f is constant in their neighborhoods).

Turning to Hurwitz's theorem, assume that f changes sign $k < n$ times in (a,b) and $t_i \in [a_i,b_i]$ $(i = 0,\ldots, k - 1)$ are nodal zeros from the preceding node's segments. If f does not vanish identically, then neither does $f \cdot \prod(t - t_i)$, but this product retains its sign, which yields $\int_a^b f(t) \cdot \prod (t - t_i)\, dt \ne 0$, in contradiction with the orthogonality conditions (Polya and Szegö 1964). The generalization of the Hurwitz theorem for T-systems is proved similarly by substituting $\prod(t - t_i)$ by a polynomial on a T-system with nodal zeros $t_0,\ldots,t_{k-1}$ and no other zeros (such a polynomial is produced by the theorem in section P12.20***). (Work out the details.)

Similar arguments are applicable by substituting the integrals by any **positive** (linear) **functionals**. A linear functional λ on a functional space is called positive if $\lambda(f) > 0$ for the elements f that do not take negative values and do not go to zero identically. This topic, related to the classic **moment problem**, is discussed in Akhiezer (1961) and Gantmacher and Krein (1950).

H12.22

The trigonometric harmonics 1, cos t, sin t do not form a T-system on a segment $[t_0, t_0 + 2\pi]$ of length 2π since a polynomial $\sin(t - t_0) = -\sin t_0 \cos t + \cos t_0 \sin t$ has three distinct zeros there. At the same time, those functions form a T- (even an ET-) system on the unit circle. For a detailed proof, readers may refer to section P12.23* below or argue as follows:

$$c + a\cos t + b\sin t = c + \sqrt{a^2 + b^2} \cdot \sin\left(t + arctg\frac{a}{b}\right),$$

which has at most two zeros, considering multiplicity, on a half-interval of length 2π. For a periodic T-system we have

$$\det \begin{pmatrix} f_0(a) & f_0(t_1) & \cdots & f_0(t_n) \\ \vdots & \vdots & & \vdots \\ f_n(a) & f_n(t_1) & \cdots & f_n(t_n) \end{pmatrix}$$

$$= (-1)^n \det \begin{pmatrix} f_0(t_1) & \cdots & f_0(t_n) & f_0(b) \\ \vdots & & \vdots & \vdots \\ f_n(t_1) & \cdots & f_n(t_n) & f_n(b) \end{pmatrix} \quad (a<t_1<\ldots<t_n),$$

and at the same time the determinants on both sides are equal for a system consisting of functions that are continuous on $[a,b]$. [Fill in the details. ***Warning***: the determinant $\det\left(f_i(a)f_i(t_1)\ldots f_i(t_{n-1})f_i(b)\right)_{i=0,\ldots,n}$ vanishes.]

H12.23

See section H12.17.

H12.24

The property of having an even number of sign changes on a circle, if this number is finite, is connected to the fact that the first point of a finite sequence on a circle is adjacent to the last one. The remaining statements are proved similarly to section H12.21, with one nuance. The theorem from section P12.20*** (what makes its application legitimate here?) assures at least $2n - 1$ changes of sign, and the requirement of evenness raises this lower estimate to $2n$. [Complete the details; also, a similar proof is in Arnol'd (1994).]

H12.25

This statement is a special case of the one from section P12.24***, taking into account section P12.23* and the orthogonality relations

$$\int_0^{2\pi} \cos m\theta \cdot \sin n\theta \, d\theta = 0,$$

$$\int_0^{2\pi} \cos m\theta \cdot \cos n\theta \, d\theta = \int_0^{2\pi} \sin m\theta \cdot \sin n\theta \, d\theta = 0 \quad \text{for } m \neq n \ (m, n = 0, 1, \ldots).$$

Actually, this special case can be obtained using a simpler technique than the general one, without the theorem from section P12.20***. Indeed, keeping to the same plan as in section H12.24, a trigonometric polynomial of order $k \leq N$, with any given $2k$ nodal zeros and no other zeros in $[0, 2\pi)$, may be obtained using the statements in sections P12.23* and P12.18*. (We recommend that t readers directly construct a trinomial $c_0 + c_1 \cos \theta + c_2 \sin \theta$ with prescribed zeros $0 \leq \theta_1 < \theta_2 < 2\pi$.)

A proof of this kind was published in a work of Tabachnikov (Arnol'd 1994 and references therein). Hurwitz's proof, using essentially similar arguments, but also supplementary computations with trigonometric functions, is in Polya and Szegö (1964). Also, we suggest that readers familiar with elements of complex analysis find a proof connected to quite different ideas (belonging to Hurwitz). Let us restrict our consideration to Fourier series with coefficients decaying at a rate of geometric progression [Hurwitz himself considered trigonometric polynomials (Polya and Szegö 1964)]. The sum of such a series is a restriction to a unit circle of the real part of a complex-analytic function defined in an origin-centered disc of radius greater than 1; this function is given as $f(z) = \sum_{n \geq N} (a_n - ib_n)z^n$. (Verify it!) If $f(z)$ has no (complex) zeros on a unit circle, a vector $f = u + iv$ makes, by the **argument principle** of complex analysis, at least N full counterclockwise turns, so it will be at least $2N$ times on the imaginary axis (which corresponds to the vanishing of the real part). However, if $f(z)$ has (complex) zeros on a unit circle, we may go around each of them by half-arcs of small radii. Note that on those arcs f rotates in the opposite (clockwise) direction (create a figure). Therefore, on the rest of the contour, lying on a unit circle, f makes even more than N full counterclockwise turns. We leave the details to the reader. Also, we leave it up to the reader to determine the limits of the method.

H12.26

A continuous function bounded below and growing on infinity on a finite-dimensional normed space has a minimum on any closed set has a minimum. For finite F, the distance from f to the elements of the vector space spanned on F is such a function, so that it has a minimum on a closed set H (this minimum is positive when $f \notin H$); therefore, a polynomial of best approximation exists. On the other hand,

a polynomial on $F = \{t^n: n = 0,1,2,\ldots\}$, of best approximation for f on I, does not exist unless $f|_I$ itself is a polynomial since a residual $f - p$ of its polynomial approximation can be made as small as desired on all of a fixed compact set (Weierstrass approximation theorem).[15] p is a polynomial of best approximation if and only if $f - p$ is an element of an affine hyperplane $f + H$ of the least deviation from zero on I because of the *(linear) translation-invariance property* of the distance: dist (f, g) = dist $(f + h, g + h)$, $\forall f, g, h$ (since obviously $\max_{t\in I} |f(t) - g(t)| = \max_{t\in I} |(f(t) + h(t)) - (g(t) + h(t))|$).

The **distance functions** defined by norms [by the formula dist $(f, g) = \|f - g\|$] are translation-invariant. Readers can easily find examples of invariant distance functions that cannot be defined in this way and also of distance functions that are noninvariant.

H12.27

Sufficiency. If $\max_{a\le t\le b} |f(t) - q(t)|$ for a polynomial $q(t)$ on F, then $p(t) - q(t)$ takes, in turn, positive and negative values at points $t_0 < \ldots < t_{\#F}$, so they would have at least $\#F$ distinct zeros in $[a,b]$.

Necessity. Let m be a maximal number of points $a \le t_0 < \ldots < t_{m-1} \le b$ such that $f(t) - p(t)$ takes values $+d$ and $-d$ in turn at these points. For definiteness, let $f(t_0) - p(t_0) = -d$. Define the points $a = t'_0 < \ldots < t'_m = b$ $(t_{j-1} < t'_j < t_j, j = 1,\ldots, m - 1)$ so that $f(t) - p(t) < d$ on $[t'_{2j}, t'_{2j+1}]$ and $f(t) - p(t) > -d$ on $[t'_{2j+1}, t'_{2j+2}]$. If $m < n + 2$, then there is (by the theorem in section P12.20***) a polynomial $q(t)$ on F with nodal zeros $t'_1, \ldots, t'_{m-1}$ and no others, negative on $[a, t'_1)$. Verify that $\max_{a\le t\le b} |f(t) - p(t) - \varepsilon q(t)|$ for small $\varepsilon > 0$, so $p(t)$ cannot be the polynomial of best approximation.

A proof of the Chebyshev theorem close to the original is in Berezin and Zhidkov (1959).

H12.28

Sufficiency. The uniqueness of a polynomial of best approximation is not a trivial fact because the balls related to the defined distance between continuous functions are not strictly convex (the spheres contain line segments).[16] (Give examples of

[15] Readers familiar with a considerably more general **Stone-Weierstrass theorem** can provide multiple further examples using many other infinite sets F.

[16] Readers may verify, for the distance function defined by a norm, that any line segments contained in the ball lie in its interior (perhaps except the ends) if and only if the triangle inequality is strict: dist (x,y) < dist (x,z) + dist (y,z) for noncollinear $x - z$ and $x - y$.

affine planes of any dimensions of continuous functions on I containing entire convex planar regions of elements of least deviation from zero.) For T-systems, the uniqueness follows from the Chebyshev theorem. To use it, verify that if p, q are polynomials of best approximation, the same will hold for their convex linear combinations. Then show that $p = q$ by applying the Chebyshev theorem to that combination.

Necessity. According to section H12.19, if $F = \{f_0,\ldots,f_n\}$ is not a T-system, then for the vector space L of functions on an appropriate $n + 1$-point set $\{t_0,\ldots,t_n\}$, a linear operator $C \xrightarrow{A} L$ that maps a vector of coefficients $(c_0,\ldots, c_n) \in C$ onto $(\sum c_i f_i(t_0),\ldots, \sum c_i f_i(t_n)) \in L$ is not an isomorphism. Therefore, there exists, on the one hand, a nonzero polynomial $p(t)$ on F vanishing on $\{t_0,\ldots,t_n\}$, and, on the other hand, a nonzero element $(y_0,\ldots, y_n) \in L$, orthogonal to $A(C)$ [that is, $\sum y_j q(t_j) = 0$ for the polynomials $q(t)$]. Scale $p(t)$ so that $\max\limits_{a \le t \le b} |p(t)| \le 1$. Let g be a continuous function with $\max\limits_{a \le t \le b} |g(t)| = 1$ and $g(t_j) = \text{sign } (y_j)$ $(j = 1,\ldots, n)$. Prove that any $cp(t)$ with $|c| \le 1$ are polynomials of best approximation for a function $f(t) = g(t) \cdot (1 - |p(t)|)$ on $[a,b]$.

A proof of the Haar theorem close to the original is in Berezin and Zhidkov (1959).

H12.29

If a polynomial $T(t)$ is of the least deviation from zero on $[-a,a]$ over the polynomials of even degree n with leading coefficient 1, then so is its even component $\frac{T(t)+T(-t)}{2}$. By virtue of uniqueness, $T(t) = \frac{T(t)+T(-t)}{2}$ is even. Similarly, $T(t) = \frac{T(t)-T(-t)}{2}$ is odd for odd n.

$T_n(t)$ may have at most $n - 1$ internal extrema since its derivative is a polynomial of degree $n - 1$.

H12.30

A function $\cos(n \arccos t)$ has $n + 1$ extrema on $[-1,1]$. (List them.) The required explicit representation (which shows that this function is a polynomial of degree n in t) can be established by induction on n using the recursion relation $T_{n+1} = 2t \cdot T_n - T_{n-1}$, $n = 1,2,\ldots$ (following an identity $\cos(n + 1)\theta = 2\cos\theta \cos n\theta - \cos(n - 1)\theta$).[17]

[17] More advanced readers may find a different proof based on the theory of spherical functions.

The preceding recursion relations on T_n may be written as a matrix equation

$$\begin{pmatrix} 0 & 2 & 0 & \dots & & \\ 1 & 0 & 1 & 0 & \dots & \\ 0 & 1 & 0 & 1 & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} T_0(t) \\ T_1(t) \\ T_2(t) \\ \dots \end{pmatrix} = 2t \cdot \begin{pmatrix} T_0(t) \\ T_1(t) \\ T_2(t) \\ \dots \end{pmatrix}$$

with the (infinite) matrix on the left-hand side belonging to the so-called **3-diagonal Jacobi matrix** family. Similar equations using other matrices of this family define series of **orthogonal polynomials** that are close to the Chebyshev ones in many respects. Readers may learn from Akhiezer (1961), Gantmacher and Krein (1950), Stoker (1950), and McLachlan (1964) about the ties of those polynomials with **continued fractions** and **positive functionals** and applications to the classic **moment problem** and **vibration theory** (including **Mathieu's equation theory**). In addition, familiarity with this subject will enable the reader to understand much better the engine of the proof of the Erdös-Mordell-Tóth geometric inequality ($\forall n \geq 3$) discussed previously in section H9.13 (the "Convexity and Related Classic Inequalities" problem group) and, especially, better understand how continued fractions are used there. Consider this exercise: find the eigenvalues of the $n \times n$ corner minors of the preceding 3-diagonal matrix; also, solve Problem P7.9** from the previously discussed "2 × 2 Matrices That Are Roots of Unity" problem group by a different method employing the Chebyshev polynomials.

H12.31

Since a polynomial $\bar{T}_n(t) = 2^{1-n}T_n(t)$ has a leading coefficient 1, $\mu_n(-1,1) = 2^{1-n}$. (What follows from this equality?) A linear map $[a,b] \xrightarrow{\lambda} [-1,1]$ extends the distances by $2/(b-a)$ times, so a polynomial $\bar{T}_n(\lambda(t))$ that deviates from zero by $\max\limits_{a \leq t \leq b} |\bar{T}_n(\lambda(t))| = 2^{1-n}$ has as its leading coefficient $[2/(b-a)]^n$. Using the Chebyshev and Haar theorems yields

$$\mu_n(a,b) = 2^{1-n} \Big/ \left(\frac{2}{b-a}\right)^n = 2 \cdot \left(\frac{b-a}{4}\right)^n.$$

(Complete the details.)

H12.33

We suggest proceeding by either of the following two methods:

(1) Algebraic method: $y - Ax \perp \operatorname{im} A \Leftrightarrow y - Ax \in \ker {}^tA$. (Why?)
(2) Analytic method: Applying the analytic version of the least-squares problem, we must minimize a nonnegative quadratic function $q(x_0, \dots, x_n) = \sum\limits_j \left(\sum\limits_i x_i f_i(t_j) - y_j\right)^2$. Do this by solving a system of linear equations $\partial q/\partial x_k = 0$: $k = 0,\dots, n$.

H12.34

Lemma 1 from section E6.3 shows that $\ker A = \ker ({}^tA \circ A)$.

H12.35

${}^tA \circ A$ is always (symmetric and) positive semidefinite (why?); thus, we have positive definiteness exactly when $\det {}^tA \circ A \neq 0$. In the least-squares problem for a T-system $\{t^n,\ldots,t,1\}$ the matrix $\begin{pmatrix} \sum t_j^{2n} & \cdots & \sum t_j^n \\ \vdots & & \vdots \\ \sum t_j^n & \cdots & m \end{pmatrix}$ arises as ${}^tA \circ A$, and the equalities that are being proved correlate with the condition of uniqueness "dim $Y \geq$ dim C" from section P12.32**.

Also, we may apply a purely algebraic claim "*A is nondegenerate* $\Leftrightarrow \det {}^tA \circ A > 0$" to a rectangular Vandermonde matrix $A = \begin{pmatrix} t_1^n & \cdots & 1 \\ \vdots & & \vdots \\ t_m^n & \cdots & 1 \end{pmatrix}$, with no reference to the least-squares problem. This method is also applicable to a more general matrix V, which degenerates for $m \leq n$ but does not degenerate if $m > n$ and there exists an $n + 1$-element sequence $j_0 <\ldots< j_n$ of indices j such that either all s_j with these indices are nonzero and the ratios t_j/s_j are distinct or all t_j with these indices are nonzero and the ratios s_j/t_j are distinct.

H12.36

A direct computation yields

$$x_0 = \frac{\sum u_j^2 \sum y_j - \sum u_j \sum u_j y_j}{m \sum u_j^2 - \left(\sum u_j\right)^2}, \qquad x_1 = \frac{m \sum u_j y_j - \sum u_j \sum y_j}{m \sum u_j^2 - \left(\sum u_j\right)^2}.$$

The fact that the line $y = x_0 + x_1 u$ passes through the indicated points may be established by direct verification using their coordinates. Also, it may be established as a special case of the claim from section P12.37**.

H12.37

This follows from the equations $\sum_i x_i \sum_j u_{kj} u_{ij} = \sum_j u_{kj} y_j \, (k = 0,\ldots,n)$ under an extra condition of equality of u_{0j} and the coefficients at x_0 to 1. (Complete the details.)

H12.38

Readers may observe from the foregoing explicit formulas in section P12.34** the homogeneity of x_i: substituting $u_{0j} \mapsto r_0 u_{0j}, \ldots, u_{nj} \mapsto r_n u_{nj}$, $\forall j$ implies $x_0 \mapsto {r_0}^{-1} x_0, \ldots, x_n \mapsto {r_n}^{-1} x_n$. We think, however, that proceeding by this method makes completing the proof difficult. Therefore, we suggest a different approach. Let us introduce, for $\alpha > 0$ and $b = (b_0, \ldots, b_n)$ with $b_i > 0$, a finite set $S_{\alpha,b} \subset \mathbb{R}^{n+1}_{u_0,\ldots,u_n}$ with the following properties: (1) the least-squares problem is solvable uniquely using $(u)_j \in S_{\alpha,b}$ and any y_j ($j = 1, \ldots, \#S_{\alpha,b}$) and (2) for any coordinate orthant O_k ($k = 1, \ldots, 2^{n+1}$), the number of points located in $S'_{\alpha,b,k} = \left(S_{\alpha,b} \setminus \prod_b\right) \cap O_k$ is not less than α: $\#S'_{\alpha,b,k} \geq \alpha \cdot \#S_{\alpha,b}$. Verify that grids $G^{n+1}(N,a)$ with $N_0, \ldots, N_n > 1$ are sets of this kind (say, for $b = a/2$ and $\alpha \approx 2^{-n-2}$). Produce the estimate $|x_i| \leq (\alpha^{-1/2} + 1) \cdot C \cdot {b_i}^{-1}$ for the solutions of the least-squares problem using $(u)_j \in S_{\alpha,b}$ and y_j with $|y_j| \leq C$.

H12.39

The answer is $m(d) = 1 + 2 + \ldots + (d+1) = (d+1)(d+2)/2$, that is, $m(d)$ is equal to the number of linearly independent monomials in (s,t) of degrees smaller than or equal to d. Indeed, for any $(s_j,t_j) \in \mathbb{R}^2$, $j = 1, \ldots, m$ with $m < (d+1)(d+2)/2$, there exists a nonzero polynomial $p_m(s,t) = \sum_{k+l \leq d} c_{kl} s^k t^l$ such that $p_m(s_j,t_j) = 0$, $\forall j$. (Why?) Thus, an affine space of solutions of the least-squares problem is at least one-dimensional for any given values y_j. On the other hand, show that a polynomial of degree d with a zero set containing a set S_d of the integral points of a triangle $\{s \geq 0,\ t \geq 0,\ s + t \leq d\} \subset \mathbb{R}^2_{s,t}$ (how many of those points are there?) does not exist, and, therefore, the least-squares problem is solvable uniquely, using the points $(s_j,t_j) \in S_d$.[18]

The least-squares problem is uniquely solved using a fixed set of points $\{(s_j,t_j) \in \mathbb{R}^2: j = 1, \ldots, m\}$ if and only if the rank of the matrix $A = \left(s_j^k t_j^l\right)_{\substack{k+l=0,\ldots,d, \\ j=1,\ldots,m}}$ equals $m(d)$.

---

[18] This means the nondegeneracy of the matrix $\left(s_j^k t_j^l\right)_{\substack{k+l=0,\ldots,d, \\ j = 1, \ldots, m(d)}}$ [$(s_j,t_j) \in S_d$]. But we must draw the reader's attention to the distinctive row scaling, which causes, for large degrees (practically, seven and larger), serious trouble for numerical computer processing. Commonly, the least-squares polynomial fitting using routine matrix computations becomes impractical very quickly with the growth of both the number of independent variables and the polynomials' degrees. One way to improve the situation consists in rescaling the rows by factors ρ^{-k-l} (however, this may bring the matrix close to a matrix of a smaller rank). Also, shifting the arguments to a neighborhood of the origin can be useful. A different way is to use the **Tikhonov-Phillips regularization**. Finer numerical procedures require a sophisticated analysis using methods of algebraic geometry.

Therefore, an algebraic algorithm for a determination of the solvability uniqueness, given $m \geq m(d)$ points (s_j,t_j), may consist of computing or estimating the aforementioned rank. Geometrically, equality of this rank to $m(d)$ means that the preceding set of (s_j,t_j) cannot belong to any algebraic curve of degree d (since coefficients of the curve would give a nontrivial solution of the linear system $Ax = 0$). Therefore, for $d = 2$ and six distinct points (s_j,t_j), a geometric algorithm for the determination of the solvability uniqueness may include the following steps:

- If four of the points belong to a straight line, then all six points belong to a quadric. (Why?)
- If four of the points do not belong to a straight line, then take a subset of five points. These five points belong to a unique quadric. (Why?) Derive an equation of this quadric $p_5(s,t) \equiv \sum_{k+l\leq 2} c_{kl}s^k t^l = 0$ (the coefficients are determined up to a common factor; why?).
- Finally, check whether the remaining sixth point satisfies this equation.

The reader may apply the following method of finding an equation of a quadric passing through five distinct points, with no four of them lying on a straight line. Take a subset of four points. The quadrics passing through four distinct points not lying on a straight line are presented as $p_4(s,t) \equiv c'p'_4(s,t) + c''p''_4(s,t) = 0$, where $p'_4(s,t) = 0, p''_4(s,t) = 0$ are distinct quadrics passing through these points (prove it!); the quadratic polynomials $p'_4(s,t)$ and $p''_4(s,t)$ may be found as products of linear factors. (List the cases respectively to locations of the points.) Determine the ratio c': c'' so that the quadric $p_4(s,t) = 0$ contains the remaining fifth point. [If three of those five points lie on a straight line $l_3(s,t) = 0$ and the other two lie on a straight line $l_2(s,t) = 0$, then the desired quadric is $l_2 \cdot l_3 = 0$.]

This method was introduced by a German geometer named Julius Plücker (1801–1868). Interested readers will find much information on this in Klein (1926) and Prasolov and Soloviev (1997).

H12.40

First obtain the mean values and covariance matrix of Y_j:

$$\bar{Y}_j - \bar{y}_j = \overline{Y_j - \bar{y}_j} = \int_{\mathbf{R}^m} (y_j - \bar{y}_j)p(y,c)\,dy_1 \ldots dy_m = 0,$$

$$\left\langle Y_j - \bar{y}_j,\ Y_k - \bar{y}_k \right\rangle = \int_{\mathbf{R}^m} (y_j - \bar{y}_j)(y_k - \bar{y}_k)p(y,c)\,dy_1 \ldots dy_m = \sigma^2 \cdot \delta_{jk}.$$

Next, the componentwise means of random vectors $\overline{v(y)} = \left(\overline{v_1(y)}, \ldots, \overline{v_r(y)}\right)$ $[v = (v_1,\ldots, v_r)]$ commute with linear transformations of those vectors: $T\,\bar{v} = \overline{Tv}$,

for a numerical matrix T [$Tv = (\sum T_{1l}v_l, \sum T_{2l}v_l, \ldots)$]. Therefore, taking into account that $({}^tA \circ A)^{-1}({}^tA)\bar{y} = ({}^tA \circ A)^{-1}({}^tA)Ac = c$, we will obtain (in accordance with section P12.34**)

$$\overline{\Phi(Y)} - c = \overline{\Phi(Y) - c} = \overline{({}^tA \circ A)^{-1}({}^tA)Y - ({}^tA \circ A)^{-1}({}^tA)\bar{y}}$$
$$= \overline{({}^tA \circ A)^{-1}({}^tA)(Y - \bar{y})} = ({}^tA \circ A)^{-1}({}^tA)\overline{(Y - \bar{y})} = ({}^tA \circ A)^{-1}({}^tA)0 = 0,$$

which proves the unbiasedness of the maximal likelihood estimate for c. Now consider the Gram matrix $G = \left(\left\langle e^{(\mu)}, e^{(\nu)}\right\rangle\right)_{\mu,\nu=1,\ldots}$ for vectors $e^{(1)}, \ldots$. Verify that the Gram matrix for vectors $v^{(1)} = \sum T_{1\nu}e^{(\nu)}, v^{(2)} = \sum T_{2\nu}e^{(\nu)}, \ldots$ [with a numerical matrix $T = (T_{\mu\nu})$] is $T \circ G \circ {}^tT$. Therefore, and taking into account the symmetry of the matrix $({}^tA \circ A)^{-1}$ (why is it symmetric?), the covariance matrix for $\Phi_0(Y), \ldots, \Phi_n(Y)$ is

$$(\langle \Phi_i(Y) - c_i,\ \Phi_k(Y) - c_k\rangle)_{i,k=0,\ldots,n} = ({}^tA \circ A)^{-1}({}^tA)\left(\left\langle Y_j - \bar{y}_j,\ Y_k - \bar{y}_k\right\rangle\right)_{j,k=1,\ldots,m}$$
$$A({}^tA \circ A)^{-1} = ({}^tA \circ A)^{-1}({}^tA)(\sigma^2 E_m)A({}^tA \circ A)^{-1} = \sigma^2({}^tA \circ A)^{-1}$$

(where E_m is the identity $m \times m$ matrix), which completes the proof.

H12.41

Exponential functions form a T-system (according to section P12.10*); therefore, the assumptions in Problem P12.40*** are realized (ker $A = 0$). Application of section P12.40*** yields the variance of the estimate $\Phi(Y) = \Phi_0(Y)$ when $n = 0$, and it yields the variances and correlation coefficient[19] of $\Phi_0(Y), \Phi_1(Y)$ when $n = 1$:

$$n = 0: \qquad \sigma_\Phi^2 = \sigma^2 \Big/ \sum e^{-2t_j/T},$$

$$n = 1: \qquad \sigma_{\Phi_k}^2 = \sigma^2 \frac{\sum e^{-2t_j/T_{1-k}}}{\left(\sum e^{-2t_j/T_0}\right)\left(\sum e^{-2t_j/T_1}\right) - \left(\sum e^{-t_j\left(T_0^{-1}+T_1^{-1}\right)}\right)^2},$$

$k = 0, 1,$

$$\rho_{\Phi_0\Phi_1} = \frac{\sum e^{-t_j\left(T_0^{-1}+T_1^{-1}\right)}}{\sqrt{\left(\sum e^{-2t_j/T_0}\right)\left(\sum e^{-2t_j/T_1}\right)}}.$$

[19] Correlation coefficients ρ_{ij} form a dimensionless covariance matrix: $\rho_{ij} := \mathrm{cov}_{ij} \Big/ \sqrt{\mathrm{cov}_{ii} \cdot \mathrm{cov}_{jj}}$.

Readers can straightforwardly get from it asymptotic expressions $(\sigma_{\Phi_k}/c_k)^2 = O(\varepsilon^{-2})$, $\rho_{\Phi_0\Phi_1} = 1 + O(\varepsilon^2)$ because $\varepsilon = |T_1 - T_0|/T_0 \to 0$. The first expression confirms that the indistinctness of the exponential terms makes their amplitudes practically inestimable (estimable within the infinite error), and the second just states that the estimates become entirely correlated (dependent).

Explanation

E12.1

The line segment between $t'_0 < \ldots < t'_n$ and $t''_0 < \ldots < t''_n$ is defined as the trajectory of a uniform movement $\tau \mapsto \overrightarrow{t}(\tau) = \overrightarrow{t'} + \tau(\overrightarrow{t''} - \overrightarrow{t'}) : \quad \tau \in [0, 1]$ $\left(\overrightarrow{t} = (t_0, \ldots, t_n)\right)$. It lies completely in the set $\{t_0 < \ldots < t_n\}$; in other words, $t_{i+1}(\tau) - t_i(\tau)$ are positive at any points of the segment [0,1] because they vary linearly and are positive at its ends (create a figure). In turn, a continuous positive function on an arc has a positive minimum there because the arc is a compact set.

E12.4

We will prove that im A has zero intersections with the coordinate subspaces of l nonzero coordinates c_i if we establish that for any $c_{i_0}, \ldots, c_{i_m}$ there is an element of im A with these coordinates of indices $i_0, \ldots, i_m$, respectively. (Why?) We leave it to the reader to verify that the positiveness of all t_j implies the existence of b with any prescribed coordinates of Ab of indices $i_0, \ldots, i_m$ (which is equivalent to the linear independence of any $m + 1$ rows of matrix A). **QED**.

Now, let $V_{n+1}(\overrightarrow{t}, \overrightarrow{k}) \neq 0$ for any n and $V_{n+1}(\overrightarrow{t}, \overrightarrow{k}) > 0$ for $n < n'$, and we must show that $V_{n'+1}(\overrightarrow{t}, \overrightarrow{k}) > 0$ ($0 < t_0 < \ldots < t_{n'}$, $k_0 < \ldots < k_{n'}$). Obviously, this holds for $n' = 0$. For $n' > 0$, directing $t_{n'}$ to infinity while keeping the remaining points fixed implies $V_{n'+1} \to \infty$. (Decompose the determinant $V_{n'+1}$ with respect to its last row and apply the inductive hypothesis.) Therefore, we have found cases of $V_{n'+1} > 0$. By virtue of continuity and since always $V_{n'+1} \neq 0$, this yields that always $V_{n'+1} > 0$. **QED**. (We leave it to the reader to fill in the details.)

E12.5

We have

$$\det\left(p_i(t_j)\right)_{\substack{i=i_0,\ldots,i_m\\ j=0,\ldots,m}} = \det\left(\sum_{l=k_{i-1}+1}^{k_i} a_l f_l(t_j)\right)_{\substack{i=i_0,\ldots,i_m\\ j=0,\ldots,m}}$$
$$= \sum_{k_{i_{r-1}}<l_{i_r}\le k_{i_r},\, r=0,\ldots,m} a_{l_{i_0}}\ldots a_{l_{i_m}} \det\left(f_{l_i}(t_j)\right)_{\substack{i=i_0,\ldots,i_m\\ j=0,\ldots,m}}.$$

E12.11

Proof of Statement 1. Decomposing $W_{m+1}(g_0,\ldots,g_m) = \begin{pmatrix} 1 & g_1 & \cdots & g_m \\ 0 & g'_1 & \cdots & g'_m \\ \vdots & \vdots & & \vdots \\ 0 & g_1^{[m]} & \cdots & g_m^{[m]} \end{pmatrix}$ with respect to its first column shows that $W_m(g'_1,\ldots,g'_m) \neq 0$ for $1 \le m \le n$. By the inductive hypothesis, $g'_1,\ldots,g'_n$ form an EM-system, so polynomials on $g'_1,\ldots,g'_m$ have at most $m-1$ zeros in I, considering multiplicity ($1 \le m \le n$). Derivatives of polynomials on $g_0,\ldots,g_m$ are polynomials on $g'_1,\ldots,g'_m$. Now complete the proof using the Rolle theorem.

$W_{n+1}(f_0,\ldots,f_n) \neq 0$ does not imply the Chebyshev property of $f_0,\ldots,f_n$, as, for example, $f_0 = \cos t, f_1 = \sin t$ cannot form T-system on any segments of lengths $\ge\pi$ (or intervals, or half-intervals, of lengths $>\pi$). Readers should realize why the preceding proof does not cover such cases.

Proof of Statement 2. First establish the following lemma.

Lemma *For m times differentiable functions $f_0,\ldots,f_m$ and sequences of points $t_0 \le \ldots \le t_{j-1} < t_j < \ldots < t_{j+r-1} < t_{j+r} \le \ldots \le t_m$ such that $t_{j+1} - t_j = \ldots = t_{j+r-1} - t_{j+r-2}$,*

$$\frac{W_{m+1}(f_i)(t_j)_{i,j=0,\ldots,m}}{N(r)h^{\binom{r}{2}}} = W_{m+1}(f_0,\ldots,f_m)(t_0,\ldots,t_{j-1},\underbrace{t_j,\ldots,t_j}_{r\text{ times}},t_{j+r},\ldots,t_m) + o(1)$$

as $h \to 0$, where $h = t_{j+1} - t_j = \ldots = t_{j+r-1} - t_{j+r-2}$ and $N(r)$ is a positive constant that depends on r.

This lemma is proved by expanding the elements of the $(j+1)$th,$\ldots$,$(j+r-1)$th columns in Taylor series at t_j up to the terms of order of h^{r-1} and then using the multilinearity and skew symmetry of determinants with respect to columns. (Work

out the details.) Applying this lemma and taking into account that $W_{m+1}(f_{i_0},\ldots,f_{i_m})(t_0,\ldots,t_m) \neq 0$ and $W_{m+1}(f_{i_0},\ldots,f_{i_m})(t) \neq 0$, deduce that these determinants have the same sign for $t_0 = t,\ldots, t_m = t + mh$, for small enough $h > 0$. Extend this property of having the same sign to any $t_0 <\ldots< t_m$ by using continuity, and then to any $t_0 \leq\ldots\leq t_m$ using the preceding lemma ("in the backward direction") and the nonvanishing of the determinants.

E12.14

Case of W > 1. Considering multiplicity, q has at most W zeros. Therefore, this number is finite, so the zeros are **isolated**. In the interval $U_j = (t_j - \delta, t_j + \delta)$ we have $q = const \cdot (t - t_j)^{k_j} + o\left((t - t_j)^{k_j}\right)$, where k_j are multiplicities of zeros t_j. These representations, with $const \neq 0$, are achievable since the multiplicities do not exceed the order of smoothness (the number of derivatives) that is not smaller than W. (Actually, $\sum k_j \leq W$.) Define a closed segment $J \subset I$ so that $|q|$ is large enough on $\mathcal{N}J$ (which is achievable due to the limit behavior of q, but it could be unachievable if, for example, q had a horizontal asymptote!). On the compact set $J\setminus(\bigcup U_j)$ we have $|q| \geq \sigma > 0$. Therefore, $q_h := q + h$ have no zeros outside $\bigcup U_j$ for $|h| < \sigma$. On the other hand, direct computations show that for small $\varepsilon,\delta > 0$, q_h with $h \neq 0, |h| < \varepsilon$ have one simple zero in U_j for odd k_j and no zeros or two simple zeros for even k_j. Also, we suggest unifying those computations in the following lemma.

Lemma *Let, for an integer $k \geq 1$, a function $f(x) = x^k + o(x^k)$ be k times differentiable in a neighborhood of $x = 0$ (for $k = 1$, continuity of the derivative is also assumed[20]). Then an equation $y = f(x)$ with a small y has:*

- *A unique small positive root for $y > 0$*
- *A unique small negative root for $y > 0$ and even k*
- *No small roots for $y < 0$ and even k*
- *A unique small negative root for $y < 0$ and odd k*

All of these roots are simple.

(Proving this lemma is a fairly good exercise in analysis. However, readers who feel they lack experience may have a look at the proof of this lemma at the end of section E12.14.)

[20] Readers may give examples of differentiable functions $f = x + o(x)$ with derivatives discontinuous at $x = 0$, when the equation $y = f(x)$ has a multiple solution for arbitrarily small y. The applicability of this lemma to our purposes is provided by the following condition: *the number of derivatives is $\geq W > 1$*; therefore, the first derivative is continuous.

To sum up, for a small $\varepsilon > 0$ the members of a family $\{q_h\colon |h| < \varepsilon\}$ have the same, modulo 2, number of zeros (considering multiplicity).

The same holds after substituting a source function q by $q_y = q + y$ ($y \in \mathbb{R}$): for a small $\varepsilon_y > 0$ the members of a family $\{q_{y+h}\colon |h| < \varepsilon_y\}$ have the same, modulo 2, number of zeros (considering multiplicity). (Why?)

Finally, take a covering of a segment $-1 \le y \le 0$ by intervals $(y - \varepsilon_y, y + \varepsilon_y)$. There exists, because of compactness, a finite subcovering, so q and $q - 1$ have the same, modulo 2, number of zeros (considering multiplicity), which completes the proof.

Advanced readers may prefer the kind of argument that is similar to that in section H12.1. A number of zeros, modulo 2, of q_y cannot change locally, so that a subset $\{y\} \subseteq \mathbb{R}$ on which it is constant is open and closed. Therefore, it is all of $\mathbb{R}$ or empty.

Proof of lemma. For small x, f has the same sign as x^k, so we can obtain the statements of the lemma concerning the existence (or nonexistence) of the roots applying the continuity of f. (How?) We will prove the uniqueness and the simplicity of small positive roots for $y > 0$; in the remaining cases, they may be established using similar arguments. If, on the contrary, there are distinct roots (a multiple root), the derivative must vanish at a point that is intermediate between those roots (resp. at that point) by the Rolle theorem (or by the Lagrange intermediate value theorem, which is equivalent to it-depending on the reader's preference). On the other hand, $f' > 0$ for small $x > 0$ since $f = x^k + v(x)$ with $v = o(x^k)$, and so $f' = kx^{k-1} + v'$ with $v' = o(x^{k-1})$. (Indeed, for $k > 1$ we will find, using the Taylor-series expansion, that

$$v = o(x^k) \Leftrightarrow \left\{v(0) = \ldots = v^{[k]}(0) = 0\right\} \Rightarrow \left\{v'(0) = \ldots = (v')^{[k-1]}(0) = 0\right\}$$
$$\Leftrightarrow v' = o(x^{k-1}).$$

For $k = 1$ we have $v'(0) = \lim_{x \to 0} [v(x)/x] = 0$, so $v' = o(1)$ because its continuity is assumed.) This completes the proof.

E12.17

Proof of Bézout theorem. We suggest using the **resultant of the polynomials**, as in Walker (1950) and Arnol'd (1989 Historical). The resultant of the two

polynomials $f(x) = \sum_{i=0}^{m} v_{m-i} x^i$, $g(x) = \sum_{j=0}^{n} w_{n-j} x^j$ is the determinant of the matrix R

$$= \begin{pmatrix} v_0 & \cdots & \cdots & v_m & & \\ & \ddots & & & \ddots & \\ & & v_0 & \cdots & \cdots & v_m \\ w_0 & \cdots & \cdots & w_n & & \\ & \ddots & & & \ddots & \\ & & w_0 & \cdots & \cdots & w_n \end{pmatrix} : \operatorname{res}(f,g) = \det R.$$ Verify that the coefficients of arbitrary polynomials $f_1(x) = \sum_{i=0}^{m-1} b_{m-1-i}x^i$ and $g_1(x) = \sum_{j=0}^{n-1} a_{n-1-j}x^j$ of degrees $n-1$ and $m-1$, respectively, satisfying the identity $fg_1 + f_1g = 0$ form a solution of a linear system, which in matrix form looks like this: $(a_0, \ldots, a_{n-1}, b_0, \ldots, b_{m-1})R = (\underbrace{0, \ldots, 0}_{m+n})$. From the preceding statements deduce that $\operatorname{res}(f,g) = 0$ if and only if $f(x)$ and $g(x)$ have a common root (in an algebraic closure of a field of the coefficients).[21]

Now consider the equations $f(x,y) = 0$, $g(x,y) = 0$ from section H12.17 as polynomial equations on x, with polynomial (in y) coefficients. Verify that $\operatorname{res}(f,g)$ as a polynomial in y has a degree of, at most, mn. If, for a fixed y, those equations have a common root x, then the resultant vanishes. Therefore, the common points of our algebraic curves project to, at most, mn points of an ordinate axis. The same will hold using independent variables $x' = ax + by$, $y' = cx + dy$ ($ad - bc \neq 0$) instead of x, y, so the common points of our algebraic curves project to, at most, mn points of any straight line in a plane. From the preceding discussion deduce that there are, at most, mn of those points. **QED**.

The proved version of the Bézout theorem yields upper estimates, but not exact values. For example, two distinct circles in $\mathbb{R}^2$ have at most two, considering multiplicity, common points. (Actually, this also may be established using the Bézout theorem. Show that common points of two distinct similar quadrics in $\mathbb{R}^2$ lie on a straight line, so there are at most $2 \cdot 1 = 2$ such points.) The complete classic version of the Bézout theorem states that there are exactly mn common points, considering multiplicity, imaginary points, and points at infinity (in short, the points in projective space $\mathbb{C}P^2$), or common components (Walker 1950). The theorems of Kushnirenko and Bernshtein (Bernshtein 1975 and references therein) generalize this result for determining the number of common points of algebraic varieties in the terms of mixed volumes of **Newtonian polyhedrons**. Also, readers will find further development of the subject and multiple references in Khovanskii (1997).

Analytic method in section H12.17. For $z = e^{i\theta}$ we have $\cos v\theta = \frac{z^v + z^{-v}}{2}$, $\sin v\theta = \frac{z^v - z^{-v}}{2i}$. Inserting this, for $v = 0, \ldots, n$, into $G(\theta) = \sum_{k=0}^{n} a_k \cos k\theta + b_k \sin k\theta$ and multiplying by $e^{in\theta}$ yields

[21] Readers may prove that actually $\operatorname{res}(f,g) = v_0^n w_0^m \prod_{i,j} (t_i - u_j)$, where t_i, u_j are roots of $f(x)$, $g(x)$, respectively (in an algebraic closure of a field of the coefficients) (Van der Waerden 1971, 1967; Lang 1965).

$$p(z) = a_0 z^n + \sum_{v=1}^{n} \left[\frac{1}{2} a_v (z^{n+v} + z^{n-v}) + \frac{1}{2i} b_v (z^{n+v} - z^{n-v}) \right],$$

so we have $u_v = \bar{u}_{2n-v} = \frac{a_{n-v} + i b_{n-v}}{2}$ for $v = 0, \ldots, n-1$ and $u_n = a_0$. Conversely, for a reciprocal polynomial $p(z) = \sum_{i=0}^{2n} u_{2n-i} z^i$ we have $u_v = \bar{u}_{2n-v}$ and, specifically, $u_n \in \mathbb{R}$, so that

$$G(\theta) - u_n = e^{-in\theta} \sum_{v=0}^{n-1} \left(u_v e^{iv\theta} + u_{2n-v} e^{i(2n-v)\theta} \right) = 2\mathrm{Re} \left(\sum_{v=0}^{n-1} u_v e^{i(n-v)\theta} \right)$$

represents a trigonometric polynomial of order n with real coefficients.

Obviously, the products of reciprocal polynomials are themselves reciprocal (also, the quotients are reciprocal polynomials, if they are polynomials at all). Derive from this that a polynomial is reciprocal if and only if its end coefficients are complex conjugate and all roots that do not belong to the unit circle can be grouped into pairs of inverse number (like $\rho e^{i\theta}$ and $\rho^{-1} e^{i\theta}$).

E12.18

If the $m \times m$ minors of this matrix vanished all together, we would have $\det(f_i(t_j))_{i,j = 0,\ldots,n} = 0$ with any t_0 and $t_{m+1}, \ldots, t_n$, as follows from the decomposition of this determinant with respect to its first,..., mth columns.

In the geometric definition of polynomials with zeros $t_1, \ldots, t_m$, the $n - m + 1$-dimensional space of the coefficients is a subspace of $\mathbb{R}^{n+1}$ orthogonal to the desired m-dimensional plane.

E12.20

Proof of theorem. Necessity. Let max $|p| > \mu$ between any adjacent zeros, and in the case where I has end(s), also between the end(s) and the closest zero(s). Define q as a Lagrange interpolating polynomial equal to zero at the nodal zeros of p, positive at its nonnodal zeros, in neighborhoods of which $p \geq 0$, negative at nonnodal zeros, in neighborhoods of which $p \leq 0$, and normalized so that max $\left| q \right| < \mu$ on I. (Work out the details.)

Sufficiency. Case of $n = 2k$. Since the set of polynomials on F defined by condition 2 is compact (indeed, it is the unit hypersphere in finite-dimensional space), there is a limit polynomial $p(t)$ that has nonnodal zeros $t_1, \ldots, t_k$ (why?) and no other zeros

(by virtue of the "necessity part" of the theorem being proved). In fact, $p(t) \geq 0$ on I; why? [Investigate signs of $p^{(v)}(t)$ in the intervals between zeros.]

Sufficiency. "Hardest case": $2k + l = n - 1$, $I = [a,b]$, $t_{k+1} = a$. In a Euclidean space $\mathbf{R}^{n+1}_{c_0,\ldots,c_n}$ the squared distance from the origin to the plane defined by independent equations $\langle c, v_j \rangle = d_j$ ($j = 1,\ldots, m$) equals

$$h^2 = (d_1, \ldots, d_m)\, G^{-1\,t}(d_1, \ldots, d_m) = \sum g^{ij} d_i d_j,$$

where $G = (g_{ij})_{i,j\,=\,1,\ldots,\,m}$ is the Gram matrix of vectors v_j, $g_{ij} = \langle v_i, v_j \rangle$. In our case, $m = n$, $d_1 = \ldots = d_{n-1} = 0$, and $d_n = 1$, so that $h^2 = g^{nn} = \frac{\text{cofactor of } g_{nn} \text{ in } G}{\det G}$. Fix a point $t_0 < b$ so that $t_1,\ldots, t_k, t_{k+2},\ldots, t_{k+l} \in [a, t_0]$ and $t_1^{(v)}, \ldots, t_{k+1}^{(v)} \in [a, t_0], \forall v$. Angles between the vectors $v = (f_0(t),\ldots, f_n(t))$, with $t \in [a, t_0]$ and $v_n = (f_0(b),\ldots, f_n(b))$, have a minimum $\alpha \in (0,\pi)$. (Why?) Hence, for the n-dimensional volume of an n-dimensional parallelepiped spanned on $v_1,\ldots, v_n$ and the $n - 1$-dimensional volume of an $n - 1$-dimensional parallelepiped spanned on $v_1,\ldots, v_{n-1}$ we have

$$\text{vol}_n(v_1, \ldots, v_n) \geq \text{vol}_{n-1}(v_1, \ldots, v_{n-1}) \cdot |v_n| \cdot \sin\, \alpha.$$

(Why?) Since $\det G = [\text{vol}_n(v_1,\ldots, v_n)]^2$ and $[\text{cofactor of } g_{nn} \text{ in } G] = [\text{vol}_{n-1}(v_1,\ldots, v_{n-1})]^2$, this yields an estimate $h \leq (|v_n| \cdot \sin \alpha)^{-1}$. (Work out the details.)

E12.24

The theorem from section P12.20*** is applicable here because the number of elements of the T-system, reduced by 1, and the number of a polynomial's nodal zeros are both even. Therefore, their difference is even.

E12.25

The desired trigonometric trinomial with zeros $0 \leq \theta_1 < \theta_2 < 2\pi$ may be found as $p(\theta) - \varepsilon$, for proper ε and trinomial $p(\theta) = 1 + \sin(\theta + \theta_0)$, with θ_0 defined so as to provide a pair of zeros in $[0, 2\pi)$ located symmetrically with respect to the point $(\theta_1 + \theta_2)/2$. (Depict and fill in the details.)

E12.26

Example of invariant distance function that cannot derive from a norm. Consider on the space X of real (or complex) sequences $x = (x_0, x_1, \ldots)$ a distance function

(verify the axioms!) $\operatorname{dist}(x, y) := \sum \frac{1}{2^n} \cdot \frac{|x_n - y_n|}{1+|x_n - y_n|}$. Obviously, it is translation-invariant, but it cannot derive from a norm since X has a finite diameter.

The topology defined by this distance function also cannot derive from a norm. To establish this, prove that a sequence $x^{(v)}$ converges to x in this topology if and only if $x_n^{(v)}$ converge to x_n for all n, then assume normability of the topology and consider the sequence $x^{(v)}/\|x^{(v)}\|$ with $x_n^{(v)} = \delta_{vn}$.

A translation-noninvariant distance function. Consider the distance function on a vector space defined by a cartographical map on a sphere (of the same dimension) with a deleted North Pole on this space. (Actually, this distance function is invariant with respect to a different group.)

E12.27

Necessity. Taking

$$u_j = \max_{t_{j-1} \le t < t_j} \{t : f(t) - p(t) = f(t_{j-1}) - p(t_{j-1})\},$$
$$v_j = \min_{t_{j-1} \le t_j} \{t : f(t) - p(t) = f(t_j) - p(t_j)\}$$

we have $u_j < v_j$, $\forall j = 1,\ldots, m-1$. (Why?) The required points $t'_1, \ldots, t'_{m-1}$ may be defined as zeros of $f(t) - p(t)$ in intervals (u_j, v_j), respectively. (Work out the details.)

E12.28

Sufficiency. If p, q are polynomials of best approximation, the same will hold for the convex linear combinations

$$|f(t) - \alpha p(t) - (1-\alpha) q(t)| \le \alpha |f(t) - p(t)| + (1-\alpha)|f(t) - q(t)| = d.$$

By the Chebyshev theorem in section P12.27***, an equation $|f(t) - r(t)| = d$ has at least $\#F$ (in fact, $\#F + 1$) distinct roots in I for a polynomial of best approximation r; therefore, an equation $f - \alpha p - (1-\alpha) q = \pm d$ with any fixed $\alpha \in (0,1)$ has at least $\#F$ distinct roots in I. However, at the roots we will have $f - p = \pm d$ and $f - q = \pm d$ – with the same signs "+" or "−" for both (why?); hence, $p = q$ at those points, so $p = q$ identically (by section P12.19*; complete the details).

Necessity. As follows from the properties of $p(t)$, $\max_{a \le t \le b} |f(t)| = 1$ and $f(t_j) = \operatorname{sign} y_j$, $\forall j = 0,\ldots, n$. From this, for any polynomial $q(t)$ on F, $\max_{a \le t \le b} |f(t) - q(t)| \ge 1$ [as otherwise $q(t_j)$ would have the same signs as y_j, which would contradict

the orthogonality of $(y_0,\ldots,y_n)$ to the polynomials on F]. At the same time, we have, for $|C| \le 1$,

$$|f(t) - Cp(t)| \le |f(t)| + |Cp(t)| = |g(t)| \cdot |1 - |p(t)|| + |Cp(t)| \\ \le 1 - |p(t)| + |p(t)| = 1,$$

which completes the proof.

E12.30

Application of the Chebyshev polynomials to find the eigenvalues of 3-diagonal n × n matrices $M = \begin{pmatrix} 0 & 2 & & \\ 1 & 0 & 1 & \\ & \ddots & \ddots & \ddots \end{pmatrix}$ and $N = \begin{pmatrix} 0 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & & \end{pmatrix}$. Let $v = {}^t(v_0,\ldots,v_{n-1})$ of $v_0 = 1$ be an eigenvector of M corresponding to an eigenvalue λ. By-component processing of an equation $Mv = \lambda v$ yields the equalities $v_i = T_i(\lambda/2)$, $i = 0,\ldots, n-1$, and $T_{n-2}(\lambda/2) = \lambda \cdot T_{n-1}(\lambda/2)$. (Why?) Hence, the characteristic polynomial of M is $\chi_M(\lambda) = 2\lambda \cdot T_{n-1}(\lambda/2) - 2T_{n-2}(\lambda/2) = 2T_n(\lambda/2)$, so the eigenvalues are $\lambda_k(M) = 2\cos\frac{(2k+1)\pi}{2n}$, $k = 0,\ldots, n-1$. (Work out the details.) As for N, a straightforward computation making use of the identity $\sin(n+1)\theta + \sin(n-1)\theta = 2\cos\theta \sin n\theta$ shows that the sequence of polynomials $U_n(t) := \frac{T'_{n+1}(t)}{n+1} = \frac{\sin((n+1)\arccos t)}{\sin(\arccos t)}$ satisfies an equation $\begin{pmatrix} 0 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & & \end{pmatrix} \begin{pmatrix} U_0 \\ U_1 \\ \vdots \end{pmatrix} = 2t \cdot \begin{pmatrix} U_0 \\ U_1 \\ \vdots \end{pmatrix}$. Therefore, a calculation similar to the previous one yields $\chi_N(\lambda) = U_n(\lambda/2)$, so the eigenvalues are $\lambda_k(N) = 2\cos\frac{k\pi}{n+1}$, $k = 1,\ldots, n$. (Fill in the details.)

A general theorem states that the corner minors A_n of a 3-diagonal matrix $A = \begin{pmatrix} a_0 & b_0 & & \\ b_0 & a_1 & b_1 & \\ & b_1 & \ddots & \ddots \\ & & \ddots & \end{pmatrix}$ with $b_i > 0$, $\forall i$, have all their eigenvalues real and simple (nonmultiple), and the eigenvalues of A_n and A_{n+1} alternate (as in **Sturm theory**) (Gantmacher and Krein 1950; Marcus and Minc 1964; and references therein). Readers may establish that these eigenvalues are real and simple (nonmultiple) with arguments similar to those used previously in section H12.21 for proving Hurwitz's theorem; also, the simplicity of the eigenvalues may be shown by utilizing (prove them first!) the inequalities $\det\begin{pmatrix} P_{n-1}(\lambda) & P'_{n-1}(\lambda) \\ P_n(\lambda) & P'_n(\lambda) \end{pmatrix} \ne 0$ for a corresponding to A sequence of the orthogonal polynomials P_n; the same inequalities imply the alternation property as well. [In an analytic language, the reason for the alternation is that P_{n-1} takes different signs at successive zeros of P_n; topologically, the alternation is due to the "displacement" of the planar vector $v_{n-1} = (P_{n-1}, P'_{n-1})$ by the vector $v_n = (P_n, P'_n)$ while v_n rotates, which places v_{n-1} on the ordinate axis, between two successive occurrences of v_n (what follows from that?). The reader may work out the details proceeding by selecting a preferred

method. The analytic one is more traditional; a far-reaching development of the topological one is discussed in Arnol'd (1985).]

Sequences of $x_n = P_n(\lambda) = T_n(\lambda/2)$ and $x_n = Q_n(\lambda) = U_n(\lambda/2)$ form a basis of solutions of an infinite linear system $\{x_{k-1} - \lambda x_k + x_{k+1} = 0\colon k = 1,2,\ldots\}$ (verify it!); in addition, zeros of P_n, Q_n alternate. A similar statement holds for linearly independent sequences of orthogonal polynomials corresponding to each 3-diagonal matrix A, discussed previously; that is, for sequences P_n, Q_n forming a basis of solutions of $\{b_{k-1}x_{k-1} + (a_k - \lambda)\cdot x_k + b_k x_{k+1} = 0\colon k = 1,2,\ldots\}$, the zeros' alternation may be shown by the inequalities $\det\begin{pmatrix} P_{n-1}(\lambda) & Q_{n-1}(\lambda) \\ P_n(\lambda) & Q_n(\lambda) \end{pmatrix} \neq 0$; fill in the details. Further generalizations and applications of these theorems in vibration theory for elastic systems are discussed in Gantmacher and Krein (1950).

E12.33

(1) Algebraic method: apply Lemma 1 from section E6.3 above and work out the details.

(2) Analytic method: the equations actually have the form $\sum_i x_i \sum_j f_k(t_j) f_i(t_j) = \sum_j f_k(t_j) y_j$, or, in matrix form,

$$\begin{pmatrix} f_0(t_1) & f_0(t_2) & \cdots \\ f_1(t_1) & & \\ \vdots & & \end{pmatrix} \circ \begin{pmatrix} f_0(t_1) & f_1(t_1) & \cdots \\ f_0(t_2) & & \\ \vdots & & \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \end{pmatrix} = \begin{pmatrix} f_0(t_1) & f_0(t_2) & \cdots \\ f_1(t_1) & & \\ \vdots & & \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix},$$

which completes the proof.

Readers experienced in multivariate derivatives might prefer organizing the computations as follows. The system may be presented in a concise form: $dq(x) = 0$. The differential on the left-hand side, which is a linear function of increment h, may be obtained in various equivalent ways, as, for example,

1. $dq(x)(h) = \frac{d}{dt}\big|_{t=0} q(x+th) = \frac{d}{dt}\big|_{t=0}\langle A(x+th) - y, A(x+th) - y\rangle = 2\langle Ax - y, Ah\rangle = 2\langle {}^tA(Ax-y), h\rangle$;
2. $dq(x)(h)$ is a linear in h component of $q(x+h) - q(x) = 2\langle Ax - y, Ah\rangle + \langle Ah, Ah\rangle$, so, $dq(x)(h) = 2\langle Ax - y, Ah\rangle = 2\langle {}^tA(Ax-y), h\rangle$;
3. The multivariate Leibnitz derivation rule brings $dq(x)(h) = 2\langle Ax - y, d(Ax-y)(h)\rangle = 2\langle Ax - y, Ah\rangle = 2\langle {}^tA(Ax-y), h\rangle$.

(Readers may suggest even more ways.) Thus, $dq(x) = 0 \Leftrightarrow {}^tA\,(Ax - y) = 0$, which completes the proof.

E12.35

We have, using statements in section E6.3, that ${}^t({}^tA \circ A) = {}^tA \circ {}^t({}^tA) = {}^tA \circ A$, which proves symmetry, and $\langle ({}^tA \circ A)x, x\rangle = \langle Ax, Ax\rangle \geq 0$, which proves positive semi-definiteness.

E12.38

If $|x_{i_0}| \leq Cb_{i_0}^{-1}$, then $|x_{i_0}| < (\alpha^{-1/2} + 1) \cdot C \cdot b_{i_0}^{-1}$. Fix i_0 such that $|x_{i_0}| b_{i_0} > C$. Take an orthant $O_k \subset \mathbb{R}^+_{,\ldots,}$, in which all $x_i u_i$ are of the same sign. In ${S'}_{\alpha,b,k}$ we will have for all j

$$\left|\sum_i x_i u_{ij}\right| = \sum_i |x_i u_{ij}| \geq \sum |x_i|\ b_i \geq |x_{i_0}| b_{i_0} \geq |y_j|,$$

and hence,

$$\left|\sum_i x_i u_{ij} - y_j\right| \geq \left|\left|\sum_i x_i u_{ij}\right| - |y_j|\right| = \left|\sum_i x_i u_{ij}\right| - |y_j| \geq |x_{i_0}| b_{i_0} - |y_j| \geq |x_{i_0}| b_{i_0} - C (\geq 0).$$

Consequently,

$$\sum_{S_{\alpha,b}} \left(\sum_i x_i u_{ij} - y_j\right)^2 \geq \sum_{S'_{\alpha,b,k}} \left(\sum_i x_i u_{ij} - y_j\right)^2 \geq \alpha \cdot \#S_{\alpha,b} \cdot (|x_{i_0}| b_{i_0} - C)^2.$$

On the other hand, the sum of squares in the left-hand side cannot exceed the similar sum that may be found for a hyperplane $y = 0$ (with zero coefficients x_i) (why?), and so,

$$C^2 \cdot \#S_{\alpha,b} \geq \alpha \cdot \#S_{\alpha,b} \cdot (|x_{i_0}| b_{i_0} - C)^2,$$

which is equivalent to $|x_{i_0}| < (\alpha^{-1/2} + 1) \cdot C \cdot b_{i_0}^{-1}$, **QED**.

E12.39

Use induction on d to show that S_d (containing $m(d)$ elements) cannot be contained in a zero set of a nonzero polynomial of degree d. For $d = 0$, the zero set is empty,

whereas $\#S_0 = 1$. Let $d > 0$ and assume that this claim has been proved already for smaller, than d, degrees. A straight line $s + t = d$ contains $d + 1$ elements of S_d and by the Bézout theorem (section H12.17) this line will be contained in the zero set of some polynomial $p_{m(d)}(s,t) = \sum_{k+l \le d} c_{kl} s^k t^l$ of degree d if S_d is also contained there. (Complete the details.) Consequently, $p_{m(d)}(s,t)$ will be divided by a linear polynomial $s + t - d$ (why?), and a set $S_d \backslash \{s + t = d\} = S_{d-1}$ will be contained in a zero set of the quotient, in contradiction with the hypothesis of induction, **QED**.

If four points $A_j = (s_j,t_j)$ belong to a straight line $l_4(s,t) = 0$, and $l_2(s,t) = 0$ is a straight line passing through A_5, A_6, then a quadric $l_2 \cdot l_4 = 0$ passes through all six. Next, if quadrics $q'_5(s,t) = 0$, $q''_5(s,t) = 0$ pass through the same five points, they have, by the Bézout theorem, a common component; that is, $q'_5(s,t)$ and $q''_5(s,t)$ have a common linear factor or are proportional (and the quadrics coincide; complete the details[22]). Let them have a common linear factor: $q'_5(s,t) = l(s,t) \cdot l'(s,t)$, $q''_5(s,t) = l(s,t) \cdot l''(s,t)$. If the straight line $l(s,t) = 0$ contains, at most, three of five common points, then at least two remaining belong to both $l'(s,t) = 0$ and $l''(s,t) = 0$, hence, these lines coincide, so, the quadrics coincide, **QED**. (We leave it to the reader to fill in the details.)

The quadrics passing through four distinct points (s_j,t_j), $j = 1,\ldots,4$, not lying on a straight line, are linear combinations of two distinct quadrics passing through these points because a 6×4-matrix $\left(s_j^k t_j^l\right)_{\substack{k+l=0,1,2, \\ j=1,2,3,4}}$ has rank four. (Why this conclusion is a right one?) The readers may find this rank as follows. A linear transformation $(as + bt,\ cs + dt) \mapsto (s,t)$ $(ad - bc \neq 0)$ converts this matrix to the matrix $\left((as_j + bt_j)^k (cs_j + dt_j)^l\right)_{\substack{k+l=0,1,2, \\ j=1,2,3,4}}$ that has the columns corresponding to polynomials in (s,t) of degrees $k + l$, respectively. First, prove that the usual monomials $s^k t^l$ of degree $k + l = d$ $(d = 0,1,2)$ are linear combinations of the polynomials $(as + bt)^k \cdot (cs + dt)^l$ of the same degree $k + l = d$. In fact, a similar claim holds for any number of variables:

Lemma 1 *A linear transformation* $\left(\sum_\nu a_{1\nu} z_\nu, \ldots, \sum_\nu a_{n\nu} z_\nu\right) \mapsto (z_1, \ldots, z_n)$ *(*$\det(a_{\mu\nu}) \neq 0$*) yields the set of the monomials* $\prod z_\mu^{k_\mu}$ *of any fixed degree* $\sum k_\mu = d$ *to a set of linearly independent polynomials.*

(Proving this lemma is a good exercise in linear algebra. Also, the readers may have a look at a proof at the end of section E12.39.) By this lemma, $\left((as_j + bt_j)^k (cs_j + dt_j)^l\right)_{\substack{k+l=0,1,2, \\ j=1,2,3,4}}$ has the same rank as $\left(s_j^k t_j^l\right)_{\substack{k+l=0,1,2, \\ j=1,2,3,4}}$. Make a

[22] Readers may easily prove the implication "*a straight line* $l(s,t) = 0$ *is contained in an algebraic curve* $p(s,t) = 0$" ⇒ "$p(s,t)$ *is divided by* $l(s,t)$," by making an affine change of the independent variables so that this line would become the ordinate axes $t = 0$. (The similar arguments allow an immediate multivariate generalization substituting a straight line with an affine hyperplane.)

change of coordinates by a linear rotation so that the second of the new coordinates (which we denote t_j, as before) would be distinct at the given four points. And so, the claim about rank is reduced to its special case:

Lemma 2 *6×4 matrices $\left(s_j^k t_j^l\right)_{\substack{k+l=0,1,2,\\ j=1,2,3,4}}$ with distinct t_j have rank four if (and only if) the points (s_j,t_j) do not lie on a straight line.*

Proof. For s, insert a Lagrange interpolating polynomial (section P12.19*) $s(t)$ on T-system $\{1,t,t^2,t^3\}$ taking values s_j at t_j, respectively, into the matrix. It may be zero or have degree 0, 1, 2, 3, depending on the location of the points $A_j = (s_j,t_j)$. Polynomials equal to zero or ones of degrees 0, 1 correspond to cases where these points lie on a straight line (why? create a figure), and inserting $s(t)$ shows that only three columns (corresponding, say, to 1, t, t^2) are linearly independent. (Why?) If deg $s(t) = 2$ or deg $s(t) = 3$, then we have four linearly independent columns (corresponding, say, to 1, t, t^2, st, or, respectively, 1, t, t^2, s; complete the details). **QED**.

This completes the proof. In fact, the polynomials $p'_4(s,t)$ and $p''_4(s,t)$ may be defined in the following way:

- If the four points are vertices of a quadrangle, as products of pairs of polynomials of degree one, defining the sides or the diagonals (all four must be distinct).
- If A_1, A_2, A_3 lie on a straight line $l(s,t) = 0$, A_1, A_4 lie on a straight line $l'(s,t) = 0$, and A_2, A_4 lie on a straight line $l''(s,t) = 0$, as $p'_4 = l \cdot l'$ and $p''_4 = l \cdot l''$.

Proof of Lemma 1. The rows of the matrix $A = (a_{\mu v})_{\mu,v\,=\,1,\ldots,\,n}$ correspond to scalar multiplications by linearly independent vectors $a_\mu = (a_{\mu v})_{v\,=\,1,\ldots,\,n}$: $Az = \begin{pmatrix} \langle a_1, z\rangle \\ \vdots \\ \langle a_n, z\rangle \end{pmatrix}$. Thus, the linear dependence of the transformed monomials would mean an identity $\sum_{k_1+\ldots+k_n=d} \alpha_{k_1,\ldots,k_n} \prod_{\mu=1}^{n} \langle a_\mu, z\rangle^{k_\mu} = 0$ with some $\alpha_{k_1,\ldots,k_n} \neq 0$. Apply a double induction: the external on $n = \dim z$ and the internal on degree d. Inserting z from an orthogonal complement to a_n yields the identity $\sum_{k_1+\ldots+k_{n-1}=d} \alpha_{k_1,\ldots,k_{n-1},0} \prod_{\mu=1}^{n-1} \langle a'_\mu, z\rangle^{k_\mu} = 0$ ($z \in a_n{}^\perp$), where a'_μ are orthogonal components of a_μ along $a_n{}^\perp$. Hence, by the hypothesis of the external induction, $\alpha_{k_1,\ldots,k_n} = 0$ if some $k_\mu = 0$. Dividing the rest of the sum by $\prod_{\mu=1}^{n} \langle a_\mu, z\rangle$ yields an identity of a smaller degree on an open and everywhere dense set of z and so, by continuity, on the whole space. The proof is completed with a reference to the hypothesis of the internal induction. Readers should complete the details. Considering $A = E$ shows the linear independence of the monomials themselves.

E12.40

Start by computing the mean value and variance of a univariate Gaussian distribution: $\left(\sqrt{2\pi}\sigma\right)^{-1}\int_{-\infty}^{\infty} y\, e^{-y^2/2\sigma^2}dy = 0$, $\left(\sqrt{2\pi}\sigma\right)^{-1}\int_{-\infty}^{\infty} y^2 e^{-y^2/2\sigma^2}dy = \sigma^2$. (The first of these equalities is obtained by integrating an odd function. The second one is obtained with integration by parts on $[0,\infty)$ using the equality $\left(\sqrt{2\pi}\sigma\right)^{-1}\int_{0}^{\infty} e^{-y^2/2\sigma^2}dy = 1/2$. Work out the details.[23]) The desired multivariate equalities straightforwardly follow from the univariate ones applying the Fubini theorem.

Completing the Solution

S12.4

The subspace $L = \text{im}\, A$ is of dimension equal to or less than $m + 1$ (which is the dimension of the domain of definition of the operator A). Consider L as a subspace of the coordinate space $\mathbb{R}^{m+l+1} = \mathbb{R}^{m+1} \bigoplus \mathbb{R}^l$; restrict the projection onto the ($m + 1$-dimensional) coordinate plane $\mathbb{R}^{m+1}$, of a kernel $\mathbb{R}^l$, to L. The dimension of the image of this restriction will equal $\dim L - \dim (L \cap \mathbb{R}^l)$. Therefore, $L \cap \mathbb{R}^l = 0$ when L is projected onto all of $\mathbb{R}^{m+1}$. (Fill in the details.)

In turn, L will be projected onto all of $\mathbb{R}^{m+1}_{i_0,\ldots,i_m}$ if and only if the rows of matrix A, of indices $i_0,\ldots,i_m$, are linearly independent (why?), showing the equivalence of (1) holding any inequalities as $V_l(\overrightarrow{t}, \overrightarrow{k}) \neq 0$ and (2) linear independence of any $m + 1$ rows of matrix A. Apply induction on l. Using the inequalities $V_{l-1}(\overrightarrow{t}, \overrightarrow{k})>0$ (for $0 < t_1 <\ldots< t_{l-1}$ and natural $k_1 <\ldots< k_{l-1}$) as the inductive hypothesis, first, derive the inequalities $V_l(\overrightarrow{t}, \overrightarrow{k}) \neq 0$; then reproduce the inductive hypothesis completely, proceeding as suggested in section E12.4 to establish that $V_l(\overrightarrow{t}, \overrightarrow{k})>0$.

To realize the first step of this program, factorize A as $A = A^l_{t_1,\ldots,t_l} := A^{(l)}_{t_l} \circ \ldots \circ A^{(1)}_{t_1}$, where $A^{(j)}_t$ is an $(m + j + 1) \times (m + j)$ matrix $\begin{pmatrix} -t & & \\ 1 & \ddots & \\ & \ddots & -t \\ & & 1 \end{pmatrix}$ (assuming that

[23] Iterating this calculation provides values for all **central moments** of a normal (Gaussian) distribution: $\left(\sqrt{2\pi}\sigma\right)^{-1}\int_{-\infty}^{\infty} y^{2p} e^{-y^2/2\sigma^2}dy = \sigma^{2p}(2p-1)!!$; also, $\left(\sqrt{2\pi}\sigma\right)^{-1}\int_{0}^{\infty} y^{2p+1} e^{-y^2/2\sigma^2}dy = \sigma^{2p+1}(2\pi)^{-1/2}(2p)!!$ ($p = 0,1,\ldots$; $(-1)!! = 0!! = 1$).

unfilled entries equal zero). The **left kernel** of the matrix $A_t^{(j)}$ [24] is spanned on vector $(1,t,\ldots,t^{m+j})$. Therefore, since the left kernel of $A = A^l_{t_1,\ldots,t_l}$ is l-dimensional (why?), and, obviously, $A = A^l_{t_{\sigma(1)},\ldots,t_{\sigma(l)}}$ for any permutation of indices σ: $\{1,\ldots,l\} \to \{1,\ldots,l\}$, the left kernel of A is spanned on vectors $(1,t_j,\ldots,t_j^{m+l+1})$, $j = 1,\ldots,l$.[25] Hence, since the rows of A have the form $vA^{l-1}_{t_1,\ldots,t_{l-1}}$, where v are rows (of the same indices) of $A^{(l)}_{t_l}$, the linear dependence of the rows of indices $i_0,\ldots,i_m$ would imply that a nontrivial linear combination of these rows $\sum \alpha_k v_{i_k}$ belongs to the left kernel of $A^{l-1}_{t_1,\ldots,t_{l-1}}$, so the rows of the $(m+l)\times(m+l)$ matrix $M = \begin{pmatrix} & v_{i_0} & \\ & \vdots & \\ & v_{i_m} & \\ 1 & \cdots & t_1^{m+l-1} \\ \vdots & & \vdots \\ 1 & \cdots & t_{l-1}^{m+l-1} \end{pmatrix}$ are linearly dependent, so $\det M = 0$. (Complete the details.) Therefore, prove that $\det M \neq 0$, or, which is the same thing, $\det {}^tM \neq 0$. If $i_0 = 1$, then this problem may be reduced to proving that $\det M' \neq 0$ for M' obtained from ${}^tM$ by deleting its first row and column (because $t_l \neq 0$; work out the details); similarly, if $i_1 = 2$, then the problem may be further reduced to proving that $\det M'' \neq 0$ for M'' obtained from M' by deleting its first row and column; and so on. If the first column of M, or M', etc. has 1 as its first nonzero element, then move the rows lying above this element to the end, making the row containing 1 the first, and, if the column has $-t_l$ as its second element, then multiply the first row by t_j and add to the second one. Readers may verify that the process combining these elementary operations results in the matrix $\begin{pmatrix} 1 & & & \\ & \ddots & & * \\ & & 1 & \\ & 0 & & B \end{pmatrix}$, where B is an $(l-1)\times(l-1)$ matrix of a determinant equal to $\pm\sum_{d,\vec{k}} t_l^d \cdot V_{l-1}(\overrightarrow{t}, \overrightarrow{k})$. Applying the inductive hypothesis and taking into account that $t_l > 0$ gives $\det B \neq 0$, completing the proof.

Remark. As attentive readers have noticed, the linear independence of any $m+1$ rows of matrix A and, therefore, the presence of $l+1$ nonzero coefficients in polynomials having l positive roots, appeared in our proof with no regard to the distinction of t_l from $t_1,\ldots,t_{l-1}$, indicating a generalization of the Chebyshev (and Descartes) properties of monomials for multiple zeros. Such a generalization is discussed in sections P12.8*, P12.9*, and P12.10*.

[24] The left (right) kernel of an $m \times n$ matrix A is the space of all row (resp. column) vectors v of dimension m (resp. n) such that $vA = 0$ (resp. $Av = 0$).

[25] Also, the left kernel may be calculated as the kernel of the **adjoint operator** A^* (which acts on row vectors as $A^*v = vA$), using the orthogonality of $\ker A^*$ with $\operatorname{im} A$. (Work out the details.)

S12.6

On the proof of the following statement: *Obeying the Descartes rule of signs by a system* $\{f_0,\ldots f_n\} \Rightarrow$ *the determinants* $\det\left(f_i(t_j)\right)_{\substack{i=i_0,\ldots,i_m,\\ j=0,\ldots,m}}$ *have equal signs for any subsequences* $0 \le i_0 <\ldots< i_m \le n$ *(*$m \le n$*), while* $t_0 <\ldots< t_m$ *are fixed*. We must verify that the signs of the determinants are the same as that for a special subsequence of indices $0,\ldots, m$. Indeed, none of a series of successive replacements of one index, starting from an arbitrary subsequence $i_0,\ldots,i_m$, as

$$(i_0,\ldots,i_m) \;\to\; (0,i_1,\ldots,i_m) \;\to\; (0,1,i_2,\ldots,i_m) \;\to\ldots\to\; (0,\ldots,m),$$

can change the sign (as discussed in section H12.6). **QED**.

S12.8

Proof of the following statement: the determinants W for an ET-system $\{f_0,\ldots f_n\}$ have the same signs for any $t_0 \le\ldots\le t_n$ when f_j have continuous nth derivatives. When $n = 0$, $W(f_0)(t_0) = f_0(t_0) \ne 0$ for any $t_0 \in I$, so they have the same sign by virtue of continuity f_0. When $n > 0$, the nonvanishing of the determinants W and the property of these determinants considered in the lemma from section E12.11 make their sign the same for two sequences of arguments: $t_0 \le\ldots\le t_n$, and $t_0' <\ldots< t_n'$, obtained from the first one by replacing the repetitions $\ldots< t_k =\ldots= t_{k+m} <\ldots$ with arithmetic progressions $t_k <\ldots< t_k + mh$ with a small $h > 0$. (Work out the details.) In turn, the equality of W's signs for any sequences of arguments without repetitions is implied by the continuity of W on the set of such sequences. (Fill in the details using the continuity of differentiable functions f_j.) QED.

Readers have probably noticed that the continuity of the nth derivatives did not appear in this proof for $n > 0$. Thus, the proved claim is applicable to systems of functions without this property, for example, $F = \{f_0, f_1\}$, where $f_0(t) = \begin{cases} t^2\cos(1/t) + C, & \text{when } t \ne 0 \\ 0, & \text{when } t = 0 \end{cases}$, with a large enough positive constant C, and $f_1(t) = e^t$. [Actually, F forms an ED-system on $I = (-\infty,\infty)$, as defined in section P12.9*; readers may work out the details using the second statement in section P12.11* below.]

S12.11

On the proof of Statement 1. The arguments from sections H12.11 and E12.11 can in no way prove that F is an ET-system on I as soon as the maximal-order

Wronskian does not have zeros on I because each f_i of F may take zero values in I (indeed, consider $F = \{\cos t, \sin t\}$ on a segment I of a length not smaller than π) and, if some f_0 does not have zeros, the derivatives of all of f_i/f_0 may have zeros, and so on.

S12.14

Completing the case of $W > 1$. For small $\varepsilon_y > 0$ the functions q_{y+h} with $|h| < \varepsilon_y$ have the same, modulo 2, number of zeros (considering multiplicity) as q_y because q_y has the same number of zeros as $f_m \cdot q_y$, so this number is W at most. (Why?) Also, q_y is W times differentiable and $|q_y| \to \infty$ at the endpoints of I (using the same arguments as for q). Together, these properties formally imply the required result, as explained in section E12.14. **QED**.

Remark on the lemma from section E12.14. As readers have probably noticed, the claim of this lemma related to $k = 1$ is a special (for dimension one) case of the implicit (or inverse) function theorem (more exactly, a part of it, without considering the differentiability of the inverse function $y \mapsto x$).

A differentiable function $f = x + o(x)$ with the derivative discontinuous at $x = 0$, such that an equation $\varepsilon = f(x)$ has at least k distinct small positive roots for any small $\varepsilon > 0$, may be created as $f = x + x^2\sin(T/x)$, using a large enough constant T. We will prove this by finding such T that for any small $\varepsilon > 0$ there exists $\delta \in (0,\varepsilon)$ such that $\varepsilon - \delta + (\varepsilon - \delta)^2 \geq \varepsilon$ and the set of T/x includes a segment of length $\geq 2\pi k$, when the points x fill an area including the segment $[\varepsilon - \delta, \varepsilon]$. (Why? First make a figure, then answer!) We leave the computations to the reader; show that if $T > 2\pi k$, then the required δ exist for $\varepsilon \leq [T/(2\pi k)]^2 - T/(2\pi k)$.

S12.17

(1) On the algebraic method. A proof of linear independence of h_i using Lemma 1 in section E12.39. Functions h_i are restrictions of monomials of degree n, $x^i y^{n-i}$, to the unit circle. Thus, if $\sum c_i h_i = 0$, then the linear combination (which is a polynomial) $\sum c_i x^i y^{n-i}$ will vanish on the whole unit circle, so it will vanish everywhere (due to its homogeneity; work out the details), and therefore, by virtue of Lemma 1 in section E12.39, the coefficients c_i must be zeros. **QED**.

Irreducibility of $f(x,y) = x^2 + y^2 - 1$ over a coefficient field of any characteristic distinct from two. Indeed, otherwise we would have $x^2 + y^2 - 1 = (a_1x + b_1y + c_1)(a_2x + b_2y + c_2)$; with this, we may scale the coefficients in the factors so that $c_1 = 1$ and $c_2 = -1$. Then $a_1 = a_2 = \pm 1$, and a similar statement holds regarding b_1, b_2. (Work out the details.) Hence, we would obtain $0 = a_1b_2 + a_2b_1 = \pm 2$, meaning simply the equality of the characteristic to two.

Readers familiar with the ideas of projective geometry may use different arguments to prove the irreducibility of $x^2 + y^2 - 1$ over the reals. A polynomial identity such as $p(x) = q(x) \cdot r(x)$ (of any number of variables and over any coefficient field) is equivalent to the identity $P(x_0,x) = Q(x_0,x) \cdot R(x_0,x)$, where for a polynomial of degree k $p(x) = \sum_{i+j+\ldots \le k} a_{ij\ldots} x_1^i x_2^j \ldots$, $P(x_0,x)$ is a homogeneous polynomial $P(x_0,x) = \sum_{i+j+\ldots \le k} a_{ij\ldots} x_0^{k-i-j-\ldots} x_1^i x_2^j \ldots$. [How are zero sets of $p(x)$ and $P(x_0,x)$ linked to each other?] Therefore, we must verify the irreducibility of the quadratic form $x^2 + y^2 - z^2$. Readers know that a quadratic form (of any number of variables) has a unique matrix representation $(x,y,\ldots) \cdot A \cdot {}^t(x,y,\ldots)$ with respect to any fixed basis, using a symmetric matrix A; but in addition, there are infinitely many similar representations with respect to the same basis, using asymmetric matrices A. Show that a quadratic form may be factorized if and only if it possesses the preceding representation with a matrix (perhaps asymmetric) of rank one (obtainable by multiplying a column vector by a row vector; why?). However, the symmetric matrix representation is obtained from any asymmetric one by a symmetrization $A \to \frac{1}{2}(A + {}^tA)$, so its rank cannot exceed two for a factorable form (as the rank of a sum does not exceed the sum of ranks). But $\mathrm{rk}\begin{pmatrix} 1 & & \\ & 1 & \\ & & -1 \end{pmatrix} = 3$, completing the proof.

Also, **readers may use geometrical arguments showing that a circle is not a union of straight lines**!

A homogeneous polynomial in x, y identically vanishes, being divisible by $x^2 + y^2 - 1$. Indeed, it vanishes on the whole unit circle, and so vanishes everywhere (due to its homogeneity).

A planar set consists of at most n points if its orthogonal projections onto any straight lines consist of at most n points. Indeed, for any fixed straight line l, this set, which we denote S, lies on a union of (not necessarily distinct) straight lines $r_1,\ldots,r_n$ orthogonal to l (create a figure!), and the same holds for any straight line l' transversal to l; therefore, S is finite (to be exact, $\#S \le n^2$; why?). Therefore, S is a bounded set. Hence, if we pick the transversal line l' close to l, S will lie within the union of intersections $\bigcup_{i=1}^{n} (r_i \cap r'_i)$. (Why?) Hence, $\#S \le n$. **QED**.

The common points of two similar (i.e., obtainable from one another by a combination of a homothetic transformation and a parallel translation) quadrics in a vector space belong to a hyperplane. Indeed, verify that the difference of two quadratic equations $\langle Ax,x\rangle = r$, $\langle A(x - x_0), x - x_0\rangle = s$ is a linear (in x) equation.

(2) On the analytic method. A proof of the equivalence of the reciprocity of a polynomial to the following property: end coefficients (a_0 and a_n) are complex conjugate and the roots lying outside the unit circle are distributed into pairs of mutually inverse complex numbers. The reciprocity of $p(z) = a_0 \prod_k \left(z - e^{i\theta_k}\right) \cdot \prod_j \left(z - \rho e^{i\theta_j}\right)\left(z - \rho^{-1} e^{i\theta_j}\right)$ such that $e^{i\left(\sum \theta_k + 2\sum \theta_j\right)} = (-1)^n \bar{a}_0/a_0$ may be established by elementary computations. [**Remark**: a direct verification shows that the equation $\bar{a}/a = e^{i\theta}$ ($\theta \in \mathbb{R}$) is solvable with respect to a and defines a up to a multiplication by arbitrary nonzero real numbers.] Conversely, let a polynomial $p(z) = a_0 z^n + \ldots + a_n$ be reciprocal. It cannot have zero as its root (since $a_n = \bar{a}_0 \neq 0$), so z_0 is a root if (and only if) $\bar{z}_0^{-1}$ is. (Work out the details.) Thus, $p(z)$ is

divisible by a linear binomial $az + \bar{a} = a\left(z - e^{i\theta}\right)$ such that $\bar{a}/a = e^{i\theta}$ or by a quadratic trinomial $a(z - \rho e^{i\theta})$ $(z - \rho^{-1}e^{i\theta})$ such that $\bar{a}/a = e^{2i\theta}$ and $\rho \neq 1$, allowing the proof to be completed by induction, considering that a quotient of division of reciprocal polynomials is reciprocal itself. (We leave it to the reader to work out the details.)

S12.20

Sufficiency. The "hardest case": $k = 0$, $l = n - 1$, $I = [a,b]$, $t_1 = a$. A limit polynomial has no zeros except $t_1 = a, t_2, \ldots, t_{n-1}$. Indeed, $\forall v$, $p^{(v)}$ have the same sign in any of the intervals $(t_1^{(v)}, t_2)$, $(t_2, t_3), \ldots, (t_{n-1}, b)$, so the limit polynomial p will have the same sign in $(a, t_2), \ldots, (t_{n-1}, b)$, respectively. (Work out the details.) Therefore, $p(b) \neq 0$, p cannot have any nodal zeros except $t_1 = a, t_2, \ldots, t_{n-1}$, while extra nonnodal ones are prohibited by the "sum of indices" condition. **QED**.

Sines of angles between vectors $v = (f_0(t), \ldots, f_n(t))$ and $v_n = (f_0(b), \ldots, f_n(b))$ for $t \leq t_0 < b$ are bounded from zero. Indeed, those vectors are not proportional for $t < b$ (because f_j form a T-system), so the angles between them are distinct from 0 and π, so for a compact segment $a \leq t \leq t_0$ bounded away from b the sines of those angles are bounded from zero (by virtue of continuity; work out the details).

For any vectors $v_1, \ldots, v_n$ of an n-dimensional Euclidean space, the determinant of the Gram matrix $G = (\langle v_i, v_j \rangle)_{i,j = 1, \ldots, n}$ is equal to the squared volume of the parallelepiped spanned on them. Readers who have not yet come across this fact may easily prove it as follows. For any orthobasis $e_1, \ldots, e_n$, consider a linear operator A translating e_i into v_i, $\forall i$. We have $G = {}^t\mathrm{AA}$, where A is the matrix for A with respect to basis $e_1, \ldots$, and so $\det G = (\det \mathrm{A})^2 = [\mathrm{vol}_n(v_1, \ldots, v_n)]^2$. **QED**. [We leave it to the reader to fill in the details using the equality of the oriented parallelepiped's volume to the determinant, following from the basic volume's properties (n-linearity and skew symmetry) known from standard university courses.]

S12.25

A twice-differentiable regular closed curve on a plane that has a positive curvature will be free of self-intersections and bound a convex (planar) body if the tangent's polar angle θ changes by 2π for the complete traversal of it. Readers may prove this by steps **A**–**G** as follows.

A. Absence of self-intersections (and self-tangencies). Assume the opposite is true. Using for (x_0, y_0) a point of self-intersection (or self-tangency) and turning the

axes so that $\theta_0 = 0$ yields the equations $\int_0^{\theta_1} R(\theta)\sin\theta\,d\theta = 0$, $\int_{\theta_1}^{2\pi} R(\theta)\sin\theta\,d\theta = 0$ for some $0 < \theta_1 < 2\pi$. If $\theta_1 \le \pi$ ($\theta_1 \ge \pi$), then the first (resp. second) equation will contradict the positiveness of R. **QED**. (We leave it to the reader to fill in the details.)

More advanced readers familiar with differential topology know a generalization of the preceding claim, which connects the number of full turns of the normal vector (the **degree of the Gauss map**) to an algebraic sum of self-intersections (**the Whitney number**), and its far-reaching further generalizations.

B. Prove that our curve intersects with any of its tangents with multiplicity two at the point of tangency. [Use Cartesian coordinates with the origin at this point and the tangent for the abscissas' axis, which allows one to describe the curve in a neighborhood of this point by an equation of the form $y = \pm\kappa x^2/2 + o(x^2)$. Work out the details.]

C. Prove that the number of intersections modulo 2 between a straight line and our curve, considering multiplicity, does not depend on this line. For this, use the two following one-parameter families of straight lines consisting of variations of some (arbitrarily fixed) line: (1) parallel translations and (2) rotations around a fixed point that is not lying on the curve. Apply the arguments from section E12.14 to the function: $\theta \mapsto$ *the parameter value corresponding to the line of the family passing* through *the point* $(x(\theta),y(\theta))$ *of the curve*. [The lemma from section E12.14 is applicable since all zeros of this function have multiplicities not exceeding the order of its smoothness (two) – as established in step **B**; in turn, the finiteness of the multiplicities shows that the zeros are isolated, and thus the function has a finite number of zeros; we leave it to the reader to complete the proof.]

There are straight lines that do not intersect with our curve at all (as the curve is bounded; work out the details). Consequently, any straight line intersects with it (a finite and) an even number of times, considering multiplicity.

D. Establish the following geometric version of the Lagrange intermediate value theorem.

For an open differentiable regular planar curve γ: $[0,1] \to \mathbb{R}^2$ *(*$\dot\gamma \ne 0$*) and the chord* l *connecting the points* $A = \gamma(0)$ *and* $B = \gamma(1)$*, the tangent at some intermediate (of* $0 < t_0 < 1$*) point* $C = \gamma(t_0)$ *is parallel to* l.

This obviously implies the usual analytic version of the Lagrange theorem (for scalar functions of one real variable) but cannot be easily reduced to it because the curve may lack a representation by the graph of such a function. (Illustrate it by appropriate figures.) Readers may prove this geometric theorem as follows. Let a straight line l have a common point with the curve. There exists a parallel translation of l by a maximal distance such that the translated line still has a common point with the curve. (Prove.) This line runs tangent to the curve at all its common points with the curve. Indeed, under the transversal intersection, the line could be translated "a little bit more" still having a common point with the curve, which contradicts the maximality of the distance. (Illustrate it by a figure, and use the lemma from section E12.14 to fill in the details.) **QED**.

The tangent to a differentiable regular curve $t \mapsto (x(t),y(t))$ at the point (x_0,y_0) corresponding to $t = t_0$ may be defined by the parametric equations $(x,y) = (u(t),v(t))$, where $u = x_0 + \dot{x}_0(t - t_0)$, $v = y_0 + \dot{y}_0(t - t_0)$ ($\dot{x}_0 = \dot{x}(t_0)$, $\dot{y}_0 = \dot{y}(t_0)$). Without loss of generality, $\dot{x}_0 \neq 0$ (what does this mean geometrically?), so $v(t) = y_0 + (\dot{y}_0/\dot{x}_0)(u(t) - x_0)$. A different straight line passing through (x_0,y_0) is defined by the equation $y = y_0 + k(x - x_0)$ ($k \neq \dot{y}_0/\dot{x}_0$); hence, its point having an abscissa $u(t)$ possesses the ordinate $w(t) = y_0 + k\dot{x}_0(t - t_0)$. Thus we will have $w(t) - v(t) = a(t - t_0)$, with $a = k\dot{x}_0 - \dot{y}_0 \neq 0$, so $y(t) - v(t) = o(t - t_0) = o(w(t) - v(t))$. **QED**. (We leave it to the reader to fill in the details.)

E. By the results obtained in steps **C** and **D**, a straight line may have at most two (considering multiplicity) common points with our curve. Indeed, there would be four such points if there were three (according to **C**), while four of those would yield three distinct points of the curve, where the tangents are parallel to the same straight line (according to **D**; create corresponding figures for three cases of multiple/singular intersections); but this contradicts a monotone change in the tangent's polar angle θ. (We leave it to the reader to work out the details.)

F. Now we are able to introduce a formal concept reflecting an intuitive proposition: *A curve bounds a planar body*. A commonly accepted approach (due to French mathematician Camille Jordan) consists of the following. By step **E**, a rectilinear ray has no, or one, or two, considering multiplicity, intersections with our curve; the start point of the ray is considered *internal* (*external*) for one (resp. no or two) intersection(s).

A similar approach is applicable to closed curves of arbitrary curvatures, to nonsmooth curves, and even to multidimensional closed hypersurfaces. That is, the famous **Jordan-Lebesgue-Brouwer theorem** states that a hypersurface homeomorphic to a unit hypersphere divides $\mathbb{R}^n$ into two connected domains (being common boundary) one of the domains is bounded, the other is unbounded; the rays of a general position, with starting points in the bounded (unbounded) domain, have an odd (resp. even) number of common points with the hypersurface.

G. Denote by S the set ("body") of all internal points of the curve as defined in step **F**. It is convex because any rays with the starting points on the line segment connecting some points $A,B \in S$ touch the curve once and only once. (Why?)

Also, readers may establish the convexity of S by a different method. As proved in step **E**, our curve lies entirely in one of two closed half-spaces defined by each of its tangents. (Work out the details.) Show that a point is internal if and only if it belongs to all of these half-spaces. Therefore, S is an intersection of half-spaces, so it is convex (as discussed previously in section P9.2* from the "Convexity and Related Classic Inequalities" problem group). **QED**.

On the proof of the four-vertex theorem: verification of the orthogonality relations $\int_0^{2\pi} R_0(\theta)\,d\theta = 0$, $\int_0^{2\pi} R_0(\theta)\cos\theta\,d\theta = 0$, $\int_0^{2\pi} R_0(\theta)\sin\theta\,d\theta = 0$. The first one follows directly from the definition of R_0 because $\int_0^{2\pi} R_0\,d\theta = \int_0^{2\pi}\left(R - (2\pi)^{-1}\int_0^{2\pi} R\,d\theta\right)d\theta = \int_0^{2\pi} R\,d\theta - \left((2\pi)^{-1}\int_0^{2\pi} R\,d\theta\right)\int_0^{2\pi} d\theta = 0$. The second and third ones are due to the periodicity (by the definition) of x, y as functions of θ, namely, $x(0) = x(2\pi)$ and $y(0) = y(2\pi)$.

S12.26

A continuous function bounded below and growing on infinity on a finite-dimensional normed space has a minimum on any closed set. Indeed, this function takes, at some points inside a ball of a large enough radius, smaller values than anywhere outside this ball. Therefore, we must verify the existence of a minimum on a bounded closed set that is compact, due to finite dimensionality.[26] Taking into account the continuity of the function completes the proof. (Work out the details.)

A similar claim is invalid for any infinite-dimensional normed space, which is clear from the following counterexample. Consider a bounded closed set S that is a disjoint union of a countably infinite number of closed balls. (Verify the existence of this set!) Set a function φ equal to one on the union of the balls' spheres and extend it outside S so that it is continuous and satisfies the inequality $\varphi \geq 1$. Inside the balls extend φ continuously so its minimum on the nth ball is n^{-1}. The restriction φ to any closed set containing S (e.g., S itself or the whole space) does not have a minimum.

We can introduce a more sophisticated example showing that a closed affine hyperplane in an infinite-dimensional normed space (even complete) may lack a point of a minimal distance from the origin. Take, in the space $C([0,1])$ of continuous functions on the segment $[0,1]$, with a norm of the maximum of the absolute value of a function, a hyperplane $\left\{x \in C([0,1]) : \int_0^1 x(t)\chi(t)\,dt = 1\right\}$, with $\chi(t) := \begin{cases} 1, & \text{for } 0 \leq t < 1/2 \\ -1, & \text{for } 1/2 \leq 1 \\ \text{anything, for } t = 1/2 \end{cases}$. (Work out all the details, and also think about why a similar thing could not happen in a Hilbert space.)

The existence of a polynomial of best approximation to a given continuous function on any fixed finite system of continuous functions. For a vector subspace of the space of continuous functions on a closed segment, the maximum of the absolute value of a function is a norm. Any norm defines a distance function, which is continuous and growing (to infinity) on infinity; with respect to this norm, $\|f_n - f\| \to 0$ & $\|g_m - g\| \to 0 \Rightarrow \|f_n - g_m\| \to \|f - g\|$, and $\|f - g\| \to \infty$ if $\|g\| \to \infty$. (Why?) Therefore, the proof may be completed by applying the previous claim to this function and the subspace spanned on $F \bigcup \{f\}$ (with the notations in section P12.26***); work out the details.

[26] Readers not yet familiar with the compactness of bounded closed sets in finite-dimensional spaces may verify it using quite elementary means. Advanced readers probably know an inverse theorem proved by the famous twentieth-century Hungarian mathematician Frigyes (Frederic) Riesz: a locally compact normed space is finite-dimensional. For a proof and further discussions, see Riesz and Sz.-Nagy (1972), Dieudonné (1960), Rudin (1973), and Banach (1932).

S12.28

Examples of affine planes in $C([a,b])$ (of any, including infinite, dimensions) containing entire convex regions of elements of the least deviation from zero. Consider the planes generated by sets of $f \Subset C([a,b])$ satisfying the linear equation $f(a) = 1$ and the inequality $|f| \leq 1$. (These sets are the desired regions; fill in the details.)

Equivalence of the balls' strict convexity with the strictness of the triangle inequality. The strictness of the triangle inequality implies, for noncollinear x, y,

$$\begin{aligned} \|x\|, \|y\| \leq 1 \quad &\Rightarrow \quad \|\alpha x + (1-\alpha)y\| < \|\alpha x\| + \|(1-\alpha)y\| \\ &= \alpha\|x\| + (1-\alpha)\|y\| \ \leq \alpha + (1-\alpha) = 1 \end{aligned}$$

(when $0 < \alpha < 1$), showing the balls' strict convexity. (Why?) Conversely, this convexity implies the strict triangle inequality. Indeed, consider noncollinear x, y; denote $\|x\| = a$, $\|y\| = b$. Since $a,b > 0$ (why?), a point $x + y$ is a convex linear combination of points $u = (a + b)x/a$, $v = (a + b)x/b$, with positive coefficients, namely, $x + y = (au + bv)/(a + b)$; or, equally, $x + y$ belongs to the interior of the line segment of these endpoints. (Create a figure!) Since we have $\|u\| = \|v\| = a + b$, the equality "$\|x + y\| = a + b$" cannot hold because it contradicts the balls' strict convexity. (Why?)

S12.31

By the Chebyshev theorem (section P12.27***), $2^{1-n}cT_n(t)$ has the least deviation from zero over the polynomials on the segment $[-1,1]$ of degree n with leading coefficient c. In turn, by Haar's theorem (section P12.28***), this polynomial is unique. (Work out the details.) Use similar arguments to prove the similar minimum property of $2^{1-n}cT_n(\lambda(t))$ for the domain $[a,b]$, where λ is a linear map $[a,b] \to [-1,1]$. The rest of section P12.31*** may be done straightforwardly. (We leave it to the reader to fill in the details.)

S12.32

Deriving the unique solvability condition for the least-squares problem in $\mathbb{R}^{n+1}_{u_1,\ldots u_n,y} = \{u_0 = 1\} \subset \mathbb{R}^{n+2}_{u_0,\ldots u_n,y}$ from a similar condition for the least-squares problem in $\mathbb{R}^{n+2}_{u_0,\ldots u_n,y}$. If a point of $\mathbb{R}^{n+1}_{u_0,\ldots u_n}$ belongs simultaneously to a hyperplane $P \neq \{u_0 = 0\}$ passing through the origin and to an affine hyperplane $H = \{u_0 = 1\}$, then this

is equivalent to its orthogonal projection into $\mathbb{R}^n_{u_1,\ldots u_n}$ belonging to the $n-1$-dimensional affine plane of this space, which is the projection of $P \bigcap H$. (Create a figure and work out the details.)

Alternatively, readers may perform the reduction using algebraic arguments as follows. $n+1$-dimensional vectors $v_j = (1,u_{1j},\ldots,u_{nj})$, $j = 1,\ldots,m$, are linearly dependent if and only if the vectors $w_1 = v_1$, $w_2 = v_2 - v_1,\ldots$, $w_m = v_m - v_1$ are linearly dependent, which in turn is equivalent to the linear dependence of $m-1$ n-dimensional row vectors $z_2,\ldots z_m$ obtained from $w_2,\ldots w_m$, respectively, by crossing out their first (zero) elements (why?), which in turn is equivalent to containing m n-dimensional vectors $(u_{1j},\ldots,u_{nj})$ by an $m-2$-dimensional affine plane in $\mathbb{R}^n_{u_1,\ldots u_n}$. (Work out the details and complete the proof.)

S12.36

For $m = 2$, a straight line $y = x_0 + x_1 u$, which is the solution of the least-squares problem, contains both (u_1,y_1) and (u_2,y_2), so it contains any points expressed by a linear combination of these points with coefficients summed to one. (Why?)

S12.37

For $m = n+1$, a plane that is the solution of the least-squares problem contains the points $(u_{1j},\ldots,u_{nj},y_j)$, $j = 1,\ldots,m$, so it contains any points expressed by a linear combination of these points with coefficients summed to one (by arguments similar to those in section S12.36; fill in the details).

For an arbitrary $m \geq n+1$, the result can be obtained by dividing the kth equation from section H12.37 by $\sum_l u_{kl}$, which yields $\sum_j \alpha_{kj} y_j = x_0 + x_1 \sum_j \alpha_{kj} u_{1j} + \ldots + \sum_j \alpha_{kj} u_{nj}$. (Work out the details.)

More advanced readers might prefer a different line of reasoning as follows. With the matrix notations $A = {}^t(u_{ij})$ and $\Gamma = (\alpha_{kj})$, the first column of $\Gamma \circ A$ is filled with ones, so the only thing left is to verify that $\forall y$: ${}^tA \circ Ax = {}^tAy \Rightarrow \Gamma \circ Ax = \Gamma y$. The proportionality of the rows of Γ to the corresponding rows of ${}^tA$ yields $\ker \Gamma \supseteq \ker {}^tA$, proving the desired claim since, obviously, $\forall y$: ${}^tA \circ Ax = {}^tAy \Rightarrow T \circ Ax = Ty$ for any matrices T of m columns, such that $\ker T \supseteq \ker {}^tA$ (and only for them). **QED**.

Remark. As readers may verify directly, $\Gamma = X \circ {}^tA$ for the matrix $X = (x_{ki})$ of the entries $x_{ki} = \delta_{ki} \Big/ \sum_l u_{kl}$. Actually, for any fixed linear operator L, a linear

operator M defined on the same domain and such that $\ker M \supseteq \ker L$ (and, obviously, only this one) can be presented by a composition $M = X \circ L$, using an appropriate linear operator X defined on $\operatorname{im} L$ (the exact image of L). (Obviously, X is unique; why?). To prove this, show the existence of a commutative diagram of linear operators

$$\begin{array}{ccccc} \operatorname{dom} L & \underset{M}{\searrow} & \operatorname{dom} L/\ker L & \overset{\bar{L}}{\underset{X}{\swarrow}} & \operatorname{im} L \\ & & \downarrow{\scriptstyle Y} & & \\ & & \operatorname{im} M & & \end{array}$$

The upper row of this diagram corresponds to L, that is, $\bar{L}$ translates $x + \ker L$ into Lx ($x \in \operatorname{dom} L$). The existence of Y is brought about by an inclusion $\ker M \supseteq \ker L$. Finally, X exists as defined by the composition $X = Y \circ \bar{L}^{-1}$. (Work out the details.)

Readers may find X with the help of a least-squares-solution-like formula $X = M \circ {}^tL \circ (M \circ {}^tL)^{-1}$ (defining ${}^tL$ on $\operatorname{im} L$ only; work out the details).

S12.39

Geometric algorithm. A polynomial on a finite-dimensional vector space (over any field F) is divisible by $l(x) = \sum a_i x_i - b$ (in the polynomial ring $F[x]$) if (and only if) it vanishes on the hyperplane $\sum a_i x_i = b$. This follows from the unique factorization property of $F[x]$ and the primality of an element $l(x)$. Indeed, readers may prove this statement by changing variables, as was done in section S3.2 ("A Combinatorial Algorithm in Multiexponential Analysis" problem group discussed previously).

For a finite subset $\{A_1, \ldots\}$ in a Euclidean space there exists a Cartesian coordinate system such that all abscissae (the first coordinates) A_{i1} are distinct. (This is true even for an infinite subset with a cardinality smaller than continuum. This subset may or may not be everywhere dense.) Indeed, the union of hyperplanes Π_{ij} orthogonal to the vectors $A_i - A_j$, respectively, cannot cover all of the space, so it cannot include all of the unit sphere. (Why? Work out the details, and then complete the proof for a finite set of A_i. Readers familiar with duality may do the same for the sets of any cardinalities smaller than the continuum using the following argument: the hyperplanes (vector subspaces of codimension one) in a finite-dimensional vector space form a set that has cardinality of the continuum, because obviously this set is bijective to the unit hypersphere in the dual vector space (the space of all linear functionals), in which the diametrically opposite points are identified with each other, so there exists a hyperplane Π distinct from all Π_{ij}. For a dimension equal to two, a straight line Π is not covered with the union of Π_{ij}, and for any dimension Π contains a straight line that is not covered with the union of $\Pi \bigcap \Pi_{ij}$, by induction. We leave it to the reader to work out the details.)

On the proof of Lemma 2 in section E12.39: given a set of pairs (t_j,s_j) with distinct t_j, the degree of a Lagrange interpolating polynomial is less than or equal to one if and only if the points (t_j,s_j) are located on a straight line. Indeed, this follows from the uniqueness of the Lagrange polynomial and the elementary fact that polynomials of degrees not exceeding one, and only such polynomials, have rectilinear graphs. (We leave it to the reader to fill in the details.)

A different proof of the vanishing of Lagrange polynomial terms of degrees greater than one can be obtained using the explicit formulas for this polynomial's coefficients in section P1.11*** (from the problem group "Jacobi Identities and Related Combinatorial Formulas"), taking into account that any symmetric polynomial of degree k in $x_1,\ldots,x_n$ can be expressed by a linear combination of polynomials $\sum x_j^i$, $i = 0,\ldots, k$, and the Jacobi identities in section P1.1**. (We leave it to the reader to work out the details.)

References

Abragam, A.: The Principles of Nuclear Magnetism. Oxford University Press, London/Toronto (1961)

Adams, J.F.: Lectures on Lie Groups. W.A. Benjamin, New York/Amsterdam (1969)

Ahlfors, L.: Complex Analysis. Science/Engineering/Math, 3rd edn. McGraw-Hill, New York (1979)

Akhiezer, N.I.: Классическая проблема моментов и некоторые вопросы анализа, связанные с ней. "Физматгиз" Press, Moscow (1961). [English transl. The Classical Moment Problem and Some Related Questions in Analysis. Oliver and Boyd Press, Edinburgh/London (1965)]

Akhiezer, N.I., Glazman, I.M.: Теория линейных операторов в гильбертовом пространстве. "Гостехиздат" Press, Moscow (1950). [English transl. Theory of Linear Operators in Hilbert Space. Ungar, New York (1963) (Reprinted by Dover)]

Alperin, J.L., Bell, R.B.: Groups and Representations. Graduate Texts in Mathematics, 1st edn. Springer, New York (1995)

Andrews, G.E.: The Theory of Partitions. Addison-Wesley, Reading/London/Amsterdam/Don Mills/Sydney/Tokyo (1976)

Andrews, G.E., Eriksson, K.: Integer Partitions. Cambridge Univeristy Press, Cambridge (2004)

Apostol, T.M.: Calculus, V.1: One-Variable Calculus, with an Introduction to Linear Algebra, 2nd edn. Wiley, New York (1967)

Apostol, T.M.: Calculus, V.2: Multi-Variable Calculus and Linear Algebra with Applications to Differential Equations and Probability, 2nd edn. Wiley, New York (1969)

Apostol, T.M.: Modular Functions and Dirichlet Series in Number Theory. Springer, New York (1990)

Arkhangelsky, A.V., Ponomharev, V.I.: Основы общей топологии в задачах и упражнениях. "Наука" Press, Moscow (1974). [English transl. Fundamentals of General Topology: Problems and Exercises. Elsevier Science (1979)]

Arnol'd, V.I.: Обыкновенные дифференциальные уравнения. "Наука" Press, Moscow (1975). [English transl. Ordinary Differential Equations. MIT Press, Cambridge (1978)]

Arnol'd, V.I.: Дополнительные главы теории обыкновенных дифференциальных уравнений. "Наука" Press, Moscow (1978). Геометрические методы в теории обыкновенных дифференциальных уравнений. Regular & Chaotic Dynamics, 2nd edn. MCNMO/VMK NMU (1999). [English transl. Geometrical Methods in the Theory of Ordinary Differential Equations. Springer, Berlin/Heidelberg (2004)]

Arnol'd, V.I.: Теорема Штурма и симплектическая геометрия. Функц. анализ и его прилож. **19**(4), 1–10 (1985). [English transl. The Sturm theorem and symplectic geometry. Functional Anal. App. **19**(4)]

Arnol'd, V.I.: Сто задач (One Hundred Problems). МФТИ Press, Moscow (1989). Russian

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8, © Springer Science+Business Media, LLC 2013

Arnol'd, V.I.: Гюйгенс и Барроу, Ньютон и Гук: пионеры математического анализа и теории катастроф от эвольвент до квазикристаллов. "Наука" Press, Moscow (1989 Historical). [English transl. Huygens and Barrow, Newton and Hooke: Pioneers in Mathematical Analysis and Catastrophe Theory from Evolvents to Quasicrystals. Birkhaüsser, Boston (1990)]

Arnol'd, V.I.: Математические методы классической механики. 3rd edn. "Наука" Press, Moscow (1989). [English transl. Mathematical Methods of Classical Mechanics (Graduate Texts in Mathematics, No 60). 2nd edn. Springer (1989)]

Arnol'd, V.I.: Теория катастроф. 3rd edn. "Наука" Press, Moscow (1990). [English transl. Catastrophe Theory. 3rd edn. Springer (1992)]

Arnol'd, V.I.: Математический тривиум. УМН **46**(1), 225–232 (1991). [English transl. Mathematical trivium. Russian Math. Surveys **46**(1)]

Arnol'd, V.I.: Topological Invariants of Plane Curves and Caustics. University Lecture Series, vol. 5. American Mathematical Society, Providence (1994)

Arnol'd, V.I.: Лекции об уравнениях с частными производными (Lectures on Partial Differential Equations). "Фазис" Press, Moscow (1997). [English transl. Lectures on Partial Differential Equations (Universitext). 1st edn. Springer (2004)]

Arnol'd, V.I.: Что такое математика (What is Mathematics)? МЦНМО Press, Moscow (2002). Russian

Arnol'd, V.I.: Задачи для детей от 5 до 15 лет (Math Problems for the Kids (ages 5–15)). МЦНМО Press, Moscow (2004). Russian

Arnol'd, V.I., Gusein-Zade, S.M., Varchenko, A.N.: Особенности дифференцируемых отображений Т.1. "Наука" Press, Moscow (1982). [English transl. Singularities of Differentiable Maps. V.1. The Classification of Critical Points, Caustic and Wave Fronts. Boston (1985)]

Arnol'd, V.I., Novikov, S.P. (eds.): Динамические системы IV. Итоги науки и техники. Современные проблемы математики. Фундаментальные направления, Т.4. ВИНИТИ Press, Moscow (1985). [English transl. Dynamical Systems IV. Symplectic Geometry and its Applications. Springer (1997)]

Arnol'd, V.I., Kozlov V.V., Neishtadt, A.I.: Динамические системы III (Математические аспекты классической и небесной механики). Итоги науки и техники. Современные проблемы математики. Фундаментальные направления, Т.3. ВИНИТИ Press, Moscow (1985). [English transl. Mathematical Aspect of Classical and Celestial Mechanics (Encyclopedia of Mathematical Sciences, V. 3.) 2nd edn. Springer (1993)]

Arrowsmith, D.K., Place, C.M.: Ordinary Differential Equations. A Qualitative Approach with Applications. Chapman & Hall, London/New York (1982)

Artin, E.: Geometric Algebra. Interscience, New York/London (1957)

Artin, M.: Algebra. Prentice Hall, New York (1991)

Atiyah, M.F., MacDonald, I.G. (1994). Introduction to commutative algebra. Addison-Wesley Publishing Co., Reading (Mass.), London, Don Mills, Ont.

Aubin, J.-P., Ekeland, I.: Applied Nonlinear Analysis. Wiley-Interscience, New York/Chichester/Brisbane/Toronto/Singapore (1984)

Balakrishnan, A.V.: Applied Functional Analysis. Springer, New York/Heidelberg/Berlin (1976)

Banach, S.: Théorie des operations linéaries. Paris (1932). [English transl. Theory of Linear Operations (Dover Books on Mathematics). Dover (2009)]

Barwise, J. (ed.): Handbook of Mathematical Logic, II. North-Holland, Amsterdam/New York/Oxford (1977)

Beckenbach, E.F., Bellman, R.: Inequalities. Springer, Berlin/Göttingen/Heidelberg (1961)

Belitskii, G.R., Lubich, Yu.I.: Нормы матриц и их приложения. "Наукова Думка" Press, Kiev (1984). [English transl. Matrix Norms and their Applications (Operator Theory Advance and Applications, V. 36). Springer (1988)]

Bellman, R.: Introduction to Matrix Analysis. McGraw-Hill, New York/Toronto/London (1960)

Berezin, I.S., Zhidkov, N.P.: Методы вычислений, ТТ. 1–2. "Физматгиз" Press, Moscow (1959–1960). [English transl. Computing Methods. Franklin Book Company (1965)]

Berger, M.: Géométrie. Cedic/Nathan, Paris (1977). French

Bernshtein, D.N.: Число корней системы уравнений. Функц. анализ и его прилож. **9**(3), 1–4 (1975). [English transl. The number of roots of a system of equations. Functional Anal. App. **9**(3), 183–185]
Bishop, R.L., Crittenden, R.J.: Geometry of manifolds. Academic, New York/London (1964)
Blaschke, W.: Einführung in die Differentialgeometrie. Springer, Berlin/Göttingen/Heidelberg (1950)
Bochner, S., Martin, W.T.: Several Complex Variables. Princeton University Press, Princeton (1948)
Borel, E.: Probability and Certainty. Walker Sun Books, SB-20. Physics and Mathematics. Walker, New York (1963)
Borevich, Z.I., Shafarevich, I.R.: Теория чисел. 2nd edn. "Наука" Press, Moscow (1972). [English transl. Number Theory. Academic Press, New York (1986)]
Bott, R., Tu, L.W.: Differential Forms in Algebraic Topology. Graduate Texts in Mathematics. Springer, New York (1982)
Bourbaki, N.: Éléments de mathématique. Premiére partie. Livre III. Topologie générale. Hermann, Paris (1960)
Bourbaki, N.: Éléments de mathématique. Variétés différentielles et analytiques. Hermann, Paris (1967–1971)
Bourbaki, N.: Éléments de mathématique. Groups et algébres de Lie. Hermann, Paris/Masson/New York/Barcelona/Milan/Mexico/Rio de Janeiro (1968–1982)
Bourbaki, N.: Éléments de mathématique. Théorie des ensembles. Hermann, Paris (1970a)
Bourbaki, N.: Éléments de mathématique. Algèbre: chapitres 1 à 3, 4 à 6, 7 à 9. Hermann & Cie, Paris (1970b)
Bourbaki, N.: Éléments de mathématique. Algèbre commutative. Hermann, Paris (1972)
Bourbaki, N.: Éléments de mathématique. Fonctions d'une variable réelle. Hermann, Paris (1976)
Bourbaki, N.: Éléments de mathématique. Espaces vectoriels topologiques, 2nd edn. Masson, Paris (1981)
Bredon, G.E.: Introduction to Compact Transformation Groups. Academic, New York/London (1972)
Briskin, M., Elichai, Y., Yomdin, Y.: How can singularity theory help in image processing. In: Gromov, M., Carbone, A. (eds.) Pattern Formation in Biology, Vision and Dynamics, pp. 392–423. World Scientific, Singapore (2000)
Bruce, J.W., Giblin, P.: Curves and Singularities: A Geometrical Introduction to Singularity Theory. Cambridge University Press, Cambridge (1993)
Brudnyi, A., Yomdin, Y. Remez Sets (preprint)
Burago, D.M., Zalgaller, V.A.: Геометрические неравенства. "Наука" Press, Leningrad (1980). [English transl. Geometric Inequalities (Grundlehren Der Mathematischen Wissenschaften, 285 a Series of Comprehensive Studies in Mathematics). Springer (1988)]
Carmo do, M.P.: Differential Geometry of Curves and Surfaces. Prentice Hall, New Jersey (1976)
Cartan, H.: Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes. Hermann, Paris (1961). [English transl. Elementary Theory of Analytic Functions of One or Several Complex Variables (Dover Books on Mathematics). Dover (1995)]
Cartan, H.: Calcul Différentiel. Formes Différentielles. Hermann, Paris (1967). [English transl. Differential Calculus. Houghton Mifflin Co (1971), and Differential Forms (Dover Books on Mathematics). Dover (2006)]
Cassels, J.W.S.: An Introduction to Diophantine Approximation. Cambridge University Press, Cambridge (1957)
Cassels, J.W.S.: An Introduction to the Geometry of Numbers. Springer, Berlin/Göttingen/Heidelberg (1959)
Cassels, J.W.S.: Rational Quadratic Forms. Academic, London/New York/San Francisco (1978)
Céa, J.: Optimisation. Théorie et algorithmes. Dunod, Paris (1971)
Charlier, C.L.: Die Mechanik des Himmels. Walter de Gruyter, Berlin/Leipzig (1927)

Chentzov, N.D., Shklarsky, D.O., Yaglom, I.M.: Геометрические нервенства и задачи на максимум и минимум (Geometric Inequalities and Problems on Maximum and Minimum). "Наука" Press, Moscow (1970). Russian
Chevalley, C.: Theory of Lie Groups I, II, and III. Princeton University Press, Princeton/New Jersey (1946–1955)
Chinn, W.G., Steenrod, N.E.: First Concepts of Topology. New Mathematical Library; The School Mathematics Study Group. Random House, New York/Toronto (1966)
Choquet, G.: Ľenseignement de la géometrie. Hermann, Paris (1964). French
Coddington, E.A., Levinson, N.: Theory of Ordinary Differential Equations. McGraw-Hill, New York/Toronto/London (1955)
Cohen, P.J.: Set Theory and the Continuum Hypothesis. Dover Books on Mathematics. Dover, New York (2008)
Collatz, L.: Funktionalanalysis und Numerische Mathematik. Die Grundlehren Der Mathematischen Wissenschaften, vol. 120. Springer, Berlin/Göttingen/Heidelberg (1964)
Courant, R.: Differential and Integral Calculus, vol. 1–2. Wiley-Interscience, New York (1992)
Courant, R., Hilbert, D.: Methods of Mathematical Physics. Wiley, New York (1953–1962)
Courant, R., Robbins, H.: What is Mathematics?: An Elementary Approach to Ideas and Methods. Oxford University Press, London (1941)
Cramér, H.: Mathematical Methods of Statistics. Princeton University Press, Princeton (1946)
Dayson, F.J.: Missed opportunities. Bull. AMS **78**, 635 (1972)
Dieudonné, J.: Foundation of Modern Analysis. Academic, New York/London (1960)
Dieudonné, J.: Algébre Linéarie et Géometrie Élémentarie (Linear algebra and elementary geometry). Troisiéme Ėdition, Paris (1968 French)
Dubrovin, B.A., Fomenko, A.T., Novikov, S.P.: Современная геометрия – Методы и приложения. "Наука" Press, Moscow (1986). [English transl. Modern Geometry - Methods and Applications, Part I: The Geometry of Surfaces of Transformation Groups, and Fields (Graduate Texts in Mathematics, No 93). Springer (1990)]
Dummit, D.S., Foote, R.M.: Abstract Algebra, 3rd edn. Wiley, New York (2003)
Dunford, N., Schwartz, J.T.: Linear Operators. Part I. General Theory. Wiley-Interscience, New York (1957)
Dunford, N., Schwartz, J.T.: Linear Operators. Part II. Spectral Theory. Self Adjoint Operators on Hilbert Space. Interscience, New York/London (1963)
Edwards, R.E.: Functional Analysis. Theory and Applications. Holt, Rinehart and Winston, New York/Chicago/San Francisco/Toronto/London (1965)
Edwards, C.H., Penney, D.E.: Elementary Differential Equations with Boundary Value Problems, 4th edn. Prentice Hall, New Jersey (1999)
Ekeland, I., Temam, R.: Convex Analysis and Variational Problems. North-Holland/American Elsevier, Amsterdam/Oxford/New York (1976)
Elichai, Y., Yomdin, Y.: Normal forms representation: A technology for image compression. SPIE. **1903**, Image and Video Processing, 204–214 (1993)
Engelking, R.: General Topology (Monografie Matematyczne, 60). manuscript of the 2nd edn. Państowe Widawnictwo Naukowe, Warszawa (1985). [English transl. of the 1st edn. General Topology. Taylor & Francis Press (1977)
Erdelyi, A. (ed.): Higher Transcendental Functions, vol. 1–3. McGraw-Hill, New York/Toronto/London (1953)
Evgrafov, M.A.: Аналитические функции (Analytic Functions), 3rd edn. "Наука" Press, Moscow (1991). Russian
Favard, J.: Cours de Géométrie Différentielle Locale. Gauthier-Villars, Paris (1957)
Feigenbaum, M.J.: Quantitative universality for a class of nonlinear transformation. J. Stat. Phys. **19**, 25–52 (1978)
Feigenbaum, M.J.: The universal metric properties of non-linear transformations. J. Stat. Phys. **21**, 669–706 (1979)
Feller, W.: An Introduction to Probability Theory and its Applications, vol. 1–2, 2nd edn. Wiley/Chapman&Hall, New York/London (1966)
Fermi, E.: Thermodynamics. Prentice Hall, New York (1937)

Feynman, R.P., Leighton, R.B., Sands, M.: The Feynman Lectures on Physics. V. 1. Addison-Wesley, Reading/Palo Alto/London (1963)

Fichtengoltz, G.M.: Курс дифференциального и интегрального исчисления (Course in Differential and Integral Calculus), vol. 1–3, 7th edn. "Наука" Press, Moscow (1970). Russian

Finnikov, S.P.: Курс дифференциальной геометрии (A Course of Differential Geometry). "ГИТТЛ" Press, Moscow (1952). Russian

Florian, A.: Zu einem Satz von P. Erdös. Elemente der Mathematik **13**(3), 55–58 (1958)

Folland, G.B.: Real Analysis: Modern Techniques and their Applications. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, New York (1999)

Fraenkel, A.A., Bar-Hillel, Y., Levy, A.: Foundations of Set Theory, 2nd edn. North-Holland, Amsterdam (1973)

Franklin, J.: Sur le développment du prodit infini $(1-x)(1-x^2)\ldots$. C.R. Acad. Fr. **92**, 448–450 (1881)

Fuchs, L.: Infinite Abelian Groups. V.1. Academic, New York/London (1970)

Fuchs, D.B.: О раскрытии скобок, об Эйлере, Гауссе, Макдональде и об упущенных возможностях. "Квант" ("Kvant"). **8**, 12–20 (1981). [English transl. On collecting like terms, on Euler, Gauss and MacDonald, and on missed opportunities. In: Fuchs, D.B., Tabachnikov, S.: Mathematical omnibus: Thirty lectures on classic mathematics; lecture 3. American Mathematical Society (2007)]

Fuchs, D.B.: Когомологии бесконечномерных алгебр Ли. "Наука" Press, Moscow (1984). [English transl. Cohomology of Infinite-Dimensional Lie Algebras. Contemporary Soviet Mathematics. Consultants Bureau, New York (1986)]

Fulton, W., Harris, J.: Representation Theory: A First Course. Graduate Texts in Mathematics/ Readings in Mathematics. Springer, New York (1991)

Gantmacher, F.R.: Теория матриц, 3rd edn. "Наука" Press, Moscow (1967). [English transl. Theory of Matrices. V.'s 1–2. Chelsea (2000)]

Gantmacher, F.R., Krein, M.G. Осцилляционные матрицы и ядра, и малые колебания механических систем. "Гостехиздат" Press, Moscow-Leningrad (1950). [English transl. Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems. US Atomic Energy Commission, Washington (1961)]

Gardner, R.J.: Geometric Tomography. Cambridge University Press, Cambridge/New York/ Melbourne (1995)

Gelbaum, B.R., Olmsted, J.M.H.: Counterexamples in Analysis. Holden-Day, San Francisco/ London/Amsterdam (1964)

Gelfand, I.M.: Лекции по линейной алгебре, 4th edn. "Наука" Press, Moscow (1971). [English transl. Lectures on Linear Algebra (Dover Books on Mathematics). Dover (1989)]

Gelfand, I.M., Shilov, G.E., Vilenkin, N.Ya., Graev, N.I.: Обобщённые функции, TT. 1–5. "Наука" Press, Moscow (1959–1962). [English transl. Generalized Functions. V.'s 1–5. Academic Press (1964)]

Gilmore, R.: Catastrophe Theory for Scientists and Engineers. Willey-Interscience, New York/ Chichester/Brisbane/Toronto (1981)

Givental, A.B..: Полиномиальность электростатических потенциалов (Polynomiality of electrostatic potentials). УМН **39**(5), 253–254 (1984). Russian

Glazman, I.M., Lubich, Yu.I.: Конечномерный линейный анализ в задачах. "Наука" Press, Moscow (1969). [English transl. Finite-Dimensional Linear Analysis: A Systematic Presentation in Problem Form (Dover Books on Mathematics). Dover (2006)]

Godbillon, C.: Géométrie différentielle et mécanique analytique. Collection Méthodes. Hermann, Paris (1969)

Gödel, K.: The Consistency of the Axiom of Choice and the Generalized Continuum Hypothesis with the Axioms of Set Theory. Ann. Math Studies, vol. 3. Princeton University Press, Princeton/New Jersey (1940)

Golubitsky, M., Guillemin, V.: Stable Mappings and their Singularities. Graduate Texts in Mathematics. Springer, New York (1974)

Grosswald, E.: Topics in the Theory of Numbers. Birkhäuser, Boston (1983)
Guillemin, V., Adams, M.: Measure Theory and Probability. Birkhaüsser, Boston (1996)
Guillemin, V., Pollack, A.: Differential Topology. Prentice Hall International, New Jersey (1974)
Gupta, H.: Selected Topics in Number Theory. Abacus, Tunbridge Wells (1980)
Hadamard, J.: Leçons de Géometrie Élémentarie (Lessons in elementary geometry). V.1 (1898): Géometrie plan (Plane geometry), V.2 (1901): Géometrie dans Léspace (Solid geometry). "Librairie Armand Colin," Paris (1898–1901 French)
Hadamard, J.: The Psychology of Invention in the Mathematical Field/The Mathematician's Mind. Dover/Princeton University Press, New York (1954/1996)
Hadamard, J.: Lessons in Geometry. American Mathematical Society, Providence (2008). Har/Cdr ed
Hall, M.: Combinatorial Theory. Blaisdell Press, Waltham/Toronto/London (1967)
Hall, M.: Theory of Groups, 2nd edn. Chelsea, New York (1975)
Halmos, P.R.: A Hilbert Space Problem Book. D. Van Nostrand, Princeton/New Jersey/Toronto/London (1967)
Halmos, P.R.: Finite-Dimensional Vector Spaces. Springer, New York (1974)
Hardy, G.H.: A Course in Pure Mathematics. Cambridge University Press, Cambridge (1908)
Hardy, G.H., Rogosinski, W.W.: Fourier Series. Cambridge Tracks in Mathematics and Math Physics, vol. 38. Cambridge University Press, Cambridge (1956)
Hardy, G.H., Wright, E.M.: An Introduction to the Theory of Numbers, 4th edn. Clarendon, Oxford (1960)
Hardy, G.H., Littlewood, J.E., Polya, G.: Inequalities. Cambridge University Press, Cambridge (1934)
Hartman, P.: Ordinary Differential Equations. Wiley, New York/London/Sydney (1964)
Hausdorff, F.: Grundzüge der Mengenlehre. Von Veit, Leipzig (1914). [English transl. Set Theory, 2nd edn. Chelsea, New York (1962)]
Haviv, D., Yomdin, Y.: Model based representation of surfaces (preprint)
Helgason, S.: Differential Geometry and Symmetric Spaces. Academic, New York/London (1962)
Helgason, S.: The Radon Transform. Progress in Mathematics, vol. 5. Birkhaüsser, Boston/Basel/Stuttgart (1980)
Helgason, S.: Groups and Geometric Analysis. Integral Geometry, Invariant Differential Operators, and Spherical Functions. Academic (Harcourt Brace Jovanovich), Orlando/San Diego/San Francisco/New York/London/Toronto/Montreal/Tokyo/São Paulo (1984)
Herman, G.T.: Image Reconstruction from Projections. The Fundamentals of Computerized Tomography. Academic, New York/London/Toronto/Sydney/San Francisco (1980)
Herman, G.T., Tuy, H.K., Langenberg, K.J., Sabatier, P.C.: Basic Methods of Tomography and Inverse Problems. Adam Gilger, Bristol/Philadelphia (1987)
Herstein, I.N.: Noncommutative Rings. Carus Mathematical Monographs, 5th edn. Mathematical Association of America Textbooks, Washington (2005). Book 15
Hilbert, D.: Grundlagen der Geometrie. Leibzig, Berlin (1930). [English transl. Foundations of Geometry, 2nd edn. Open Court (1988)]
Hilbert, D., Cohn-Vossen, S.: Anschauliche Geometrie. Berlin (1932). [English transl. Geometry and the Imagination, 2nd edn. Chelsea, New York (1990)]
Hille, E.: Lectures on Ordinary Differential Equations. Addison-Wesley, Reading/Menlo Park/London/Don Mills (1969)
Hille, E., Phillips, R.S.: Functional Analysis and Semi-Groups. AMS Colloquium Publications, vol. XXX1. American Mathematical Society, Providence (1957). Rev. ed
Hirsch, M.W.: Differential Topology. Graduate Texts in Mathematics, vol. 33. Springer, Berlin/Heidelberg/New York (1976)
Hörmander, L.: An Introduction to Complex Analysis in Several Variables. D. Van Nostrand, Princeton/New Jersey/Toronto/New York/London (1966)
Hörmander, L.: The Analysis of Linear Partial Differential Operators I. Distribution Theory and Fourier Analysis. Springer, Berlin/Heidelberg/New York/Tokyo (1983)

Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge/New York/New Rochelle/Melbourne/Sydney (1986)
Humphreys, J.E.: Introduction to Lie Algebras and Representation Theory. Graduate Texts in Mathematics, vol. 9. Springer, New York (1973)
Hurwitz, A., Courant, R.: Vorlesungen ueber Allgemeine Funktionentheorie und Elliptische Funktionen./Geometrische Funktionentheorie. Springer, Berlin/Göttingen/Heidelberg/New York (1964)
Infeld, L.: Whom the Gods Love: The Story of Evariste Galois. National Council of Teachers of Mathematics, Reston (1948)
Ireland, K., Rosen, M.: A Classical Introduction to Modern Number Theory. Graduate Texts in Mathematics, vol. 87. Springer, New York/Heidelberg/Berlin (1982)
Itskovich, G., Roytvarf, A.: Method of successive determination of parameters in the multiexponential analysis problem. Proc. Int. Seminar "Inverse problems in Geophysics", vol. 181, Novosibirsk (1996)
Itskovich, G., Roytvarf, A.: Signal processing method for determining the number of exponential decay parameters in multiexponentially decaying signals and its application to nuclear magnetic resonance well logging. US 005764058A (1998)
Jacobi, C.G.J.: Disquisitiones analyticae de fractionibus simplicibus. Berlin (1825)
Jacobi, C.G.J.: Vorlesungen über Dynamik. Aufl. 2. G. Reimer, Berlin (1884). [English transl. Mathematical Aspect of Classical and Celestial Mechanics (Texts and Readings in Mathematics, V.51). 2nd rev. edn. Hindustan Book agency (2009)]
Jech, T.: Set Theory, 3rd edn. Springer, Berlin (2006)
Kac, M.: Some Stochastic Problems in Physics and Mathematics. Lectures in Pure and Applied Science, vol. 2. Magnolia Petroleum, Dallas (1956)
Kac, M.: Probability and Related Topics in Physical Sciences. Lectures in Applied Mathematics Series, vol. 1a. American Mathematical Society, Providence (1957)
Kac, M.: Statistical Independence in Probability, Analysis and Number Theory. Carus Mathematical Monographs, American Mathematical Society (1959)
Kac, V.G.: Infinite-dimensional algebras, Dedekind's η-function, classical Möbius function and the very strange formula. Adv. In Math. **30**, 85–136 (1978). Preprint MIT
Kac, V.G.: Infinite-Dimensional Lie Algebras, 3rd edn. Cambridge University Press, Cambridge (1994)
Kac, M., Santalo, L.A.: Integral Geometry and Geometric Probability. Cambridge University Press, Cambridge (2004)
Kak, A.C., Slaney, M.: Principles of Computerized Tomographic Imaging. IEEE Press, New York (1988)
Kamke, E.: Diferentialgleichungen. Lösungsmethoden und Lösungen, 1. Gewöhnliche Diferentialgleichungen, 3rd edn. Leipzig (1977)
Kantorovich, L.V., Akilov, G.P.: Функциональный анализ, 2nd edn. "Наука" Press, Moscow (1972). [English transl. Functional Analysis. Pergamon Press, Oxford, XIV (1982)]
Kargapolov, M.I., Merzljakov, Yu.I.: Основы теории групп, 2nd edn. "Наука" Press, Moscow (1977). [English transl. Fundamentals of the Theory of Groups (Graduate Texts in Mathematics, No 62). Springer (1979)]
Karlin, S., Studden, W.J.: Tchebycheff Systems: With Application in Analysis and Statistics. Interscience Publishers A. Divison of Willey, New York/London/Sydney (1966)
Kazarinoff, N.D.: Geometric Inequalities. Random House, New York/Toronto (1961)
Kelley, J.L.: General Topology. Graduate Texts in Mathematics. Springer, New York (1975)
Khovanskii, A.G.: Малочлены. "Фазис" Press, Moscow (1997). [English transl. Fewnomials. Translations of Mathematical Monographs 88, AMS, Providence/Rhode Island (1991)]
Kirillov, A.A.: Элементы теории представлений. "Наука" Press, Leningrad (1972). [English transl. Elements of the Theory of Representations (Grundlehren Der Mathematischen Wissenschaften). Springer (1976)]

Kirillov, A.A., Gvishiani, A.D.: Теоремы и задачи функционального анализа, 2nd edn. “Наука” Press, Moscow (1988). [English transl. from the 1st edn. Theorems and Problems in Functional Analysis (Problem Books in Mathematics). Springer, NY/Heidelberg/Berlin (1982)]

Klein, F.: Elementarmathematik vom Höheren Stamdpunkte aus Erster Band. Arithmetik. Algebra. Analysis. (Dritte Auflage). Springer, Berlin (1924). [English transl. Elementary Mathematics from an Advanced Standpoint: Arithmetic, Algebra, Analysis (Dover Books on Mathematics). Dover Publ. Inc (2004)]

Klein, F.: Elementarmathematik vom Höheren Stamdpunkte aus Erster Band. Geometrie. (Dritte Auflage). Springer, Berlin (1925). [English transl. Elementary Mathematics from an Advanced Standpoint: Geometry (Dover Books on Mathematics). Dover Publ. Inc (2004)]

Klein, F.: Vorlesungen über die Entwicklung der Mathematic im 19. Jahrhundert. Teil 1. Für den Druck bearbeitet von Courant, R., Neugebauer, O. Springer, Berlin (1926). [English transl. Development of Mathematics in the Nineteenth Century (Lie Groups Series, No 9). Applied Mathematics Group publishing (1979)]

Kline, M.: Mathematics and the Physical World. Dover Books Explaining Science. Dover, New York (1981)

Kline, M.: Mathematics: The Loss of Certainty. Galaxy Books. Oxford University Press, USA (1982)

Kline, M.: Mathematics and the Search for Knowledge. Oxford University Press, USA (1986)

Kobayashi, S., Nomizu, K.: Foundations of Differential Geometry, V.I. Interscience, New York/London (1963)

Kobayashi, S., Nomizu, K.: Foundations of Differential Geometry, V.II. Interscience, New York/London/Sydney (1969)

Kolmogorov, A.N.: Основные понятия теории вероятностей, 2nd edn. “Наука” Press, Moscow (1974). [English transl. Foundations of the probability theory, 2nd English edn. Chelsea Publ. Co, NY (1956)]

Kolmogorov, A.N., Fomin, S.V.: Элементы теории функций и функционального анализа, 4th edn. “Наука” Press, Moscow (1976). [English transl. Elements of the Theory of Functions and Functional Analysis (Dover Books on Mathematics). Dover Publ. Inc (1999)]

Kosinski, A.A.: Differential Manifolds. Dover Books on Mathematics. Dover, New York (2007)

Kostrikin, A.I., Manin, Yu.I.: Линейная алгебра и геометрия. МГУ Press, Moscow (1980). [English transl. Linear Algebra and Geometry (Algebra, Logic and Applications). Taylor & Francis Press (1997)]

Krasnoselskii, M.A., Vainikko, G.M., Zabreyko, P.P., Rutitskii, Ya.B., Stetzenko, V.Ya.: Приближённые решения операторных уравнений. “Nauka” Press, Moscow (1969). [English transl. Approximate Solutions of Operator Equations. Wolters-Nordhoff, Groningen (1972)]

Krein, S.G., ed. coaut.: Функциональный анализ (Справочная математическая библиотека) 2nd rev. edn. “Наука” Press, Moscow 1972). [English transl. of the 1st edn. Functional Analysis. Foreign Technology Division MT-65-573. U.S. Dept. Commerce, Nat. Bur. Standards, Washington/DC; microfiche distr. by Clearinghouse for Federal Sci. Tech. Inform., Springfield/VA, MR 38 #2560 (1968)]

Krein, M.G., Nudelman, A.A.: Проблема моментов Маркова и экстремальные задачи. “Наука” Press, Moscow (1973). [English transl. The Markov Moment Problem and Extremal Problems, Translations of Math. Monographs, V.50. American Mathematical Society, Providence/Rhode Island (1977)]

Kreyszig, E.: Differential Geometry. Dover Books on Mathematics. Dover, New York (1991)

Kuratowski, K.: Topology, V.1. Academic/Polish Scientific, New York/London/Warszawa (1966). rev. ed

Kuratowski, K., Mostowski, A.: Set Theory, with an Introduction to Descriptive Set Theory. Studies in Logic and the Foundations of Mathematics, vol. 86, 2nd edn. North-Holland, Amsterdam (1976)

Kutateladze, S.S., Rubinov, A.M.: Двойственность Минковского и её приложения (Minkowski Duality and its Applications). “Наука” Press, Novosibirsk (1976). Russian

Lakatos, I.: Proofs and Refutations: The Logic of Mathematical Discovery. Cambridge University Press, Cambridge (1976)
Lakatos, I.: The Methodology of Scientific Research Programmes, V.1: Philosophical Papers; Mathematics, Science and Epistemology, V.2: Philosophical Papers. Cambridge Univercity Press, Cambridge (1980)
Landau, E.: Elementare Zahlentheorie (Vorlesungen über Zahlentheorie, B.1). Leipzig (1927). [English transl. Elementary Number Theory (with added exercises by Bateman, P.T., & Kohlbecker, E.E.). AMS, Chelsea (1999)]
Landau, E.: Grundlagen der Analysis. Akademische Verlagsgesellschaft M.B.H., Leipzig (1930). [English transl. Foundation of Analysis (Graduate Studies in Mathematics). American Mathematical Society, Chelsea (2001)]
Landau, E.: Einführung in die Differentialrechnung und Integralrechnung. Groningen, Batavia, Noordhoff (1934). [English transl. Differential and Integral Calculus, 3rd edn. American Mathematical Society, Chelsea (2001)]
Landau, L.D., Lifshitz, E.M.: Теоретическая физика, Т.1: Механика. "Наука" Press, Moscow (1973). [English transl. Mechanics, 3rd edn. Butterworth-Heinemann (1976)]
Landau, L.D., Lifshitz, E.M.: Теоретическая физика, Т.5: Статистическая физика. "Наука" Press, Moscow (1976). [English transl. Statistical Physics, 3rd edn. Pergamon Press, Oxford/New York (1979)]
Landau, L.D., Lifshitz, E.M.: Теоретическая физика, Т.8: электродинамика сплошных сред "Наука" Press, Moscow (1982). [English transl. Electrodynamics of continuous media, 2nd edn. Butterworth-Heinemann (1984)]
Landau, L.D., Lifshitz, E.M.: Теоретическая физика, Т.6: Механика сплошных сред. "Наука" Press, Moscow (1986). [English transl. Fluid Mechanics, 2nd edn.. Pergamon Press, Oxford/New York (1987)]
Landau, L.D., Lifshitz, E.M.: Теоретическая физика, Т.7: Теория упругости. "Наука" Press, Moscow (1987). [English transl. Theory of Elasticity, 3rd edn. Pergamon Press, Oxford/New York (1986)]
Lang, S.: Algebra. Addison-Wesley, Reading/London/Amsterdam/Don Mills/Sydney/Tokyo (1965)
Lang, S.: $SL(2, \mathbb{R})$. Addison-Wesley, Reading/London/Amsterdam/Don Mills/Sydney/Tokyo (1975)
Lang, S.: Math Talks for Undergraduates. Springer, New York (1999)
Lang, S.: Introduction to Differentiable Manifolds. Springer, New York (2002)
Lange, K., Fessler, J.A.: Globally convergent algorithms for maximum a posteriori transmission tomography. IEEE Trans. on Image Proc. **4**(10), 1430–1438 (1995)
Lavrentiev, M.A., Shabat, B.V.: Методы теории функций комплексного переменного (Methods of the Theory of Functions of Complex Variable). "Наука" Press, Moscow (1973). Russian
Lefschetz, S.: Differential Equations: Geometric Theory. Interscience, New York/London (1957)
Lelong-Ferrand, J.: Les fondements de la géometrie. Presses Universitaires de France (1985 French)
Lenhard, H.-C.: Verallgemeinerung und Verschärfung der Erdös-Mordellschen Satz für Polygone. Archiv der Mathematik **12**, 311–314 (1961)
Leuenberger, F.: Zum Mordellschen Beweis einer Ungleichung von Erdös. Elemente der Mathematik **17**, 15–17 (1962)
Levitan, B.M., Sargsyan, I.S.: Введение в спектральнуя теорию: самосопряжённые обыкновенные дифференциальные операторы. "Наука" Press, Moscow (1970). [English transl. Introduction to Spectral Theory: Selfadjoint Ordinary Differential Operators. American Mathematical Society (1975)]
Loeve, M.: Probability Theory. Graduate Texts in Mathematics, vol. 1–2, 4th edn. Springer, New York (1977–1978)

Lubich, Yu.I.: Линейный функциональный анализ (Linear functional analysis). Итоги науки и техники. Современные проблемы математики. Фундаментальные направления, Т.19: Функциональный анализ – 1. ВИНИТИ Press, Moscow (1988 Russian)
Luce, R.D., Raiffa, H.: Games and Decisions. Introduction and Critical Survey. Wiley/Chapman&Hall, New York/London (1957)
MacDonald, L.G.: Affine root systems and Dedekind's η-function. Invent. Math. **15**, 91–143 (1972)
MacDonald, I.G.: Symmetric Functions and Hall Polynomials. Oxford University Press, New York (1979)
Marcus, M., Minc, H.: A Survey of Matrix Theory and Matrix Inequalities. Allyn and Bacon, Boston (1964)
Massey, W.S.: Algebraic Topology: An Introduction. Harcourt, Brace & World, New York/Chicago/San Francisco/Atlanta (1967)
McLachlan, N.W.: Theory and Application of Mathieu Functions. Dover, New York (1964)
Milnor, J.W.: Morse Theory. Princeton University Press, Princeton/New Jersey (1963)
Milnor, J.W.: Topology from the Differentiable Viewpoint. Princeton University Press, Princeton/New Jersey (1997)
Milnor, J.W., Husemoller, D.: Symmetric Bilinear Forms. Springer, Berlin/Heidelberg/New York (1973)
Moore, G.H.: Zermelo's Axiom of Choice: Its Origins, Development, and Influence. Studies in the History of Mathematics and Physical Sciences, vol. 8. Springer, New York (1982)
Moulin, H.: Théorie des jeux pour l'économie et la politique. Collection méthdes. Hermann, Paris (1981)
Munkres, J.: Topology: A First Course. Prentice Hall College Divison, New Jersey (1974)
Nagle, K.B., Saff, E.B., Snider, A.D.: Fundamentals of Differential Equations, 6th edn. Addison-Wesley, Reading (2003)
Narasimhan, R.: Complex Analysis in One Variable, 2nd edn. Birkhaüsser, Boston (2000)
Naymark, M.A.: Теория представлений групп (Group Representation Theory). "Наука" Press, Moscow (1976). Russian
Neumann von, J.: Mathematische Grundlagen Der Quantenmechanik (Die Grundlehren Der Mathematischen Wissenschaften, Band 38). Springer, Berlin (1932). [English transl. Mathematical Foundations of Quantum Mechanics. Princeton University Press (1996)]
Nirenberg, L.: Topics in Nonlinear Functional Analysis. Courant Institute of Mathematical Sciences, New York University, New York (1974)
Niven, I.: Numbers: Rational and Irrational. New Mathematical Library; The School Mathematics Study Group. Random House, New York (1961)
Niven, I., Zuckerman, H.: An Introduction to the Theory of Numbers, 3rd edn. Wiley, New York (1973)
Nomizu, K.: Lie Groups and Differential Geometry. Mathematical Society of Japan, Tokyo (1956)
Oppenheim, A.: The Erdös inequality and other inequalities for a triangle. Amer. Math. Monthly **68**, 226–230 (1961)
Pakovich, F., Roytvarf, N., Yomdin, Y.: Cauchy-type integrals of algebraic functions. Isr. J. of Math. **144**, 221–291 (2004)
Petrovski (Petrovsky), I.G.: Лекции по теории обыкновенных дифференциальных уравнений, 7th edn. "Наука" Press, Moscow (1984). [English transl. Ordinary Differential Equations. Dover (1973)]
Poincaré, H.: Calcul des Probabilitiés. L'école Polytechniquue, Paris (1912)
Polya, G.: Mathematics and Plausible Reasoning. Princeton University Press, Princeton/New Jersey (1954)
Polya, G.: Mathematical Discovery. On Understanding, Learning, and Teaching Problem SolvingV.'s 1–2. Wiley, New York (1962–1965)
Polya, G.: How to Solve it: A New Aspect of Mathematical Method, 2nd edn. Princeton University Press, New Jersey (1971)

Polya, G., Szegö, G.: Isoperimetric Inequalities in Mathematical Physics. Princeton University Press, Princeton/New Jersey (1951)
Polya, G., & Szegö, G.: Aufgaben und Lehrsätze aus der Analysis. Springer, Göttingen/Heidelberg/New York (1964). [English transl. Problems and Theorems in Analysis I, II. Springer, Reprint edition (1998)]
Poston, T., Stewart, I.: Catastrophe Theory and its Applications. Pitman, London/San Francisco/Melbourne (1978)
Prasolov, V.V., Soloviev, Y.P.: Эллиптические функции и алгебраические уравнения (Elliptic Functions and Algebraic Equations). "Факториал" Press, Moscow (1997). Russian
Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in Fortran 77. The Art of Scientific Computing, 2nd edn. Cambridge University Press, Cambridge/New York/Melbourne (1992)
Pressley, A.N.: Elementary Differential Geometry. Springer Undergraduate Mathematics Series, 2nd edn. Springer, New York (2010)
Privalov, I.I.: Введение в теорию функций комплексного переменного (Introduction to the Theory of Functions of a Complex Variable), 13th edn. "Наука" Press, Moscow (1984). Russian
Rabbot, J.M.: Тригонометрия (Trigonometry). МГУ Press, Moscow (1969). Russian
Rademacher, H.: Topics in Analytic Number Theory. Grundlehren series, vol. 169. Springer, New York (1973)
Reed, M., Simon, B.: Methods of Modern Mathematical Physics. V.1: Functional Analysis. Academic, New York/London (1972)
Rényi, A.: Dialogues on Mathematics. Holden-Day, San Francisco (1967)
Rham, G.: Variétés Différentiebles (Formes, Courants, Formes Harmoniques). Hermann&Cie, Paris (1955). French
Riesz, F. & Sz.-Nagy, B.: Leçons d'analyse fonctionnnelle. Sixtème edition, Budapest (1972). [English transl. Functional Analysis. Dover (1990)]
Robinson, A., Roytvarf, A.: Computerized tomography for non-destructive testing. EU **99971130**, 2–2218 (2001)
Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton/New Jersey (1970)
Rudin, W.: Functional Analysis. McGraw-Hill, New York/St. Louis/San Francisco/Düsseldorf/Johannesburg/London/Mexico/Montreal/New Delhi/Panama/Rio de Janeiro/Singapore/Sydney/Toronto (1973)
Rudin, W.: Principles of Mathematical Analysis. Science/Engineering/Math, 3rd edn. McGraw-Hill, New York (1976)
Rudin, W.: Real and Complex Analysis. International Series in Pure and Applied Mathematic, 3rd edn. McGraw-Hill, Singapore (1986). Science/Engineering/Math
Rumer, Yu.B., Ryvkin, M.Sh.: Термодинамика, статистическая физика и кинетика. "Наука" Press, Moscow (1972). [English transl. Thermodynamics, Statistical Physics and Kinetics. "Mir" , Moscow (1980)]
Sakharov, A.D.: Воспоминания. "Права Человека" Press, Moscow (1996). [English transl. Memoirs. Knopf, New York (1990)]
Sansone, G.: Equazione differenziale ordinaria (Ordinary differential equations), V.'s 1–2. Zanichelli (1948–1949 Italian)
Schmidt, W.M.: In: Dold, A., Eckmann, B. (eds.) Diophantine Approximations. Lecture Notes in Mathematics, vol. 785. Springer, Berlin/Heidelberg/New York (1980)
Schutz, B.F.: Geometrical Methods of Mathematical Physics. Cambridge University Press, Cambridge/London/New York/New Rochelle/Melbourne/Sydney (1982)
Schwartz, L.: Analyse Mathématique I-II. Cours Professé L'ecole Polytechnique. Hermann, Paris (1967)
Schwartz, J.T.: Differential Geometry and Topology. Courant Institute of Mathematical Sciences, New York (1968)

Schwartz, L., Huet, D.: Méthodes mathématiques pour les sciences physiques. Hermann, Paris (1961)
Serre, J.-P.: Lie Algebras and Lie Groups (Lectures Given at Harvard University). W.A. Benjamin, New York/Amsterdam (1965)
Shiryaev, A.N.: Вероятность, 2nd edn. "Наука" Press, Moscow (1989). [English transl. Probability (Graduate Texts in Mathematics, No 95). 2nd edn. Springer (1995)]
Shoenfield, J.R.: Mathematical Logic. Addison-Wesley, Reading (1967)
Sierpiński, W.: O teorii mnogoski. Wybrane zagadnienia dla szkol srednich. Państwowe Zakłady Wydawn. Szkolnych, Warszawa, Wyd. 2 (1969). [Russian transl. О теории множеств (On the theory of sets). "Просвещение" Press, Moscow (1966)]
Sivashinsky, I.C.: Неравенства в задачах (Inequalities in Problems). "Наука" Press, Moscow (1967). Russian
Smith, K.T., Solmon, D.C., Wagner, S.L.: Practical and mathematical aspects of the problem of reconstructing objects from radiographs. Bull. AMS **83**(6), 1227–1270 (1977)
Spivak, M.: Calculus on Manifolds. W.A. Benjamin, New York/Amsterdam (1965)
Spivak, M.: Calculus, 3rd edn. Publish or Perish, Houston (1994)
Stanley, R.P.: Enumerative Combinatorics, vol. 1, 2. Cambridge University Press, Cambridge (1999)
Steinhaus, G.: Kalejdoskop matematyczny. Warszawa (1938). [English transl. Mathematical Snapshots, 3rd edn. Dover (2011)]
Steinhaus, G.: Sto zadań. Państwowe Wydawn. Naukowe, Warszawa (1958). [English transl. One Hundred Problems in Elementary Mathematics. Pergamon Press, Oxford/London/New York / Paris (1963)]
Sternberg, S.: Lectures on Differential Geometry. Prentice Hall, Englewood Cliffs/New Jersey (1964)
Stoker, J.J.: Nonlinear Vibrations in Mechanical and Electrical Systems. Interscience, New York (1950)
Stoll, R.R.: Sets, Logic and Axiomatic Theories, 2nd edn. W.H. Freeman, San Francisco (1975)
Strang, G.: Introduction to Linear Algebra, 4th edn. Wellesley Cambridge Press, Wellesley (2009)
Szegö, G.: Orthogonal Polynomials, 4th edn. American Mathematical Society, Providence/Rhode Island (1981)
Sze-Tsen, H.: Homotopy Theory. Academic, New York/London (1959)
Thorpe, J.A.: Elementary Topics in Differential Geometry. Springer, New York/Heidelberg/Berlin (1979)
Titchmarsh, E.C.: The Theory of Functions, 2nd edn. Oxford University Press, London (1939)
Tóth, L.F.: Inequalities concerning polygons and polyhedra. Duke math. J. **15**, 817–822 (1948)
Trigg, C.: Mathematical Quickies. McGraw-Hill, New York/London (1967)
Van der Waerden, B.L.: Ontwakende Wetenschap. P. Noordhoff. Groningen (1950). [English transl. Science Awakening. Groningen (1954)]
Van der Waerden, B.L.: Mathematische statistik. Springer, Berlin/Göttingen/Heidelberg (1957). [English transl. Mathematical Statistics. Springer (1969)]
Van der Waerden, B.L.: Algebra II. Springer, Berlin/Heidelberg/New York (1967)
Van der Waerden, B.L.: Algebra I. Springer, Berlin/Heidelberg/New York (1971)
Venkataraman, C.S.: In: Problems and solutions. Math. Magazine **44**(1), 55 (1971)
Venkov, B.A.: Элементарная теория чисел. "ОНТИ" Press, Moscow and Leningrad (1937). [English transl. Elementary Number Theory (Translated and edited by H. Alderson). Wolters-Noordhofl, Groningen (1970)]
Vilenkin, N.Ya.: Специальные функции и теория представлений групп. "Наука" Press, Moscow (1965). [English transl. Special Functions and the Theory of Group Representations. Translations of Mathematical Monographs 22, American Mathematical Society, Providence/ Rhode Island (1968)]
Vilenkin, N.Ya.: Комбинаторика. "Наука" Press, Moscow (1969–1). [English transl. Combinatorics. Academic Press (1971)]

Vilenkin, N.Ya.: Рассказы о множествах. “Наука” Press, Moscow (1969–2). [English transl. Stories about Sets. Academic Press (1968)]
Vinogradov, I.M. (ed.): Математическая энциклопедия, TT. 1–5. “Советская Энциклопедия” Press, Moscow (1977–1984). [English transl. and update. Encyclopedia of Mathematics (1988–1994). Reidel, 10 volumes + supplements (1997-)]
Vogler, H.: Eine Bemerkung zum Erdös-Mordellschen Satz für Polygone. Anz. Österr. Akad. Wiss., Math.- naturwiss. Klasse **103**, 241–251 (1966)
von Leichtweiß, K.: Konvexe Mengen. VEB Deutscher Verlag, der Wissenschaften, Berlin (1980)
von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press, Princeton/New Jersey (1953)
Walker, R.J.: Algebraic Curves. Princeton University Press, Princeton/New Jersey (1950)
Wallace, A.H.: Differential Topology: First Steps. Dover Books on Mathematics. Dover, Mineola (2006)
Warner, F.W.: Foundations of Differentiable Manifolds and Lie Groups. Springer, New York/ Berlin/Heidelberg/Tokyo (1983)
Weil, A.: Ľintêgration dans les groupes topologiques et set applications (Integration on topological groups and its applications). Paris (1940 French)
Weyl, H.: The Classical Groups. Their Invariants and Representations. The Institute for Advanced Studies. Princeton University Press, Princeton, New Jersey (1939)
Weyl, H.: Symmetry. Princeton University Press, Princeton/New Jersey (1952)
Whitney, H.: Geometric Integration Theory. Princeton University Press, Princeton/New Jersey (1957)
Whittaker, E.T., Watson, G.N.: A Course of Modern Analysis (An Introduction to the General Theory of Infinite Processes and of Analytic Functions; with an Account of the Principal Transcendental Functions). Cambridge University Press, Cambridge (1927)
Wiener, Z., Yomdin, Y.: From formal numerical solutions of elliptic PDE’s to the true ones. Math. Comput. **69**(229), 197–235 (2000)
Williams, C.S.: Designing Digital Filters. Prentice Hall International, New Jersey (1986)
Yomdin, Y.: Discrete Remez inequality. Isr. J. of Math. (submitted)
Yomdin, Y., Zahavi, G.: High-order processing of Singular Data. Non-commutativity and singularities. Proceedings of French-Japanese symposia, IHES (2006). Edited by J.-P. Bourguignon, M. Kotani, Y. Maeda and N. Tose: Advanced studies in pure mathematics, 55, Mathematical Society of Japan, Tokyo 173–207 (2009)
Yosida, K.: Functional Analysis. Die Grundlehren Der Mathematischen Wissenschaften, vol. 123. Springer, Berlin/Göttingen/Heidelberg (1965)
Zalenko, M.S.: Периодические последовательности. (Periodic sequences). Мат. Просв. **3**(10), 195–205 (2006). Russian
Zariski, O., Samuel, P. (with the cooperation of Cohen, I.S.): Commutative Algebra, V.’s 1–2. D. Van Nostrand, Princeton/Toronto/London (1958–1960)
Zel’dovich, Y.B., Myskis, A.D.: Элементы математической физики (Elements of Mathematical Physics). “Наука” Press, Moscow (1973). Russian
Zel’dovich, Ya.B.: Высшая математика для начинающих. “Физматгиз” Press, Moscow (1963). [English transl. Higher Mathematics for Beginners. Central Books. (1974)]
Zel’dovich, Ya.B., Myskis, A.D.: Элементы прикладной математики, 2nd rev. edn. “Наука” Press, Moscow (1972). [English transl. Elements of Applied Mathematics. Central Books Ltd (1977)]
Zel’dovich, Ya.B., Yaglom, I.M.: Высшая математика для начинающих физиков и техников. “Наука” Press, Moscow (1982). [English transl. Higher Mathematics for Beginning Physicists and Engineers. Prentice Hall (1988)]
Zermelo, E.: Collected Works, V.1: Set Theory, Miscellanea/Gesammelte Werke, B.1: Mengenlehre, Varia. Springer, Heidelberg/Dordrecht/London (2010)
Zetel, S.I.: Новая геометрия треугольника (A New Geometry of the Triangle). “Учпедгиз” Press, Moscow (1962). Russian

Index

A.A. Roytvarf, *Thinking in Problems: How Mathematicians Find Creative Solutions*,
DOI 10.1007/978-0-8176-8406-8,